Resource Discovery in Heterogeneous Digital Content Environments

George Macgregor
0000-0002-8482-3973

Department of Computer & Information Sciences

University of Strathclyde

A thesis submitted for the degree of PhD by published works at the University of Strathclyde

Doctor of Philosophy

2020

Acknowledgements

Big thanks to Dr Diane Rasmussen Pennington for her helpful supervision, support and candid comments on draft chapters – and for her dog training tips. Thanks also to my second supervisor, Dr David McMenemy, for persuading me that I should finally pursue a PhD.

The research documented in works selected for inclusion in this thesis were in some instances supported by – or were the outcome of – research and development projects. In particular, I would like to acknowledge the following Jisc-funded projects: the COPAC/Clumps Technical Cooperation Project (CC-Interop), the High-level Thesaurus Project (Phases III & IV), and Principles in Patterns (PiP).

I would also like to acknowledge the co-authors of PWs 2, 4, 6, 7, 8: Emma Graham (née McCulloch) (University of Strathclyde), Fraser Nicolaides (University of Greenwich), Dennis Nicholson (University of Strathclyde - retired), Anu Joseph (University of Strathclyde).

Thanks are also extended to Aaron Tay (Singapore Management School), Dr Martin Brändle (University of Zürich), and Dr Péter Király (Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen) for providing useful feedback on my repository related research works.

Declaration

This thesis is the result of the author's original research. The work contained therein has been composed by the author and has not been previously submitted for examination which has led to the award of a degree.

Signed:

Date:

Statement of contribution

This statement confirms that of the 11 published works selected for this portfolio, 6 are single-authored. The remaining 5 are co-authored. All of co-authored works stem from projects in which I was either a principal investigator or lead author of the work. My specific contributions to each of the co-authored works has been highlighted in the analysis of each work (Chapters 3-7) but is also specified in Chapter 2.

Signed:

Date:

Abstract

The concept of 'resource discovery' is central to our understanding of how users explore, navigate, locate and retrieve information resources. This submission for a PhD by Published Works examines a series of 11 related works which explore topics pertaining to resource discovery, each demonstrating heterogeneity in their digital discovery context. The assembled works are prefaced by nine chapters which seek to review and critically analyse the contribution of each work, as well as provide contextualization within the wider body of research literature. A series of conceptual sub-themes is used to organize and structure the works and the accompanying critical commentary. The thesis first begins by examining issues in distributed discovery contexts by studying collection-level metadata (CLM), its application in 'information landscaping' techniques, and its relationship to the efficacy of federated item-level search tools. This research narrative continues but expands in the later works and commentary to consider the application of Knowledge Organization Systems (KOS), particularly within Semantic Web and machine interface contexts, with investigations of semantically aware terminology services in distributed discovery. The necessary modelling of data structures to support resource discovery – and its associated functionalities within digital libraries and repositories – is then considered within the novel context of technology-supported curriculum design repositories, where questions of human-computer interaction (HCI) are also examined. The final works studied as part of the thesis are those which investigate and evaluate the efficacy of open repositories in exposing knowledge commons to resource discovery via web search agents.

Through the analysis of the collected works it is possible to identify a unifying theory of resource discovery, with the proposed concept of *(meta)data alignment* described and presented with a visual model. This analysis assists in the identification of a number of research topics worthy

of further research; but it also highlights an incremental transition by the present author, from using research to inform the development of technologies designed to support or facilitate resource discovery, particularly at a 'meta' level, to the application of specific technologies to address resource discovery issues in a local context. Despite this variation the research narrative has remained focussed on topics surrounding resource discovery in heterogeneous digital content environments and is noted as having generated a coherent body of work.

Separate chapters are used to consider the methodological approaches adopted in each work and the contribution made to research knowledge and professional practice.

Contents

List of Figures

Chapter 1

Introduction: towards a unifying theme of resource discovery

1.1 Overview

The research works assembled for this submission of PhD by Published Works, published between 2003 and 2020, contribute to our understanding of resource discovery and address questions arising within a series of heterogeneous discovery contexts. In addition to the critical commentary that accompanies them, the 11 selected works contribute to a unifying theme of resource discovery and enable the notion of '(meta)data alignment' to be proposed. The works are accompanied by five themed chapters, all of which provide critical commentary on the research assembled and consider the works' impact in relation to the extant literature. This critical commentary also considers the methodological merits of each work and the implications for the present author's ongoing research agenda. The precise structure of the thesis is detailed in sections 1.3 and 1.4 below, and details of the published works are provided in Chapter 2, with each published work reproduced in Appendix B. It is useful to contextualize this body of work by first understanding the concept of resource discovery and how it relates to the works assembled for this thesis.

1.2 The concept of resource discovery

The potential of information to be valuable is key to its conceptualization as an 'information resource'. Information resources are expressly designed or conceived by someone, or something, to convey meaning, with the intention that the resource be used for a specific purpose. This ability to convey meaning and knowledge gives information its value. Today such information resources are typically characterized by their application of human-readable and/or machine-readable characters, such that they are conductive to replication and long-term storage [3]. The concept of 'resource discovery' is therefore central to our understanding of how users explore, navigate, locate and retrieve information resources.

Formal definitions of resource discovery can vary in their specificity. Courtney [4] states that resource discovery entails 'locating resources that are unknown' to the user. This definition emphasises the capacity of discovery tools or systems to surface both 'hidden collections' and new information relevant to users' overall information needs, but also to provide users with navigational aids to support discovery within increasingly complex digital collections.

Noted informatician Clifford A. Lynch provides an exhaustive exploration of resource discovery concepts [5]. Lynch formally defines resource discovery as a 'complex collection of activities that [...] range from simply locating a well-specified digital object on the network all the way through lengthy iterative search activities' [5]. Even simpler attempts at defining resource discovery [6] highlight the systematization of information resources as key to providing users with a 'consistent, organized view of information'. Lynch [5] elaborates by noting the process of identifying a set of potentially relevant information resources as being central to resource discovery, with the organization and ranking of resources within the set, and their expansion or filtering according to specified criteria, as being especially important to our understanding of the concept. Typical examples of discovery include the 'searching of various types of directories, catalogs or other descriptive databases'.

We can therefore state that resource discovery underpins users' information seeking behaviour by providing mechanisms through which users' can satisfy their information needs [7]. These mechanisms can be varied but, in general, support users' ability to locate the information resources which correspond to the requirements specified in a user query, which might be submitted to a resource discovery service. The query may be user generated, mediated by machine, or may even be entirely machine generated on behalf of the user depending on the type of resource being requested; but ultimately the delivery, supply or support of 'resource' from the discovery service is provided if users' query requirements are matched by the resource discovery service. These resource discovery services will typically assume the form of one or more information retrieval systems, digital libraries or digital repositories, each based largely or entirely on the use of surrogate descriptions of digital content, such as metadata or other forms of structured data.

A significant body of theoretical and philosophical work exists which seeks to provide a conceptual model of what constitutes an information resource [8]. This abstract work has been helpful in understanding the value of information, especially within corporate contexts and knowledge-based industries. However, within the domain of resource discovery, typical examples of information resources remain far more material. They include information objects such as documents, files, data, and multimedia content, all in human and machine-readable form and capable of being called within a networked environment [9, 10, 11]. These information resources are sought by users because they help fulfil an information need [12].

It is worth acknowledging that the concept of resource discovery can also encompasses several different communities of practice. For example, the emergence of ubiquitous computing (ubicomp) over recent decades [13] and, more recently, human-centered computing (HCC) has expanded the scope of resource discovery to include 'resource' as a type of computational resource, as nodes within the Internet of Things (IoT), or as distributed systems, among other types [14]. Similarly, resource discovery as a basis for delivering information resources to users has become embedded within users' information seeking behaviour [15, 16, 17, 18]. The ubiquity of the web, improved digital literacy among users and the proliferation of networked digital devices has resulted in increased user engagement with information retrieval tools, digital libraries and repositories. Discovering — and negotiating with — these information resources through every day human-computer interactions has consequently become a typical activity for any information user [19].

The purpose of this thesis is to explore a series of interlinked research topics within the broader research theme of resource discovery using the selected published works. In their taxonomy of resource discovery, Vanthournout et al. [20] note that resource discovery involves three principal actors: resource providers, resource users, and the resource discovery service itself. The works assembled for this thesis contribute to the body of knowledge on resource discovery and examine each of these taxonomic actors in different ways. Collectively the works contribute to our understanding of how provider, user and resource discovery service shapes the efficacy of resource discovery within heterogeneous digital information environments. To achieve this the works study several of the mechanisms known to underpin the concept of resource discovery. These mechanisms have represented the research focus of the present author's career and include the following interrelated areas:

1. Metadata – and more generally structured data, e.g. applications of RDF/XML, etc.;

2. Knowledge organization;

3. Distributed systems interoperability, especially the syntactic and semantic interoperability issues that arise from numbers 1 and 2, and;

4. The influence of numbers 1, 2 and 3 on information retrieval and aspects of human-computer interaction (HCI).

With users confronting rapidly changing resource discovery environments, providing a better understanding of these mechanisms is essential to ensure optimum levels of information engagement from users. Without ongoing attention there will always remain a possibility that users' information needs will go unsatisfied, with consequent implications for everyday task completion and new knowledge creation that this implies. A critical analysis and commentary of the assembled works therefore represents the main body of this thesis. It will seek to contextualise the work within extant research literature, provide commentary and critically appraise its contribution to the relevant fields of study.

1.3 Resource discovery sub-themes

The published works selected for inclusion in this thesis are listed and annotated in Chapter 2. The works span different types of research contribution; some provide conceptual or theoretical background to specific research areas or problem spaces, while many others are experimental, frequently describing the development or deployment of new technologies and/or their evaluation. Irrespective of the 'type' of contribution, the works are grouped according to categories, with each category occupying a sub-theme within the broader topic of this thesis. The works are presented largely in a chronological order, reflecting the evolving sophistication of the present author's reasoning about the specific topics explored within resource discovery, as well as maturation in the methodological approaches adopted.

The sub-themes are as follows:

- Resource discovery within digital libraries.

- Resource discovery concepts within Knowledge Organization Systems (KOS) & Semantic Web contexts.

- Human-computer interaction (HCI) & curriculum design repositories.

- Open science: resource discovery & open repositories.

The very fact that resource discovery is the unifying theme connecting these sub-themes is significant; but it can also be noted that relationships exist between the individual works, thereby demonstrating a coherence to the body of work presented. Even between some of the earliest selected published works and the latest there are clear intellectual overlaps; from the role of repositories in delivering research content to users, to works demonstrating the application of resource discovery expertise within alternative communities of practice, e.g. within a repository designed to store XCRI compliant metadata about curriculum designs being generated within a UK HEI.

A conceptual model of how the selected works relate to one another is provided in Chapter 2. This model will be described in more detailed in the following chapter and will be referred to throughout the thesis. The nature of the relationships between the selected works – represented as nodes in the model – will also be described.

1.4 Thesis structure & approach

The thesis is structured as follows: Chapter 2 provides bibliographic details for the research works which have been selected for inclusion in the thesis. This chapter also includes brief annotations and key information about each work, as well as a conceptual model describing the way in which the selected works are linked.

Chapters 3-6 will consider the works in the context of each of the resource discovery sub-themes, outlined above. As such, Chapter 3 will explore 'Resource discovery within digital libraries'; Chapter 4, 'Resource discovery concepts within KOS & Semantic Web contexts'; Chapter 5, 'Human-computer interaction (HCI) & curriculum design repositories' and, finally, Chapter 6, 'Open science: resource discovery & open repositories'. These chapters provide a commentary on the set of works assembled, to explain their background, common themes and linkages, context in the literature, methodological approaches, research contribution, limitations, impact, and fit with the present author's ongoing research agenda.

Chapter 7 will consider the methodological evolution of the present author, as displayed in the published works. Chapter 8 then uses the assembled published works to propose a unifying theory of resource discovery through the concept of *(meta)data alignment* and provides an exploration of potential future research questions arising from the works. Finally, the wider contribution of the works to academic knowledge and practice is considered in Chapter 9, as well as their collective importance.

The selected published works are presented as an appendix to the thesis and form Appendix B.

Chapter 2

Selection of published works

The published works selected for inclusion in this thesis are listed below in section 2.1. Alongside the full bibliographic details, 2.1 also provides a brief rationale for the inclusion of each work, a summary of their research contribution and the received citations at time of writing. The works are organized according to the sub-themes introduced in Chapter 1. They are as follows:

- Resource discovery within digital libraries.
- Resource discovery concepts within Knowledge Organization Systems (KOS) and Semantic Web contexts.
- Human-computer interaction (HCI) and curriculum design repositories.
- Open science: resource discovery and open repositories.

Some related additional works are also referred to a various points within the main body of the thesis. These works are distinguished from other cited literature by appearing in bold typeface, e.g. [**21**] rather than [22] for all other literature. Their details are included in the References section but are not included in the formal submission owing to restrictions of space and submission requirements. A full list of the present author's published works is also provided in **Appendix A**.

With the exception of one work, all the selected works for this thesis were peer-reviewed and published in the formal literature as journal articles or conference papers. The exception is a report of evaluative work [**21**] conducted under the auspices of a research project and delivered as a published deliverable for the project. The intellectual justification for its inclusion within this thesis will be provided in Chapter 5. Suffice to state that it comprises a detailed scholarly contribution which is commensurate in quality to the other works selected for inclusion.

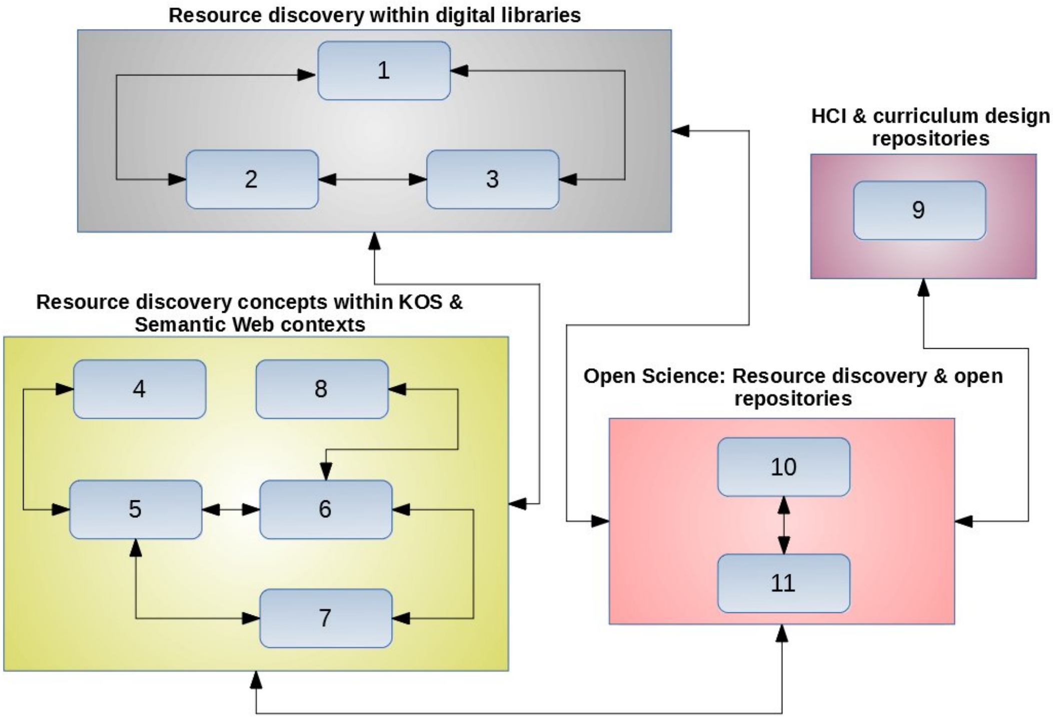

Resource discovery within digital libraries
1
2
3
HCI & curriculum design repositories
9
Resource discovery concepts within KOS & Semantic Web contexts
4
8
5
6
7
Open Science: Resource discovery & open repositories
10
11

Figure 2.1: Conceptual model of the published works assembled, the sub-themes to which they are assigned, and their interrelationships.

Chapter 1 noted that 'resource discovery' was the unifying theme connecting the sub-themes, providing additional coherence to the body of work presented, but that additional relationships also existed between individual works. A conceptual model demonstrating this coherence is provided in Fig. 2.1. This model diagrams the works listed in section 2.1 below as nodes, with the various relationships and interconnections noted between published works. Although they are not diagrammed as such, it could be suggested that the relationship between these nodes is almost hierarchical insofar as the works considered within 'Resource discovery in digital libraries' demonstrate the widest subject scope, while those in 'Open science: resource discovery and open repositories' — the final published works to be considered as part of this thesis — demonstrate a considerable narrowing in subject scope.

2.1 Bibliographic details & summaries

The bibliographic details of the published works selected for this thesis are provided below. Each work is accompanied by a brief rationale and, where applicable, its citations, as calculated by Google Scholar[1]. For shorthand reference within the thesis proper, each work is numbered using the convention 'PW' (e.g. PW1 = Published Work 1).

The contribution made by the present author to each co-authored work is also provided using CRediT (Contributor Roles Taxonomy) [2]. CRediT specifies 14 contributor roles that are typically performed in the creation of a scholarly work and standardizes their definition. These roles describe specific aspects of the production of a work, such as contributing to the 'methodology' or 'formal analysis'. Not all 14 roles are applicable to the selected co-authored works but all relevant roles are specified alongside the co-authored works for this thesis.

2.1.1 Resource discovery within digital libraries

1. **Published work 1 (PW1)** Macgregor, G. (2003) Collection-level description: metadata of the future? *Library Review*, 52 (6). pp. 247-250.

   Available: `https://doi.org/10.1108/00242530310482015`

   A brief conceptual paper exploring the role of collection-level metadata (or 'description') in supporting user resource discovery within the context of rapidly growing digital libraries and other large heterogeneous information environments. The concept of 'functional granularity' and 'information landscaping' is explored as is its role in defining resource collections by administrators of those collections but also by users.

   Citations acquired at time of writing: 26

2. **Published work 2 (PW2)** Macgregor, George and Nicolaides, Fraser (2005) Towards improved performance and interoperability in distributed and physical union catalogues. *Program*, 39 (3). pp. 227-247.

   Available: `https://doi.org/10.1108/00330330510610573`

   Evaluative research undertaken to investigate disparities in the performance of competing discovery models within digital libraries: centralized (physical) and distributed (virtual) bibliographic discovery services. Observations gleaned by the research resulted in numerous practical implications for those establishing distributed systems based on the Z39.50 information retrieval protocol and search/retrieve web services, as well as those establishing centralized systems.

[1] George Macgregor: `https://scholar.google.co.uk/citations?user=nDfa5GMAAAAJ`

[2] CRediT — Contributor Roles Taxonomy: `https://casrai.org/credit/`

This work received 'Highly Commended Paper' award from Program in 2005.

Citations acquired at time of writing: 9

Co-author contribution: The present author was responsible for the following: *writing – original draft* and *writing – review & editing.* The present author contributed equally with his co-author in the following areas: *conceptualization, formal analysis, investigation, validation, visualization.*

3. **Published work 3 (PW3)** Macgregor, George (2005) Z39.50 broadcast searching and Z-server response times: perspectives from CC-interop. *Online Information Review*, 29 (1). pp. 90-106.

   Available: `https://doi.org/10.1108/14684520510583963`

   This work explores the influence of broadcast searching on so-called 'Z-server' response times, noting that in 2019 Z39.50 still remains an important machine interface to digital libraries and OPACs (Library of Congress, 2019). The research involved search tests using 17 different Z-servers, analysis of the results and conclusions drawn on the suitability of the Z39.50 protocol in distributed discovery models. The work is a notable contribution on the study of preferable models of union discovery services (or catalogues), i.e. physical or virtual.

   Citations acquired at time of writing: 6

2.1.2 Resource discovery concepts within KOS & Semantic Web contexts

4. **Published work 4 (PW4)** Macgregor, George and McCulloch, Emma (2006) Collaborative tagging as a knowledge organization and resource discovery tool. *Library Review*, 55 (5). pp. 291-300.

   Available: `https://doi.org/10.1108/00242530610667558`

   A conceptual and literature based work exploring the emergence of 'collaborative tagging' as a mechanism to facilitate information management and resource discovery. A conceptual analysis of collaborative tagging against more formal subject retrieval mechanisms is presented (e.g. thesauri, taxonomies, ontologies, etc.) and issues with the technique highlighted. This is a highly cited work published at a critical point in the evolution of tagging approaches within online discovery tools.

Citations acquired at time of writing: 444

Co-author contribution: the following responsibilities were shared, although the present author assumed a leading role, hence the granting of first authorship: *writing – original draft* and *writing – review & editing.* The present author contributed equally with his co-author in the following areas: *conceptualization, investigation.*

5. **Published work 5 (PW5)** Macgregor, George (2009) E-resource management and the Semantic Web : applications of RDF for e-resource discovery. In: *The E-Resources Management Handbook - UKSG.* UKSG, Newbury, pp. 1-20. ISBN 9780955244803

   Available: `https://doi.org/10.1629/9552448-0-3.20.1`

   This work provides a theoretical introduction to some essential Semantic Web concepts and the resource description framework (RDF), a key enabling language of the Semantic Web. It exhibits theoretical understanding and fluency with Semantic Web technologies, applications of RDF (e.g. FOAF, SKOS, OWL, DC, RDFa) and outlines applications within digital libraries and other e-resource contexts. This provides a theoretical foundation for subsequent works included within this thesis sub-theme; although the work itself was published after these other works.

   Citations acquired at time of writing: 6

6. **Published work 6 (PW6)** Macgregor, George and Joseph, Anu and Nicholson, Dennis; Prasad, A.R.D and Madalli, Devika P., eds. (2007) A SKOS Core approach to implementing an M2M terminology mapping server. In: *International Conference on Semantic Web & Digital Libraries (ICSD 2007).* Documentation Research & Training Centre, Bangalore, India, pp. 109-120.

   Available: `https://strathprints.strath.ac.uk/2970/`

   The first of several works selected for this thesis exploring the potential of Semantic Web and Linked Data approaches to facilitating resource discovery. This particular paper describes the use of the W3C Simple Knowledge Organization System (SKOS) in the implementation of a machine interface (or API) to deliver a functioning terminology server, capable of mediating subject based searches across different knowledge organization systems. SKOS is shown to be useful to wrap terminology responses, consumable by digital libraries, repositories, etc.

Citations acquired at time of writing: 16

Co-author contribution: The present author assumed sole responsibility for the following: *writing – original draft, writing – review & editing, conceptualization, visualization.* The present author contributed equally with his co-authors in the following areas: *software, methodology.*

7. **Published work 7 (PW7)** Macgregor, G. and McCulloch, E. and Nicholson, D. (2007) Terminology server for improved resource discovery: analysis of model and functions. In: *Second International Conference on Metadata and Semantics Research*, 2007-10-11 - 2007-10-12.

   Available: `https://strathprints.strath.ac.uk/3435/`

   This work is a companion paper to the previous one and demonstrates the web service requests supported by the proposed terminology server, based on SKOS for data structuring and SOAP / SRW for machine requests. The terminology server model, employing a Dewey Decimal Classification (DDC) spine approach, is outlined, as is the system architecture and possible resource discovery use cases for the server.

   Citations acquired at time of writing: 6

   Co-author contribution: The present author contributed equally with his co-authors in the following areas: *writing – original draft, writing – review & editing, conceptualization, investigation, visualization, software, methodology.*

8. **Published work 8 (PW8)** McCulloch, E. and Macgregor, G. (2008) Analysis of equivalence mapping for terminology services. *Journal of Information Science*, 34 (1). pp. 70-92.

   Available: `https://doi.org/10.1177/0165551507079130`

   Using prior work surrounding a SKOS based terminology server as the context, this work considers the equivalence or mapping types required to facilitate interoperability in the context of a distributed terminology server. The SKOS Core Mapping Vocabulary Standard (MVS) and other mapping types are tested against terminological mappings within the terminology server. An alternative and generic suite of match types is proposed, although more detailed than the MVS proposition. It has been subsequently cited by many researchers investigating the deployment of semantically aware systems within specific knowledge domain discovery tools (e.g. astronomy) and within digital libraries more generally.

Citations acquired at time of writing: 31

Co-author contribution: Despite the second authorship, both authors contributed equally to the creation of this work. Authorship order was determined by a coin toss. The present author contributed equally in the following areas: *writing – original draft, writing – review & editing, conceptualization, methodology, investigation, data curation, validation, visualization, software, methodology.*

2.1.3 Human-computer interaction (HCI) & curriculum design repositories

9. **Published work 9 (PW9)** Macgregor, George (2012) *Principles in Patterns (PiP) : User Acceptance Testing of Course and Class Approval Online Pilot (C-CAP).* [Report]. University of Strathclyde, Glasgow.

   Available: `https://strathprints.strath.ac.uk/46510/`

   There is growing interest in the use of technology-based approaches to improve the quality, reuse potential and discovery of curriculum designs within HEIs. This work – the only non-peer-reviewed work selected for inclusion this thesis – formed part of an evaluative strand in the Principles in Patterns (PiP) project. The work is broadly concerned with 'user acceptance testing' of a technology-based curriculum design tool, devised to improve curriculum design quality but also enable the deposit of approved designs into a 'design repository' for the purposes of discovery, sharing and reuse (via XCRI compliant metadata). The general evaluative approach adopted employs a combination of standard Human-Computer Interaction (HCI) techniques and specially designed data collection instruments, including protocol analysis, stimulated recall and pre- and post-test questionnaire instruments.

   Citations acquired at time of writing: N/A

2.1.4 Open science: resource discovery & open repositories

10. **Published work 10 (PW10)** Macgregor, George (2019) Improving the discoverability and web impact of open repositories: techniques and evaluation. *Code4Lib Journal* (43).

    Available: `https://journal.code4lib.org/articles/14180`

    This work is the first of two which evaluate the effect of repository optimization techniques on the discovery potential of open repositories on the web. The work outlines the approaches implemented and reports on comparative search traffic data and usage metrics, and delivers conclusions on the efficacy of the techniques implemented. The evaluation provides persuasive evidence that specific enhancements to technical aspects of a repository can result in significant improvements to repository visibility, resulting in a greater web impact and consequent increases in COUNTER compliant content usage. All supporting and underlying data are made openly available.

    Citations acquired at time of writing: 3

11. **Published work 11 (PW11)**

    Macgregor, George (2020) Enhancing content discovery of open repositories: an analytics-based evaluation of repository optimizations. *Publications*, 8(1), 8.

    Available: `https://doi.org/10.3390/publications8010008`

    This second work within the current sub-theme is a more detailed continuation of the previous work, using a larger dataset with which to verify and validate findings. A deeper exploration of the data, alongside additional statistical analyses, is performed. As in the previous work, all relevant data are made openly available.

    Citations acquired at time of writing: Work has yet to acquire citations owing to its recent publication.

2.2 Presentation of published works

All of the published works (**PW1-PW11**) have been reproduced in **Appendix B**. Full bibliographic details are provided above in section 2.1 for each work if the final published version is sought directly from the publisher.

Using the conceptual model presented in Fig. 2.1 and the resulting sub-themes, the next seven chapters will now provide a critical commentary of the assembled published works, exploring their background, impact, context within the wider body of extant literature, methodological competency and linkage with the other works selected.

Chapter 3

Resource discovery within digital libraries

This chapter first considers contributions to the topic of resource discovery within digital library contexts. This will include consideration of the first three published works presented in this thesis: **PW1**, **PW2** and **PW3**, as will be the convention in this thesis. As earlier sections have established, resource discovery can occur in a wide variety of user contexts. The contributions presented here explore the role of collection-level metadata (or collection-level description) in supporting 'information landscaping' and ergo federated searching of digital libraries and repositories. The ability of such discovery tools to be delivered to end users is often determined by the efficacy of the technical protocols underpinning the federated searching, but also the degree of semantic interoperability observed between search targets.

3.1 Collection-level metadata & 'information landscaping'

Collection-level metadata (CLM) is predicated on a desire to improve the browsing and searching of large, multi-corpus, multi-format and often distributed digital collections or information services. Within the context of digital libraries, collection-level metadata can be deployed as an important resource discovery mechanism. CLM provides structured, open, standardized and machine-readable metadata providing a high-level description of an aggregation of individual items in both digital and physical environments [**23**]. Although such metadata can support a number of different information management functions [24, 25], the principal motivation is that collection-level metadata can support users' information seeking since such metadata provides a simple and convenient way of delivering ‘first level access’ to large information corpora [26]. CLM assists because it groups resources into convenient collections. The metadata used to describe these collections then provides a suitable access point for users to enter relevant collections and retrieve information at the item level, while simultaneously excluding less relevant collections [27]. The idea of excluding or filtering out less relevant collections is an important one within ‘information landscaping’ and will be revisited in the next section.

The relevance of CLM in facilitating discovery has increased in recent years as the volume of content within digital collections has grown, within both formal collections served by digital libraries or repositories [28, 29, 30, 31, 32, 33, 34, 35]; but also within services delivering user-generated content, such as photo-sharing platforms [36, 37]. Service discovery tools, or so-called ‘service registries’, have also adopted collection-level metadata application profiles. These have been designed to expose machine readable data about collections of resources which can then be interrogated by software applications, such as portals [38, 39].

Although there has been significant early work undertaken to define consistent approaches and standards for CLM [40], much of the subsequent research area has been informed by the analytic models proposed by Heaney [41, 42]. These models informed the development of the RSLP Collection Description schema [43], the Dublin Core Collections Application Profile [44], and the IESR Application Profile [39]. Richer implementations of Heaney’s analytic model, demonstrating greater hierarchical and associative relationships, have also been implemented, such as that proposed by the SCONE and CC-Interop projects [45, 46, 47, 48, 49], **[1]**. Those falling into the former category tend to demonstrate ‘flat file’ characteristics but have nevertheless been found to be adequate within their given applications (Fig. 3.1). In fact, the terminology service described in Chapter 4 and, in particular, within PW6 and PW7 demonstrate a `get_collections` function to assist in the identification of digital collections by service registries. This terminology service function uses the Dublin Core Collections Application Profile and the IESR Application Profile to model the terminological and collection data returned to clients.

Despite its adequacy, the Dublin Core Collection Description Terms [44] has been updated to demonstrate greater granularity as well as better alignment with the Resource Description Framework (RDF), as per the Semantic Web and Linked Open Data (LOD) conventions. More recently, the Europeana cultural heritage platform, which encompasses 60 million digital items, has incorporated CLM into the Europeana Data Model (EDM) [50], aspects of which drive some browsing functionality

Figure 3.1: Example CLM record in XML, adhering to the IESR Application Profile & incorporating shared elements from RSLP Collection Description schema & Dublin Core Collections Application Profile.

```
<?xml version="1.0"?>
<iesrd:iesrDescription
 xmlns:dc="http://purl.org/dc/elements/1.1/"
 xmlns:dcterms="http://purl.org/dc/terms/"
 xmlns:dcmitype="http://purl.org/dc/terms/dcmitype/"
 xmlns:iesr="http://iesr.ac.uk/terms/#"
 xmlns:iesrd="http://iesr.ac.uk/"
 xmlns:rslpcld="http://purl.org/rslp/terms#"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://iesr.ac.uk/ http://iesr.ac.uk/schemas/xsd/iesr.xsd">
 <dcmitype:Collection>
  <dc:identifier xsi:type="dcterms:URI">
   http://scone.strath.ac.uk/coln/7952
  </dc:identifier>
  <dc:title>
   Dept. of Computing Science and Mathematics eTheses
  </dc:title>
  <dcterms:abstract xml:lang="en">
   Electronic copies of theses produced by students from the Department of Computing
       Science and Mathematics of the University of Stirling.
  </dcterms:abstract>
  <dc:type xsi:type="dcterms:DCMIType">Collection</dc:type>
  <dc:type xsi:type="rslpcd:CLDT">
   Collection.Internet.Text.Image.Special.Form.Virtual
  </dc:type>
  <iesr:hasService xsi:type="dcterms:URI">
   http://scone.strath.ac.uk/coln/7953
  </iesr:hasService>
  <iesr:hasService xsi:type="dcterms:URI">
   http://dspace.stir.ac.uk/dspace/handle/1893/36
  </iesr:hasService>
  <dc:subject xsi:type="dcterms:LCSH">
   Computer science
  </dc:subject>
  <dc:subject xsi:type="dcterms:LCSH">
   Mathematics
  </dc:subject>
  <dc:subject xsi:type="dcterms:DDC">
   004
  </dc:subject>
  <dc:subject xsi:type="dcterms:DDC">
   510
  </dc:subject>
  <rslpcd:owner xsi:type="dcterms:URI">
   http://scone.strath.ac.uk/agnt/5393
  </rslpcd:owner>
  <dcterms:isPartOf xsi:type="URI">
   http://scone.strath.ac.uk/coln/7911
  </dcterms:isPartOf>
 </dcmitype:Collection>
</iesrd:iesrDescription>
```

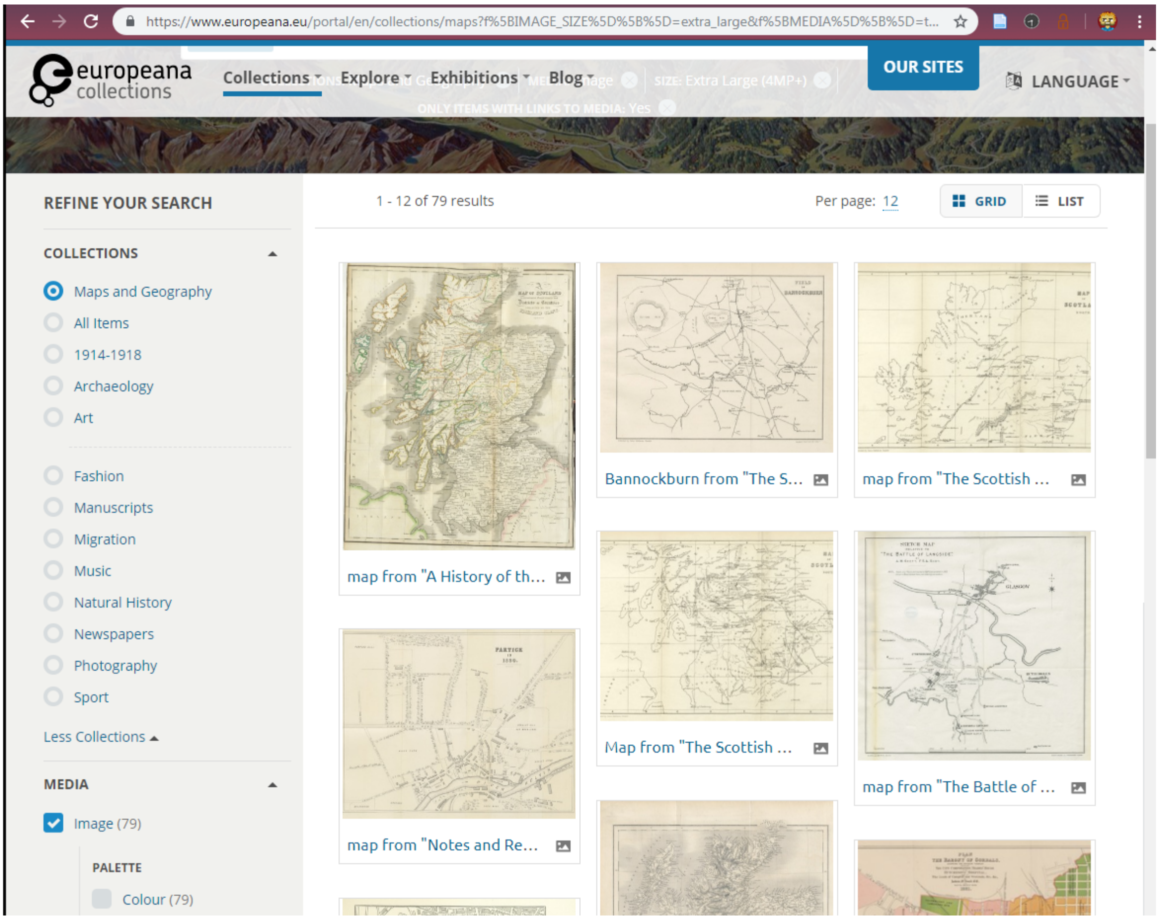


Figure 3.2: User interface of Europeana, employing the EDM to power collection browsing functionality.

for users (Fig. 3.2). Europeana has devised a metadata applications profile aligned with the Dublin Core Collections Application Profile. This profile takes advantage Heaney's analytic model and harnesses advances in Semantic Web data modelling via the Resource Description Framework (RDF) [33].

3.2 Information landscaping & 'functional granularity'

The first published work (**PW1**) presented as part of this chapter is a brief conceptual paper [**23**], which addresses the value of CLM in resource discovery and its potential for supporting users in satisfying their information needs. It was noted previously that a useful characteristic of CLM based systems is their ability to simplify the information landscape. For example, the information landscape presented to the user may be generated with specific grouping criteria. This may be based on subject strengths, language, accessibility, digital format, or even geographical location if we are referring to physical resources. Systems employing CLM therefore have the capacity to enable users to identify potentially fruitful collections worthy of item-level search, while enabling them to discard those collections which are considered to contain fewer relevant items. As we shall see from the published works presented in the next section, issues surrounding interoperability become a key issue when users traverse the landscape and reach the 'discovery level' (see Fig. 3.3), where item level retrieval tools are presented to the user, such as digital libraries, repositories or library catalogues. These interoperability issues arise owing to the distributed nature of these services.

PW1 introduces a number of concepts that were undocumented in the literature at the point of publication. PW1 therefore contributed to domain understanding of how ideas of 'functional granularity' and dynamic information landscaping can be used within CLM based systems. In other words, it describes the creation of systems that use the richness of CLM schema to deliver flexible collection grouping criteria for users (i.e. 'dynamic landscaping'). When deployed in digital libraries or repositories this approach frequently resembles the kind of faceting that might be observed on e-commerce websites or certain generic search tools. Similarities do exist; but, owing to the richness of CLM schema, far richer and more precise filtering is possible, often based on varied and extensive criteria.

PW1's contribution to understanding appears to be corroborated by its influence in subsequent works exploring approaches to CLM and CLM-based systems, e.g. [51, 52, 37, 53, 34, 35] and also provides further theoretical discussion around the idea

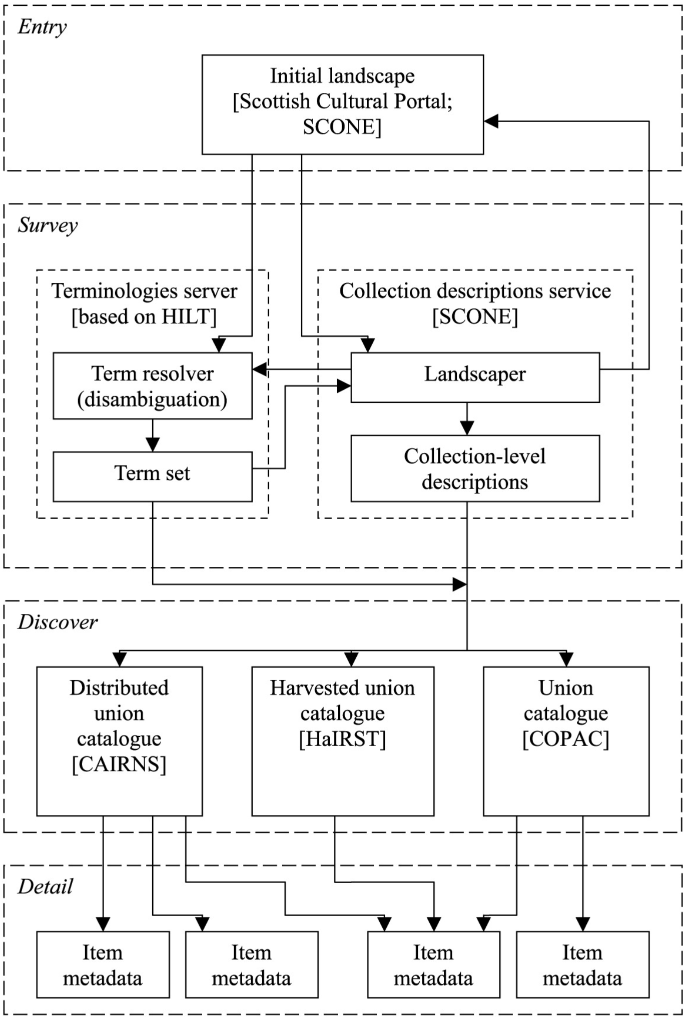

Entry
Initial landscape [Scottish Cultural Portal; SCONE]
Survey
Terminologies server [based on HILT]
Collection descriptions service [SCONE]
Term resolver (disambiguation)
Landscaper
Term set
Collection-level descriptions
Discover
Distributed union catalogue [CAIRNS]
Harvested union catalogue [HaIRST]
Union catalogue [COPAC]
Detail
Item metadata
Item metadata
Item metadata
Item metadata

Figure 3.3: A model of an information environment which employs dynamic landscaping through the use of CLM-based services, as presented in [1].

of 'functionality granularity', a method for defining the parameters of, say, a digital collection and how it should be recorded in CLM [54]. In particular, it explains the practical implementation of functional granularity as a flexible one to be performed by collection administrators — and one which naturally creates levels of granularity to be traversed by the user when replicated by digital services, thereby providing a useful navigation aid for users of large digital corpora as well as providing an effective filtering mechanism.

In critically reviewing PW1 it is evident that no research questions are explicitly articulated. As the concept of CLM and their application within digital libraries was a new one, and published literature addressing CLM was limited at the point of publication, PW1 ultimately sought to inform readers of the potential benefits of using CLM based systems in resource discovery; but also to highlight cogent conceptual questions surrounding collection modelling, collection definition, information landscaping and functional granularity. By highlighting these conceptual questions PW1 is presenting, however inexplicitly, a series of suggested future research areas. Its inclusion here as a published work is therefore designed to demonstrate the present author's abstract understanding of how CLM and its applications can be harnessed to support resource discovery strategies.

PW1 benefits from insights gained under the auspices of two research and development projects around the time of publication (SCONE and CC-interop), both of which experimented with CLM approaches to information landscaping and resource discovery, e.g. [**1**],[55]. Both projects developed prototype systems capable of being embedded within the architectures similar to that diagrammed in Fig. 3.3, and in some instances these prototypes also served as pilot services to users [56], [**57**].

Despite the conceptual and 'real world' merits of PW1, a clear limitation remains its brief nature, absence of research framework for the conceptual discussion and absence of visual models, which could have aided analysis of such an abstract topic, especially within the wider CLM research agenda. Subsequent work in this space [58, 59, 33, 53, 34] would have benefited from such a framework because -– as acknowledged by these cited works -– problems surrounding collection definition and modelling has impeded research attempting to understand how best CLM can support resource discovery. Conceptual work undertaken for the EDM [50] has arguably been the only extant work which has attempted to specify 'representational requirements' to assist in the definition and modelling of collections [33, 60], something which PW1 could have proposed, albeit embryonically.

PW1 also fails to be explicit about the conceptual boundaries of CLM. In other words, its purpose is to articulate the benefits of CLM and highlight specific conceptual questions, but a failure to specify a framework introduces uncertainty about whether these conceptual questions are exhaustive or whether there are others which have been omitted. This limitation is reinforced by the lack of any visual model to support the conceptual discussion and the absence of any caveat to control for this omission. Despite all of this — as noted above -– the work nevertheless remained a necessary contribution to scholarly discourse at the point of publication because community understanding about the potential of CLM was relatively unknown.

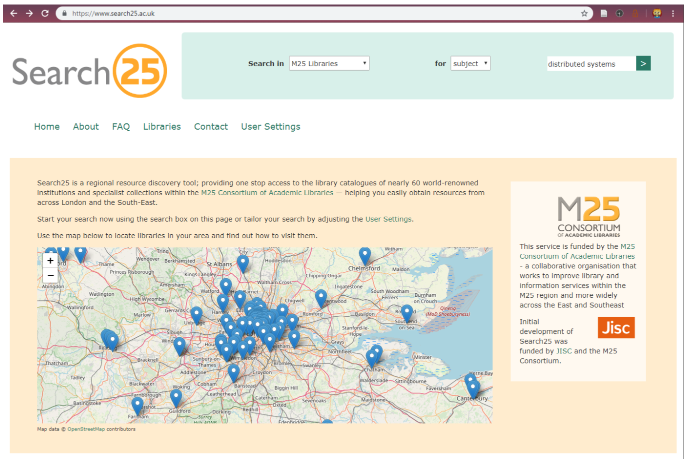
3.2.1 Federated search, clumping & interoperability issues – PW2 & PW3

The ‘discover’ and ‘detail’ stages of the landscaping process within a larger information environment is the point at which item-level discovery becomes relevant (Fig. 3.3). As noted, issues surrounding interoperability become more challenging when users reach the ‘discovery level’ and when the federated search of multiple distributed services is offered to users. The transition to distributed item-level retrieval tools such as digital libraries, repositories or library catalogues demands a level of technical and semantic interoperability between services in order to provide reliable federated item-level retrieval. Many of the interoperability challenges between services have been known for a considerable time (Borgman, 2002) but many persist and remain difficult to solve without adequate technical standardization or metadata harmonization [61, 62, 63, 64, 65].

The idea of ‘clumping’ is associated with the proliferation of digital libraries supporting the Z39.50 Information Retrieval protocol [66]. Z39.50 is a client–server, application layer protocol for searching and retrieving information from remote services over a TCP/IP network [67]. It specifies procedures and formats for a client to search database(s) hosted by a server, retrieve records, and execute related information retrieval functions. The use of Z-client software (employing the Z39.50 protocol) enables a single Z-client to connect to multiple Z-servers (or ‘clumps’). This approach allows the client to ‘broadcast’ a single search (i.e. perform a federated search or meta-search) to multiple Z-enabled services simultaneously, with results from each service retrieved and merged into a single result set for the user (e.g., Fig. 3.4 – Search25). Data retrieved by the client are typically in MARC or XML but can span a variety of formats and serializations [68].

Z39.50 remains a popular protocol despite other, arguably preferable web service approaches, especially those demonstrating RESTful characteristics [69]. Its wide adoption and software support through software toolkits such as YAZ [70] has meant that Z39.50 continues to be deployed in new digital library applications [71, 72, 73, 74, 75]. It has also found recent applications within areas as diverse as massive data sharing platforms for meteorological disaster data [76] and digital library recommender systems [77].

Studying specific interoperability issues within distributed digital libraries with regards to item-level search is the focus of **PW2** and **PW3**. These works dovetail with PW1 insofar as a) they correspond with the discovery layer diagrammed in Fig.



Search25 is a regional resource discovery tool; providing one stop access to the library catalogues of nearly 60 world-renowned institutions and specialist collections within the M25 Consortium of Academic Libraries — helping you easily obtain resources from across London and the South-East.

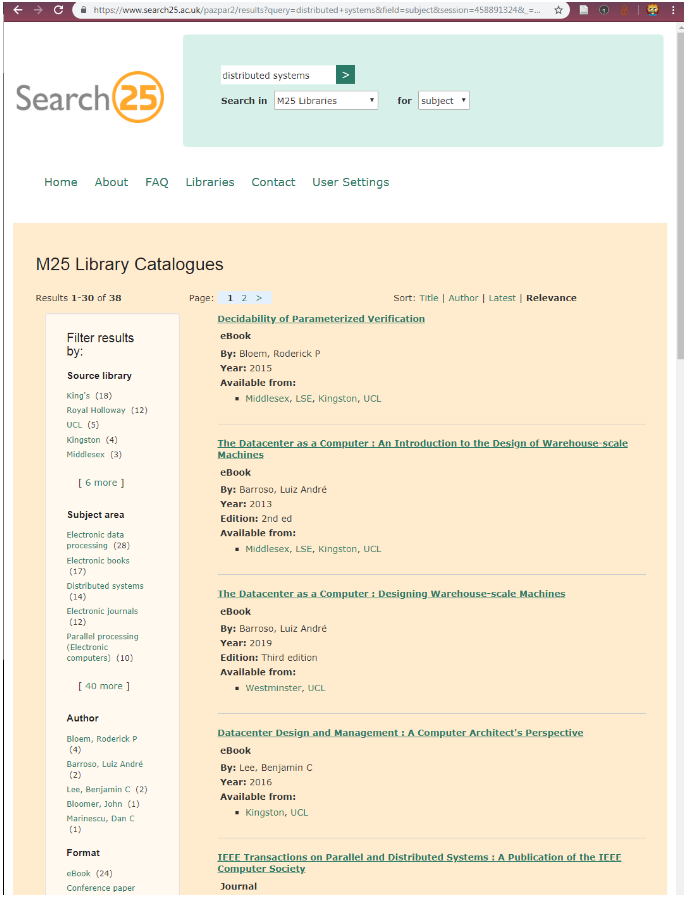
Start your search now using the search box on this page or tailor your search by adjusting the User Settings.

Use the map below to locate libraries in your area and find out how to visit them.

Map data © OpenStreetMap contributors

This service is funded by the M25 Consortium of Academic Libraries - a collaborative organisation that works to improve library and information services within the M25 region and more widely across the East and Southeast

Initial development of Search25 was funded by JISC and the M25 Consortium.


https://www.search25.ac.uk/pazpar2/results?query=distributed+systems&field=subject&session=458891324&_=...
Search25
distributed systems
>
Search in
M25 Libraries
for
subject
Home
About
FAQ
Libraries
Contact
User Settings

M25 Library Catalogues

Results **1-30** of **38** Page: **1** 2 > Sort: Title | Author | Latest | **Relevance**

Filter results by:

Source library

King's (18)
Royal Holloway (12)
UCL (5)
Kingston (4)
Middlesex (3)

[6 more]

Subject area

Electronic data processing (28)
Electronic books (17)
Distributed systems (14)
Electronic journals (12)
Parallel processing (Electronic computers) (10)

[40 more]

Author

Bloem, Roderick P (4)
Barroso, Luiz André (2)
Lee, Benjamin C (2)
Bloomer, John (1)
Marinescu, Dan C (1)

Format

eBook (24)
Conference paper

Decidability of Parameterized Verification

eBook

By: Bloem, Roderick P
Year: 2015
Available from:

- Middlesex, LSE, Kingston, UCL

The Datacenter as a Computer : An Introduction to the Design of Warehouse-scale Machines

eBook

By: Barroso, Luiz André
Year: 2013
Edition: 2nd ed
Available from:

- Middlesex, LSE, Kingston, UCL

The Datacenter as a Computer : Designing Warehouse-scale Machines

eBook

By: Barroso, Luiz André
Year: 2019
Edition: Third edition
Available from:

- Westminster, UCL

Datacenter Design and Management : A Computer Architect's Perspective

eBook

By: Lee, Benjamin C
Year: 2016
Available from:

- Kingston, UCL

IEEE Transactions on Parallel and Distributed Systems : A Publication of the IEEE Computer Society

Journal

Figure 3.4: Clumping and performing broadcast searches on the Search25 service, `https://www.search25.ac.uk/`.

3.3, b) follow the same overall conceptual thread, and c) follow on from any information landscaping functionality facilitated by CLM. The context for both works is the federated search of digital libraries, particularly the technical and semantic interoperability issues inherent in 'clumping' approaches to distributed searching. PW2 [**78**] is a co-authored work but one which the present author leads. The study design, execution, and analysis was shared between the authors, with a larger proportion assumed by the present author.

PW2 is essentially divided into two parts. The first documents system search tests conducted on a series of Z-enabled OPACs and a comparative analysis of the relative performance of the results delivered by a distributed architecture (i.e. federated or clumped) and a centralized equivalent (i.e. physical union catalogue (Copac)). Perhaps predictably, the findings of the work highlight the inherent tension between almost all distributed systems and their centralized counterparts [79]. That is, the degree of autonomy afforded to distributed systems and whether these systems fulfil the requirements and demands requested by the other systems in the cooperative. We use the term 'cooperative' here to denote systems, or in this case digital libraries or similar, which have entered into an information environment of distributed services. Such information environments specify certain technical expectations on individual services as a rule to participating in the wider cooperative. These technical expectations will typically describe the minimum syntax and semantics of the participating services [79]. This model is popular in most communities of practice where interactions between systems occurs. It allows organizations to exert control over local systems and provides freedom from any dependences to other systems while satisfying the expectations of the cooperative. In other words, providing the technical expectations of the cooperative are satisfied the distributed systems can operate as if centralized.

The predictability of PW2 arises because the findings identified low adherence to the rules of the cooperative such that collective performance of the distributed digital libraries was often poorer than the centralized one. In essence, the minimum syntactic and semantic expectations were not being satisfied by the distributed services. A lack of support for the Z39.50 Bib-1 attribute set — within which submitted queries are semantically defined -– was observed. Wide variation in metadata quality across sites and low semantic alignment to support subject-based queries was also identified (an issue to be addressed more exhaustively in Chapter 4).

It is perhaps interesting to note that CLM has been proposed as a mechanism for providing effective filtering tools, thereby helping users to reduce information overload before attempting to discover relevant item-level resources [29]. While the cooperative demonstrated superior data currency when compared to the centralized case study, what PW2 demonstrates — and subsequent works presented in later chapters — is that item level syntactic and semantic interoperability issues were compromising search quality for users. This finding conforms to the prototypical 'known issue' within distributed system theory [79]. Whilst CLM may provide filtering opportunities for users, it is clear that poor adherence to the rules of the cooperative compromises retrieval, despite the use of the Z39.50, the Bath Profile [80], and MARC with AAC2.

Despite the detailed treatment of its methodological approach to data collection and its acknowledgement of caveats, a level of methodological naivety is evident from PW2. The process of data collection for the first part of PW2 used a comparative case study approach [81], with 'analytic induction' used to create a descriptive model of the retrieval problems observed across the different Z-servers, each employing variant configurations and metadata conventions. That there is a methodological label for this approach is omitted in PW2 and instead this is only implicitly suggested rather than explicitly stated. Articulating that the approach to data collection in PW2 subscribed to a recognized qualitative approach would have added credibility to an otherwise sound methodological approach and could have provided greater validation of the overall findings.

The second part of PW2 involved data collection at two workshops. These workshops were facilitated as focus groups, the use of which were considered essential to elicit the required rich qualitative data from digital library and system practitioners. The operation of the focus groups, and the data collection techniques used, were highly successful in the field. Documentation of how the qualitative data were analysed and the conclusions drawn was less successful, as evidenced by PW2. A large volume of qualitative data were gathered during the focus groups, with group discussions transcribed, organized according to high-level themes and circulated among focus group participants for comment or correction. These steps in the data analysis are curiously omitted from PW2's methodology, thereby compromising the transparency of the methodological approach, replicability of the study and potentially undermining the safeness of the findings.

Upon reflection it is clear that whilst the data collection were sound for this portion of PW2, the weakness was a failure to treat such a rich qualitative dataset with sufficient analytical detail. Data coding were undertaken to only a shallow level (i.e. high-level themes), with a reliance on 'lumping' [82]. 'Lumping' can be an expedient way of analysing large volumes of qualitative data [82, 83] but has long been recognized in seminal works as being subject to superficiality as the process of lumping often means the coder inadvertently avoids careful scrutiny of the data [84]. Little attention was therefore paid to surfacing potentially relevant subordinate concepts or themes as part of PW2's analysis. No coding framework for the focus group data was therefore generated and ergo presented in PW2. If a more sophisticated qualitative approach had been adopted, such as Grounded Theory [85], combined with a more exhaustive approach to coding, additional insights from the data may have been exposed and a superior summary of the qualitative data analysis could have been presented in PW2. It is apposite to note that the use of qualitative methods — as well as mixed methods to deliver triangulation in research findings — is something which the present author has deployed extensively in other works, including in one of the published works selected for this thesis (i.e. PW9) [**21**]. Suffice to state that this analytical failure was never repeated in any subsequently published, or indeed unpublished, works.

The aforementioned methodological oversights from PW2 can certainly appear obvious in a historical critique but few of them realistically undermine the overall findings of the work. None were identified as problematic via the peer-review process either. Instead PW2 contributed to understanding of the management of distributed digital libraries and the interoperability problems to be solved. Its impact was that it was the first and only study of its kind to engage in such an evaluation and confirmed a negative finding, which hitherto had been only acknowledged via anecdotal evidence [**1**], [86, 87, 88]. This enabled the creation of national strategies and transferable recommendations on Z-server management and metadata practices across research institutions [**89**], which national services such as Copac implemented [66]. Although the findings were acknowledged in the literature [90] it nevertheless remains the case that such a base failure in distributed digital library management persists more recently in related communities of practice, particularly in relation to scholarly open repository implementations using the OAI-PMH protocol [91, 92, 32], a discussion point to which we will return in Chapter 6. It is consequently possible to conclude that there are sections of the digital library domain which fail to successfully launch services externally and/or successfully integrate within other systems, or aggregated services.

PW3 is related to PW2 and occupies a similar intellectual space [**93**] insofar as it again evaluates the efficacy of a group of Z-servers, with federated search being the typical research use case. The work could also be said to explore the issues surrounding 'transparency' within distributed digital libraries.

‘Transparency’ in the context of distributed systems and relevant reference models [94], is the goal of hiding the fact that the system’s processes and resources are actually physically distributed across multiple servers. In other words, distributed systems should ideally present themselves to users as if they were a single, centralized system. For this to be successful within the use case documented in PW3, it is necessary for users’ expectations of the distributed digital library to compare favourably with the centralized model. This is especially true with respect to the efficacy of the Z39.50 protocol itself, which in the early years of its adoption in digital libraries was considered sluggish [95]; but also with respect to how individual services have adhered to the syntactic and semantic rules of the cooperative.

Parallel research involving the same use case systems -– but undertaken by collaborators in order to better understand users’ expectations [96, 97] -– found evidence of what has controversially been termed ‘Google generation’ user behaviour [19]. User expectations surrounding retrieval response times, influenced by interactions with Google and similar services, meant that understanding the response times, system impediments, etc. that might undermine search performance within a distributed digital library model was necessary. The research which was conducted as part of PW3 was therefore designed to better understand the performance issues with a view to informing the development of new, more successful digital libraries.

The so-called ‘quick and dirty’ Z39.50 implementations at institutions found in PW3 once again suggested poor adherence to the rules of the cooperative and therefore lower levels of transparency. Results suggested that improved treatment of complex search queries, greater harmonization in Z-server configurations and lower time-out thresholds might deliver performance enhancements. However, in general -– and especially when Z-servers were configured correctly -– transparency could be successfully maintained and user expectations better fulfilled. Z-servers within the cooperative tended to respond rapidly and network congestion and local usage of services was not found to significantly influence Z-server performance. These findings may be considered encouraging in the context of Z-enabled digital libraries but are difficult to reconcile against users’ retrieval and HCI expectations found in related work [96, 97].

It can be posited that the influence of PW3 may have been greater had there been a more imaginative approach to presenting the evidence. Data charts are adequate for presenting such a vast volume of data but their effective interpretation remains difficult without tabulated data summarises, or indeed access to the raw datasets as would be de rigueur in 2020. The volume of data doubtlessly made the inclusion of tabulated data undesirable in this work; yet in subsequent works it has been possible to adopt creative approaches to the summarization of far larger datasets. The failure to adopt such approaches in PW3 is therefore a clear limitation of the work. The limitation is particularly obvious in this instance because specific data points are discussed in the main body of the work but cannot be verified easily by consulting the charts. A tabulated summary of data providing response times across systems, with appropriate segmentation of the data according to measures of central tendency, level of variance (SD), and IQR would have been appropriate and may have prompted additional data insights. Upon reflection it is odd that this deficit went unreported during peer-review and therefore remained unaddressed in the final published work.

Nevertheless, as the only study of its kind the contribution made by PW3 to the wider research agenda surrounding distributed digital libraries clearly reinforced the viability of Z39.50 based approaches to distributed digital library item-level retrieval, as noted previously by the continued deployment of the approach in digital library applications. Unfortunately, many of the reported semantic interoperability and Z-server configuration problems persist. For example, Kapidakis & Sfakakis [98] describe similar difficulties in delivering 'meta-search' functionality involving FRBRized digital libraries as well as low semantic interoperability between services, once again indicating that basic lessons surrounding the management of distributed architectures have not been learned.

Chapter 4

Resource discovery concepts within KOS & Semantic Web contexts

This chapter will continue the discussion of Chapter 3 by progressing onto concepts surrounding resource discovery within subject-based item-level retrieval contexts, specifically the development and deployment of terminology services and the exploitation of Semantic Web approaches to achieve this (**PW5**, **PW6**, **PW7**, **PW8**). Consideration will also be given to **PW4**. PW4 revisits some core principles in knowledge organization and the creation of knowledge organization systems (KOS) within the context of Web 2.0 collaborative tagging. PW4 warrants collective analysis alongside the other published works of this chapter owing to its focus on KOS principles which, as we shall demonstrate in later sections, are core to the terminological approaches adopted in PW5-PW8, as well as many vocabulary specifications used within the Semantic Web. Its relevance in this regard will be explained in more detail in the first section of this chapter. It could be suggested that, from PW4 onwards, a more mature academic writing style is visible the present author's published works.

PW4, PW6, and PW7 are co-authored works on which the present author leads, reflecting the leading role assumed in writing the works, forming their approach and — in the case of PW6 and PW7, during which time the present author was a research fellow — being principally responsible for leading the underlying research project. PW8 is the final co-authored work of the thesis, devised and written during the same time as PW6 and PW7. PW8 was written during a time of great collaboration and productivity. The present author is listed second in PW8; however, both authors contributed equally to the work, with the order of attribution determined by a coin toss. The methodological approach and its execution, data analysis and conclusions were therefore shared equally across both authors.

4.1 Knowledge Organization Systems

Knowledge organization systems (KOS) are conceptual and terminological devices used to present systematized interpretations of knowledge [99]. These devices exert control, not just over the way in which this knowledge is organized, but also over the terminology used to describe certain knowledge concepts (i.e. vocabulary control).

KOS typically encompass the following types, ranked here based on their semantic sophistication:

1. Term lists, such as authority files, gazetteers, glossaries, etc.

2. Hierarchical relational vocabularies, such as information retrieval thesauri and subject heading lists.

3. Taxonomic classification schemes, such as bibliographic classification schemes, taxonomies, etc.

4. Ontologies and knowledge graphs, such as both upper and domain ontologies.

It is long established that the incorporation of KOS into retrieval tools — in their various permutations — can perform an important role in improving resource discovery outcomes, e.g. [100, 101, 102]. The benefits of concept structure and the control exerted over the vocabulary used within KOS are explained within PW4 [**103**] and will not be reproduced here (see section 'Defining Controlled Vocabularies' in PW4). Suffice to state that vocabulary control (e.g. control for synonyms, homonyms, lexical anomalies, etc.) used by KOS, and the resultant control exerted over indexing, ensures the terms used to describe concepts are standardized and therefore similar or related resources are collocated for ease of discovery by the user. Further ease of discovery is promoted through hierarchical and syntactic relationships, as well as coding or notation — the latter of which continues to find uses within tools such as the MeSH Browser [1] and UNESCO Thesaurus browser [2].

Owing to the recent proliferation of KOS in supporting commercial retrieval systems (inc. web search engines) and information retrieval within an ever growing number of digital content platforms (e.g. digital repositories, digital libraries, cultural asset collections, etc.), the variety of KOS types now far exceeds the typical examples provided above [104, 105]. Recent work has sought to propose a 'taxonomy'

[1] MeSH Browser: `https://meshb.nlm.nih.gov/search`

[2] UNESCO Thesaurus Browser: `http://vocabularies.unesco.org/browser/thesaurus/en/`

for KOS types based on their relative semantic complexity, taking account of how KOS are now central to the functioning of the Semantic Web and Linked Data [106].

Naturally, digital libraries and repositories have often been at the forefront of KOS innovation, with many active and prototype systems in use. See for example [107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119]. Aside from their use to improve search performance in such digital platforms, most of the innovation in KOS integration has been for the purposes of user navigation, resource browsing (e.g. facetted browsing, facetted results refinement, etc.) or resource display. These types of resource discovery aid are underpinned by the hierarchical and associative concept relations encoded by KOS and support well understood information seeking behaviours surrounding browsing [120, 121]. They also support the alternating information seeking strategies employed by specific user communities [122, 123].

Additionally, query expansion (QE) functionality [124] is something which KOS are uniquely scoped to provide. Such use of interactive QE can assist users in their query formulation and there are numerous examples in the literature of KOS powered query expansion within digital libraries, repositories and medical literature corpora [125, 126, 127, 22, 128, 129, 130, 131]. As we shall see in later sections, functional approaches to QE were among the anticipated use cases of the terminology services described in PW6, PW7 and PW8.

4.2 KOS interoperability & terminology services

Problems surrounding the interoperability of KOS has long been an active research area [132]. As additional applications for KOS continue to be found, especially within Linked Data and biomedical contexts, the need for advances in KOS interoperability remains a focus of research activity in order to improve subject-based searching and browsing across services [119]. Digital libraries and repositories have increased the need for such interoperability in order to facilitate user access to discrete heterogeneous digital objects. Within distributed resource discovery contexts, such as those described in Chapter 3, this need is especially true since digital objects held across distributed systems will tend to be indexed and organized according to different KOS [**133**]. Encountering disparate KOSs is the inevitable reality of resource discovery within these systems because different terminologies will generally have been used to meet the subject searching and browsing requirements of local users, or to better describe the digital collection within wider metadata requirements. The impracticality of querying multiple services individually, or even acquainting oneself with the terminologies or KOS in use, is such that federated subject-based searching becomes not only necessary, but critical.

A failure to adhere to a single KOS could be described as yet another failure by systems to satisfy the semantic expectations of the 'cooperative' model, as we described previously in Chapter 3. However, KOS deployment within typical digital resource platforms frequently exists independently of other metadata interoperability requirements owing to the importance of subject-based resource discovery in users' unknown item searching [134]. A lack of discipline specificity in more general, universal schemes, such as Library of Congress Subject Headings (LSCH)[3] limits their application to similarly general, universal collections. Discipline specific services are therefore better fulfilling subject-based requirements via domain specific KOS (e.g. MeSH[4], STW Thesaurus for Economics[5], HASSET[6], the NASA GCMD[7], etc.), with scientific and biomedical applications in particular stimulating the creation of numerous domain ontologies [135, 136, 137].

Improving the ability of users to engage in federated, subject-based resource discovery of disparate discipline-specific repositories and digital libraries is the predominant focus of the published works associated with this chapter; facilitating semantic interoperability and, specifically, achieving interoperability between KOS so that subject-based federated resource discovery is possible. Some of the difficulties in achieving this — to be explored in more detail in later sections — demonstrate the immense problems in semantic interoperability. So, before exploring terminology services and KOS interoperability, it is worth highlighting the related concepts highlighted in PW4.

4.3 Semantic expressiveness of knowledge structures & collaborative tagging: PW4

PW4 is ostensibly about the emergence collaborative tagging as a popular mechanism for organizing digital content [**103**]. Collaborative tagging (or simply 'tagging') emerged in parallel with the broader trend of Web 2.0 in the mid-2000s, in which a

[3] Library of Congress Subject Headings (LCSH): `http://id.loc.gov/authorities/subjects.html`

[4] Medical Subject Headings (MeSH): `https://www.nlm.nih.gov/mesh/`

[5] STW Thesaurus for Economics: `http://zbw.eu/stw/version/latest/about`

[6] Humanities and Social Science Electronic Thesaurus (HASSET): `https://hasset.ukdataservice.ac.uk/`

[7] NASA Global Change Master Directory (GCMD): `https://wiki.earthdata.nasa.gov/display/CMR/NASA+GCMD+Keywords`

growth in user-generated content and participatory digital cultures on the Web was considered indicative of a ‘second generation’ in the World Wide Web [138]. The popularity of tagging as a new, preferable approach to organizing content was in a large part aided by several high-profile talks delivered by Internet sociologists and new media writers, particularly Clay Shirky (e.g. [139]). Shirky’s notions of how knowledge, or more specifically digital content, should be organized and ergo discovered was influenced by familiar arguments that existing approaches to KOS creation failed to reflect users’ real requirements [140]. Shirky posited that KOS were frequently delivering biased interpretations of knowledge domains and therefore organization could better be delivered by the participatory user base associated with Web 2.0 (i.e. ‘organization goes organic’ [139]).

The relevance of this to terminology services is that collaborative tagging, and the so-called ‘folksonomies’ they produced, was frequently proposed by its advocates as a way of addressing semantic interoperability problems on the web, and even replacing the Semantic Web altogether [141]. Most works published during the emergence of Web 2.0 noted the potential for user generated knowledge to contribute to aspects of the ‘web of data’, e.g. [142, 143, 144]; but also noted that the purported potential of tagging was unrealistic and failed to acknowledge the inherent challenges in providing semantics for both humans and machines, making tasks associated with the Semantic Web such as better resource discovery unachievable [145].

That tagging was contrary to well understood principles in information retrieval and knowledge organization was the focus of PW4. Within PW4 a definition of KOS is proposed (ironically referred to by the synonym, ‘controlled vocabularies’) and used to assess the efficacy of collaborative tagging. Though merely a review and a conceptual exercise to measure collaborative tagging as an effective knowledge organization mechanism, PW4 is the most cited published work presented as part of this thesis. PW4 could be described as a Zeitgeist work insofar as it was -– at the point of publication — the only review of extant literature and the only published attempt to logically assess the efficacy of tagging as a knowledge organization mechanism. For this reason it has acquired an impact arguably incommensurate with its real significance. PW4 nevertheless exposes a fluency in the construction of KOS such that the limitations of tagging could be assessed logically, theorized and articulated.

For example, early experimental work reported positive results in the use of tagging data to generate coherent knowledge structures [146, 141]. But these results have not always been borne out by subsequent work. Term noise and a lack of expressiveness, hierarchical or syntactic structure are highlighted as difficulties in harnessing crowd-sourced tagging data to generate coherent knowledge structures [147]. Dong et al's review [147] notes positive progress in the research literature but reinforces the unreliability of data mining, machine learning or semantic mapping techniques to extract meaning from tagging data. This disappointment has been found in experimental studies exploring the retrieval efficacy of tags in a variety of online settings. For instance, Lorince et al. [148] investigated the efficacy of tagging based retrieval within online music services such as Last.fm[8]. Their data suggest tags did not generally serve as retrieval aids nor did they predict listening behaviour or function in personal information management; instead they appeared to be purely a 'participative' exercise on the part of users.

Where attempts have been made to harness this participation for the purposes of augmenting the Semantic Web findings have been disappointing [149]. Markines et al. [149] concluded that it was computationally unscalable to perform the level of similarity analyses required on large-scale tagging corpora. The intriguing aspect to these disappointments has been research exploring greater user intervention to essentially annotate content more effectively (i.e. for users to create rudimentary metadata). Passant et al. [150] describe a lightweight collaborative Semantic Web framework which can underpin Web 2.0 services but which also invites users to annotate content more effectively, annotations which can then be translated into machine readable statements via RDF [150]. Similarly, Zhang & Cranshaw [151] demonstrated a prototype system designed to enrich group chat content by presenting users with opportunities to 'mark-up' their chats as a supplement to their tagging data [151].

It is worth noting that the idea of inviting greater annotation of unstructured data has been successful in the case of Wikidata[9] -– as a component of the wider participative Web 2.0 service, Wikipedia. Wikidata been able to attract sufficient volunteers to curate an extensive and growing corpus of data statements as key-value pairs, thereby relating concepts, objects and things to one or more values [152, 153]. The exposure and openness of these statements holds particular potential for Linked Open Data (LOD) and Semantic Web applications, and is already being harnessed by digital libraries, heritage platforms and repositories to improve authority data [154, 155] or augment existing structured data [156]. Experiments on extracting a crowdsourced KOS of some kind from Wikidata is presenting numerous challenges but is nevertheless proving more productive than prior attempts with tagging data [157].

[8] Last.fm: `https://www.last.fm/`

[9] Wikidata: `https://www.wikidata.org/`

This is not to state that there has not been progress in harnessing tagging data or incorporating it into digital information platforms. See indicative examples, [158, 159, 160, 161, 162, 163]. However, where it has been harnessed to support information retrieval, tests confirm the corollary of the logic outlined in PW4: that tagging data tend to reveal high recall and low precision [164, 165, 162, 166, 167]. This, in turn, echoes the theoretical and experimental work of previous decades, specially surrounding the use of early free-text indexing in the 1970s and beyond -– the conclusions of which were that free-text was most productive when combined or mapped to existing KOS [168, 169], much as the successful tagging research has demonstrated more recently. The efficacy of the folksonomies generated from collaborative tagging are simply limited owing to their lack of semantic richness and expressiveness, a phenomenon which has been the focus of numerous KOS typologies over recent years. See for example, the ISKO Encyclopedia of Knowledge Organization, which reviews numerous KOS typologies, most of which attempt to arrange KOS according to their characteristics, with semantic expressiveness forming an important criteria [170]. In fact, we can diagram the relationship between semantic expressiveness and resource discovery (Fig. 4.1). As the semantic expressiveness and complexity of a KOS increases on the X axis, a proportional increase in the resource discovery power of the KOS on the Y axis can generally be observed.

If we accept that the lack of semantic expressiveness is a problem with folksonomies, then semantic interoperability across different folksonomies will be even more difficult than with formal KOS. This leads to the need for improved and continued interoperability between formal KOS and the availability of potential solutions. It also returns us to the intellectual work highlighted earlier in this chapter surrounding the need to better support subject-based resource discovery, a line of enquiry which remains an active research area [171, 172].

4.4 Terminology services

Terminology services can assume a number of manifestations but a useful definition has been provided by Tudhope et al.: Terminology Services (TS) are a set of services that present and apply vocabularies, both controlled and uncontrolled, including their member terms, concepts and relationships. This is done for purposes of searching, browsing, discovery, translation, mapping, semantic reasoning, subject indexing and classification, harvesting, alerting etc. [173].

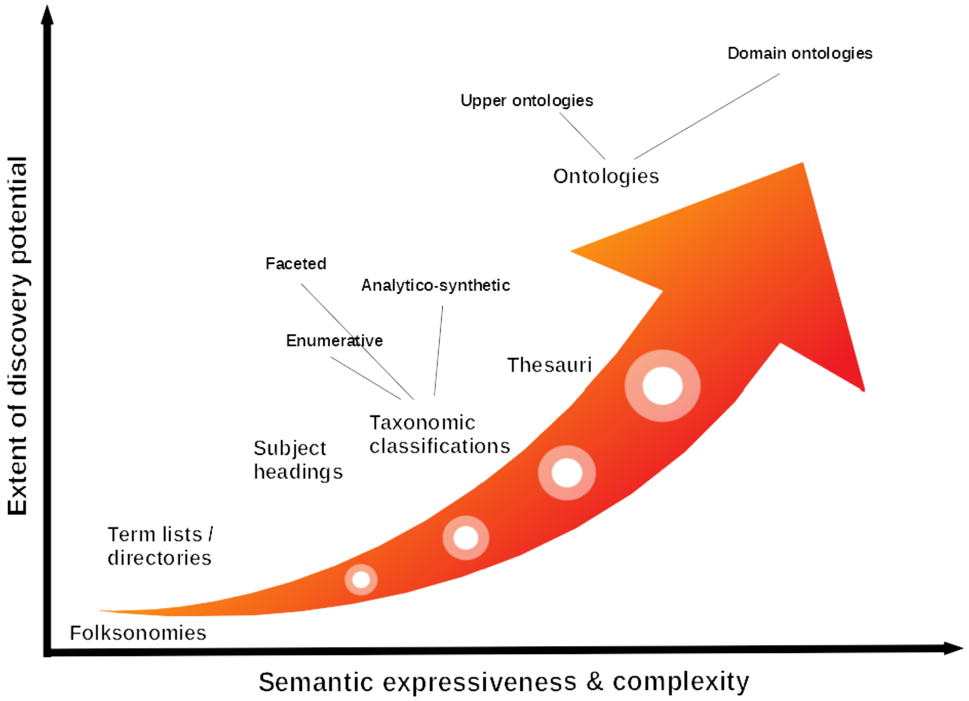


Figure 4.1: Conceptual diagram of the semantic expressiveness & complexity of major KOS types relative to their research discovery potential. Semantic expressiveness & complexity increases on the X axis results in proportional increases on the resource discovery power of the KOS on the Y axis.

Advances in the modelling and publication of KOS data has developed alongside progress in the Semantic Web and Linked Data (or 'Linked Open Data' (LOD)), providing decentralized mechanisms for publishing, sharing, reusing and facilitating access to terminological data. This is perhaps most ably demonstrated by WikiData. Terminology services nevertheless remain a necessary component of KOS interoperability efforts, partly because LOD or Semantic Web solutions cannot in themselves deliver the infrastructure required [174], but also because approaches to the integration of RDF data into local services (e.g. digital libraries, repositories, etc.) is at an insufficient level of maturity, both at the integration level but also at the HCI level [172]. Terminology services can provide the necessary infrastructure and offer opportunities for solving these issues. Their development and deployment within the sub-domains of medical informatics and bioinformatics has been particularly pronounced in recent years. A need to store, query and retrieve data held within complex biomedical terminologies — ranging from gene ontologies to SNOMED to highly specific terminologies pertaining to dermoscopy — requires high levels of efficacy. See for example: [175, 135, 176, 177, 178, 179, 180, 181, 182, 183, 136].

Most of the works assembled for the remainder of this chapter relate to the use of terminology services, the inclusion of RDF data within these services to enable KOS interoperability and terminology mapping (i.e. PW5, PW6, PW7, PW8). PW6 [**133**], PW7 [**184**] and PW8 [**185**] each explore aspects of a specific terminology service, data from which was designed to be embedded within local services. PW5 complements the other works by providing an exercise in the role of RDF in 'e-resource discovery' and is therefore related to PW6, PW7 and PW8 insofar as RDF data modelling was used in the terminology service.

4.4.1 SKOS-based M2M terminology mapping server: PW6 & PW7

The aforementioned terminology service was designed to provide machine access (M2M) via a SOAP web service to terminological data relating to disparate KOS, thereby enabling local services to harness any hierarchical or syntactic relationships for browsing or interactive query expansion (QE) [**133**]. Data served also related to various KOS-to-KOS mappings based on a Dewey Decimal Classification (DDC) switching language, or 'spine', thereby enabling subject-based queries in one discovery service to be translated into the KOS of another [**185**]. Although the terminology service was flexible to satisfy a number of KOS interoperability user cases, the principal use case was as a 'shared service' or node within wider information environments. One such example was the Jisc Information Environment (IE) [186], comprising numerous disparate digital libraries, domain specific or scientific repositories, and cultural heritage platforms each with similarly disparate KOS, thereby demanding federated subject-based searching and browsing for users. A terminology service functioning as a piece of 'shared service infrastructure' and designed to mediate subject-based queries across disparate services is necessary if resource discovery is to be successful [**187**]. The need for terminological mediation in user queries was, for example, envisaged as a critical component of the original Jisc IE architecture [188]. But it should be noted that the use of terminology services to solve this issue also presents opportunities for additional subject-based discovery aids, such as serving data to support hierarchical browsing for users, interactive QE, as well as 'recommended' documents based on related terms or concepts.

PW6 describes the proposed terminology service, focussing largely on the technical approach adopted [**133**]. This service — which was prototyped with colleagues [**189**] — operated in a M2M web service context, using Search/Retrieve Web service (SRW)[10] allowing messages from client to server to be messaged using XML over HTTP via the W3C SOAP protocol[11]. SOAP enables the 'wrapping' of XML messages within an XML envelope. Client queries for terminological data, submitted to the terminology service by SRW, would therefore be returned to the client within a SOAP envelope and modelled using an XML compliant specification, in this case an XML serialization of RDF (RDF/XML). The work outlines the various terminological calls ('server functions') the client can make to the terminology server (e.g. `Get_filtered_set`, `Get_non_DDC_records`), explores the use of the SKOS Mapping Vocabulary Specification (MVS) to modelling KOS-to-KOS mappings, and notes experimental work being undertaken using a geospatial dataset repository. PW7 continues the exposition of PW6 by delivering:

1. A fuller explanation of the server functions available.

2. A demonstration of the way in which terminological data are modelled for messaging in SOAP envelopes.

3. Example searches, and;

4. An analysis of the KOS-to-KOS mapping approach used by the terminology service. The process of 'terminology mapping', which provides the basis for

[10] Search/Retrieve Web Service (SRW) - LOC Standards: `http://www.loc.gov/standards/sru/`

[11] W3C SOAP: `https://www.w3.org/TR/soap/`

KOS-to-KOS terminological data, is discussed in more detail in a later section and is given fuller treatment in PW8 [**185**].

PW7 also highlights the additional role of CLM within the terminology service. An important use case which complements systems offering information landscaping is the notion that there may be circumstances whereby user queries are *collection-level* based rather than *item-level*. PW7 therefore demonstrates a `get_collections` function to assist in the identification of digital collections and/or services by subject(s). This function uses the Dublin Core Collections Application Profile and the IESR Application Profile to model the terminological and collection data returned to clients (see PW7, section 16.5.1).

As noted by PW6 and PW7, the modelling of terminological data were performed using the Simple Knowledge Organization System (SKOS) vocabulary specification. Pure XML specifications for modelling information retrieval thesauri were, and continue to be, available, such as Zthes [190], but they lack the expressiveness to model knowledge structures which display greater semantic sophistication. SKOS was merely an emerging specification at the time of the publication of PW6 and PW7. It had emerged to facilitate the modelling of KOSs for publication on the Semantic Web using RDF (e.g. [191]). The present author contributed case studies to the W3C to aid the development of subsequent versions of SKOS (e.g. [192], [**193**]). SKOS has since become a W3C standard and a key building block of the Semantic Web and LOD [194], and now underpins numerous digital library applications [195, 196, 197, 198, 199, 191, 200, 105, 172], [**201**].

Although largely descriptive in nature, **PW5** could be described as a typical thesis chapter insofar as it establishes the candidate's knowledge, understanding and appreciation for what is a key component of the terminology server described in PW6 and PW7: KOSs and their terminological mappings modelled as RDF in RDF/XML [**201**]. That is to state, it establishes credibility in the author's other works by demonstrating fluency in the concepts underpinning the terminology server, such as in semantically aware metadata, RDF vocabulary specifications and so forth. Of course, this is not to diminish the broader contribution of PW5, which is to provide an exposition of the wider resource discovery opportunities which can arise within semantically aware 'e-resource management' contexts, as well as the potential applications of RDF within digital libraries and repositories. The work itself references many of the other published works presented in this chapter, owing to the use of SKOS RDF within the terminology service described in PW6 and PW7. Even so, it expands on this by assembling examples of RDF applications, such as Dublin Core [202], FOAF [203], RDFa [204], OWL [205] and, of course, SKOS.

For the purposes of a terminology service, however, SKOS presented opportunities for accurately modelling KOS and maintaining their structural and semantic properties. By serializing the SKOS data as RDF/XML, terminological data could be embedded within SOAP envelopes for messaging, extensive examples of which are provided in PW6 and PW7. This, in turn, presented opportunities for the flexible and reliable re-use of terminological data by clients in local systems, as well as novel applications such as displaying terminological results to users as RDF graphs, or better contextualizing results within the semantic structures of KOSs.

An interesting aspect of these technical experiments was that -– at the point these works were published — none of the KOSs used within the terminology server had been modelled in RDF. Their modelling in SKOS and related RDF vocabulary specifications (in RDF/XML) had instead to be created from scratch, largely by the present author, resulting in high levels of efficacy in RDF, graph modelling and knowledge of its technical applications, as evidenced by PW5 [**201**]. Today such modelling would not be required since organizations, including the Library of Congress and the National Institutes of Health (NIH) have since expended great effort to model and expose their terminological data in line with LOD expectations.

The published works discussed thus far in this chapter have presented important conceptual or technical work, with PW6 and PW7 in particular setting out the technical framework for a wider research agenda. This agenda proved influential among those pursuing similar semantic interoperability research, e.g. [206, 207, 208, 209, 210]. Additional aspects of this research agenda were evaluated as part of PW8, which sought to investigate the mapping quality possible across multiple KOSs and how these mapping relationships could be characterized. The nature of PW8 and its contribution is given a more detailed treatment in the following section.

4.4.2 Terminology mapping, equivalence & term disambiguation: PW8

KOS-to-KOS mapping — or more specifically terminology mapping — is a KOS interoperability approach which has been adopted in a wide variety of resource discovery contexts with varying degrees of success [211]. The process of mapping involves imposing a degree of equivalence between the same or similar concepts within different

KOS, including any conceptual and hierarchical features [212]. The terminology service described in PW6 and PW7 derived its terminology mapping data via such a mapping approach.

PW8 provides a detailed exposition of terminology mapping and its difficulties, as well as a review of the research literature in its opening sections, so these will not be reproduced here; suffice to state that direct KOS-to-KOS mapping can be resource intensive owing to its intellectually onerous nature. Mapping research has therefore tended to explore variants of the 'terminology switching model', in which a single KOS is used as an intermediary terminology, against which all other KOSs are mapped [213]. This simplifies the management of multiple terminological mappings and reduces the resource required to maintain direct KOS-to-KOS mappings (see PW8, Fig. 1). By using a common — normally 'universal' -– KOS as the intermediary terminology, it is possible for queries submitted using the terminology of retrieval system *A* to be translated into the terminology of retrieval system *B*.

It is certainly an approach that 'simplifies' and reduces the cost of the terminology mapping; however, it remains a process entailing considerable human resource in order to yield accurate and comprehensive terminological equivalences. It is therefore interesting to note that in the conference slides accompanying PW6 [**133**] the present author noted that investigations were under way to implement 'a more distributed model, including exploring a collaborative model to maintaining and implementing mappings to the spine, such as a wiki-style model for a group of cataloguers, indexers, etc. within the Jisc IE'. While this functionality was ultimately never implemented within the prototype terminology server owing to competing project priorities, it was nevertheless planned and recognized as the only viable, long-term approach to maintaining existing mappings or implementing new ones (in lieu of sufficiently reliable machine automated techniques). It therefore remains a prescient insight into viable distributed concept mapping models, as instantiated more recently with WikiData [156, 152, 157].

The terminology mapping approach upon which the terminology service described in PW6 and PW7 was based used an approach similar to terminology switching, insofar as the Dewey Decimal Classification (DDC) was deployed as the intermediary terminology. As a universal classification, DDC offers extensive treatment of most intellectual concepts and benefits from language-independent analytico-synthetic notation capable of uniquely identifying concepts, a feature considered important for minting concept URIs within the Semantic Web — something which Panzer [214] explored in more detail -– and facilitating multilingual information retrieval. However, the terminology service approach differed to switching by instead using a so-called 'DDC spine' [215]. This spine, described in PW6, became central to functionality surrounding 'concept disambiguation' because, depending on client calls made to the terminology server, relevant terminological data could be messaged back to the client and used to resolve the presence of homographs, enabling end users to refine their query [**133**].

Owing to the semantic, hierarchical, lexical and conceptual differences between KOSs, terminology mapping can only ever provide approximate equivalence between concepts [216]. Characterizing the nature of the imposed mappings therefore becomes important in order to denote the level of equivalence achieved. The nature of this equivalence also has to be accommodated by a terminology service and communicated to clients. **PW8** enumerates the motivation behind capturing this data and serving it as an integral part of an M2M terminology service, including the ability for clients to rank results according to the degree of equivalence with users' preferred terminology, among others.

In the terminology service described in PW6 and PW7 equivalence was characterized using the SKOS Mapping Vocabulary Specification (MVS), a draft specification which has since been incorporated into SKOS proper albeit in a modified form [194] (see example in Fig. 4.2). While the MVS was deployed within the terminology service in lieu of alternatives, it was also acknowledged that the MVS equivalence types were inadequate to accommodate service-scale terminology services [**193**], lacking the necessary specificity to characterize the breadth of equivalences likely to arise across KOSs of varying sophistication. PW8 therefore sought to explore a range of alternative equivalence types for possible use within terminology services [**185**], using Chaplan's mapping types [217] as a starting point to creating a 'generic suite' of equivalence types. By using the prototype terminology server as a testbed it was possible to use Chaplan's more detailed mapping equivalences to investigate to what extent equivalence could be imposed between randomly selected concepts from the Art & Architecture Thesaurus (AAT), MeSH, LCSH and the UNESCO Thesaurus via a DDC spine and whether these equivalence types were a suitable alternative to the MVS in the Semantic Web, or within terminology services.

Whereas other works presented in this chapter presented conceptual or technical work forming part of a wider research agenda, PW8 documented a detailed comparative study which used various methodological controls to improve the validity of its conclusions. By testing mapping quality across a number of disparate KOS types the work contributed to the evolution of terminology service requirements [218], mapping

Figure 4.2: Example of the now deprecated SKOS Mapping Vocabulary Specification (MVS), deployed in conjunction with SKOS Core, and as used within prototype M2M terminology service.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core.rdf#"
xmlns:map="http://www.w3.org/2004/02/skos/mapping#"
xml:base="http://.../concepts.php">
<skos:Concept rdf:about="#363.34">
        <skos:prefLabel xml:lang="zxx">363.34</skos:prefLabel>
        <skos:altLabel xml:lang="en">Disasters</skos:altLabel>
        <skos:inScheme rdf:resource="http://.../schemes/DDC.rdf"/>
        <map:exactMatch>
                <skos:Concept rdf:about="#16117"/>
        </map:exactMatch>
        <map:exactMatch>
                <skos:Concept rdf:about="#16118"/>
        </map:exactMatch>
        <map:narrowMatch>
                <skos:Concept rdf:about="#16119"/>
        </map:narrowMatch>
        <map:narrowMatch>
                <skos:Concept rdf:about="#2256"/>
        </map:narrowMatch>
        <map:narrowMatch>
                <skos:Concept rdf:about="#762"/>
        </map:narrowMatch>
        <map:exactMatch>
                <skos:Concept rdf:about="#2696"/>
        </map:exactMatch>
        <map:exactMatch>
                <skos:Concept rdf:about="#143"/>
        </map:exactMatch>
</skos:Concept>
<skos:Concept rdf:about="#16117">
        <skos:prefLabel xml:lang="en">Disasters</skos:prefLabel>
        <skos:inScheme rdf:resource="http://.../schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#16118">
        <skos:prefLabel xml:lang="en">Emergency management</skos:prefLabel>
        <skos:inScheme rdf:resource="http://.../schemes/LCSH.rdf"/>

        <!-- example truncated -->
```

approaches [219, 220, 198, 221] and ontology mapping research [222]. It was also the most extensive mapping research at the point of publication since most prior work focussed on the mapping issues encountered when equivalence between two single KOSs was being sought (e.g. [223, 224], as opposed to the mapping of *multiple* KOSs across a terminology spine.

What critical reflection highlights is that despite being framed within the context of terminology services, the Semantic Web and the MVS, PW8 makes no proposal for how the identified equivalence match types – which the research concluded needed to be more detailed than the MVS — would be modelled as RDF as part of the wider SKOS specification. The omission of such a proposal appears anomalous since the findings may have enjoyed greater impact if the equivalence types had been operationalized for RDF, perhaps launched as a separate RDF vocabulary for integration by others into their own Semantic Web applications. As noted previously, SKOS was an embryonic specification and the suitability of the MVS was subject to debate [223], [**133**], a factor which originally motivated the research documented in PW8. To not include proposals for how this data could be modelled or re-used alongside other RDF vocabularies therefore appears short-sighted. Furthermore, it may have been excessive for a single published work, but there was an opportunity to pursue a follow-up research study which sought to deploy the proposed equivalence types within the terminology service described in PW6 and PW7 and evaluate their efficacy, again improving the impact of the conclusions of PW8.

Chapter 5

Human-computer interaction (HCI) & curriculum design repositories

This chapter will continue with many of the resource discovery concepts explored in prior chapters but within the novel context of technology-supported curriculum design (or 'tech-supported curriculum design'). This context entails questions around the creation of metadata describing interoperable curriculum data and the use of curriculum design repositories. It is within this environment that **PW9** explores the HCI issues inherent in deploying tech-support curriculum design systems within academic communities [**21**]. In this regard it is worth drawing attention to a series works occupying a similar intellectual space and which were published by the present author in tandem with PW9. These works are not presented as part of this thesis but nevertheless provide important additional narrative around the research contained within PW9. See for example: [**225, 226, 227, 228, 229, 230, 231**].

5.1 Tech-supported curriculum design

Curriculum design is central to the learning and teaching programmes offered by higher education institutions (HEIs). The creation of a curriculum design is about setting out a 'total plan for learning' [232], within which due consideration is given to the intended learning of students, the assessment methods to be drawn upon, and the overall academic rationale underpinning the proposed curriculum [233]. Curriculum design in HEIs is therefore a 'teachable moment' because it remains one of the few instances when academic lecturers concentrate on the planning and structure of their proposed teaching content [234].

The principal motivation behind tech-supported curriculum design is to harness this teachable moment to:

- Promote better curriculum design which, in turn, promotes improved academic quality, pedagogy and ergo student learning impact [235, 236].
- Capture and aggregate structured data about curriculum designs for the purposes of discovery, information management, sharing, improved interoperability across systems, and reuse in the creation of new curricula [**226**].
- Support HEIs in developing curricula which are more responsive to rapidly changing educational requirements, skill needs within industry and specialist curricula for delivery at international branch campuses, or to attract international students within an increasingly globalized HE sector [232, 237].

Those approaches to tech-supported curriculum design that demonstrate the highest levels of technical innovation can support interactive curriculum design systems, within which the 'designer' is supported in the design process. Such support may include system features to ensure designers' adherence to pedagogical best practice, while simultaneously exposing the designer to novel or existing high-impact learning designs [238]. The identification of common curriculum design issues which might otherwise cause academic quality or teaching delivery problems can also be detected [236]. For instance, Kolås & Staupe [239] describe their experiments with a 'design wizard' which promotes the curriculum designer in devising the most appropriate pedagogical approaches for any given learning or assessment method. This design wizard uses various system rules based on the pedagogical evidence-base that exists on the most effective teaching methods, assessment strategies and student engagement tactics to be used, thereby ensuring that that design data are captured appropriately and that key pedagogical quality standards are satisfied.

Despite a number of seminal works in the literature, the research landscape of tech-supported curriculum design remains embryonic, with the most notable experiments initiated in the UK and Australia (e.g. [240, 241, 242, 243, 244, 245], [**225**]. More recently some aspects of tech-supported curriculum design have been relevant to 'instructional design' [246] which, although demonstrating a focus on learning and instruction delivery, increasingly influences the design of curricula [247].

The ability to capture and aggregate curriculum designs, along with their associated (meta)data, is an important motivation of tech-supported curriculum design. It is also the aspect which most relates to the other published works assembled for this thesis.

Information management traditions in HEIs have tended to be unsatisfactory in the area of curriculum design, with many tech-supported projects noting inadequacies in the prior document management, version control, and discovery capacities of institutions [234, 248, 249]. The arrival of tech-supported curriculum design has successfully exposed these inadequacies. In its place it has introduced an information resource ethos into the systems, processes and practices surrounding curriculum design management (where tech-supported approaches have been adopted). Recognizing curriculum designs as constituting information resources, or ‘knowledge assets’, which hold ongoing value and which also require capturing, modelling, describing, sharing and reusing has been central to this ethos [**226**]. Even within projects which have explored socially orientated Web 2.0 inspired approaches, for example Cloudworks[1] at the Open University, an emphasis has been to ensure designs are captured, shared, reused and their value maximized [238].

The prospect of modelling and describing curriculum designs has motivated thinking on how best such knowledge assets should be captured [250, 239]. By capturing structured data about designs there are opportunities to ensure designs can be more easily discovered, not only within local institutional contexts, but across distributed environments where multiple curriculum design systems or repositories may co-exist. The sharing of curriculum designs across institutions and educational sectors, facilitated via standardized interoperable metadata schema, is therefore a distinguishing feature of many tech-support curriculum design approaches [251].

5.2 XCRI

The eXchanging Course Related Information (XCRI) data model, accompanied by an XML schema, has provided a basis for interoperability between disparate systems, as well as specifying core data elements for curriculum design systems and repositories [2]. Borrowing from a number of existing schema, including Dublin Core and Metadata for Learning Opportunities (MLO) [252, 253], XCRI has the potential to describe ‘course related information’, as per the example created in Fig. 5.1. However, owing to the vast nature of curriculum information, most examples in the literature have focused primarily on a smaller application profile of XCRI known as XCRI-CAP

[1] Cloudworks: `https://cloudworks.open.ac.uk/`

(Course Advertising Profile) (Fig. 5.2 – XCRI-CAP UML model). XCRI-CAP enables institutions to share data pertaining to curricula with course aggregators and discovery systems [254, 255, 256, 245]. To this end experiments were conducted using a central data hub for facilitate the exchange of XCRI data, known as the XCRI eXchange Platform (XXP) [257], with various 'value added' services build on top of XCRI data. More recently, data are driving websites such as Prospects[2], a course comparison service.

Improved semantic interpretation of curriculum data has nevertheless been established, with sector-wide recommendations to model XCRI in RDF [258] and with XCRI mapped to more semantically aware schema, such as Schema.org [259]. Disappointingly, work on an official RDF vocabulary remains unfinished [2]; but work undertaken as part of the LUCI project by Ouseena & Hyeonsook created an RDF schema for XCRI and explored ways in which curriculum data could be exposed as Linked Data [260]. Suffice to state, XCRI provides a foundation set of curriculum design metadata which can then be extended using other vocabulary specifications — and which can be harnessed by institutions to build curriculum design systems, or in the case of PW9, a prototype curriculum design system and repository [**21**].

5.3 Prototype curriculum design repository & cognitive load theory: PW9

The prototype system described in **PW9** was developed under the auspices of a wider tech-supported curriculum design project[3], which researched innovative technological approaches to curriculum design in order to exploit the 'teachable moment' described earlier in this chapter. This entailed developing technology which could better support academics in designing improved curricula, thereby leading to superior educational outcomes, as well as the discovery, aggregation and improved management of curriculum design data [**227, 228, 231**]. This latter aspect was notable for the creation of a design repository, from which designs could be discovered, reused (or 'cloned') for the purposes of creating new designs (see Fig. 5.3 & 5.4). Exemplar designs were also highlighted to users by the system to inspire users' innovation in their own design practices [**226**]. Designs were described according to a bespoke XML schema, with the capability of a subset of elements to be mapped to XCRI for discovery and interoperability [**225**].

[2] Prospects: `https://www.prospects.ac.uk/`

[3] Principles in Patterns (PiP): `https://www.principlesinpatterns.ac.uk/`

Figure 5.1: XCRI example created for MSc Digital Health Systems at the Department of Computer & Information Sciences, University of Strathclyde.

```
<?xml version="1.0" encoding="UTF-8"?>
<catalog
    xmlns:"http://xcri.org/profiles/1.2/catalog"
    xmlns:xcriTerms="http://xcri.org/profiles/1.2/catalog/terms"
    xmlns:credit="http://purl.org/net/cm"
    xmlns:dcterms="http://purl.org/dc/terms/"
    xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:mlo="http://purl.org/net/mlo"
    xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos"
    xmlns:xhtml="http://www.w3.org/1999/xhtml"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns:courseDataProgramme="http://xcri.co.uk"
    xsi:schemaLocation="
    http://xcri.org/profiles/1.2/catalog http://schema.prospects.ac.uk/xcri/xcri_cap_1_2.xsd
    http://www.w3.org/2003/01/geo/wgs84_pos http://www.craighawker.co.uk/xcri/validation/xsds/geo.xsd
    http://xcri.org/profiles/1.2/catalog/terms http://schema.prospects.ac.uk/xcri/xcri_cap_terms_1_2.xsd
    http://purl.org/net/mlo http://schema.prospects.ac.uk/xcri/mlo/mlo_xcri_profile.xsd
    http://xcri.co.uk http://schema.prospects.ac.uk/xcri/coursedataprogramme.xsd"
    generated="2019-10-14T15:58:23">
    <dc:contributor>Macgregor, George</dc:contributor>
    <dc:description>University of Strathclyde. This data is released under Open Government Licence (OGL) Version
        3.0 - http://www.nationalarchives.gov.uk/doc/open-government-licence/</dc:description>
        <provider>
            <mlo:hasPart>Department of Computer and Information Sciences</mlo:hasPart>
            <mlo:hasPart>Department of Physics</mlo:hasPart>
            <mlo:hasPart>Department of Pure and Applied Chemistry</mlo:hasPart>
            <dc:description>A place of useful learning</dc:description>
                <dc:identifier>https://www.strath.ac.uk/</dc:identifier>
            <dc:identifier xsi:type="courseDataProgramme:ukprn">10099999</dc:identifier>
            <dc:title>University of Strathclyde</dc:title>
                <mlo:url>https://www.strath.ac.uk/studywithus/postgraduatetaught/</mlo:url>
            <course>
                <mlo:isPartOf>Department of Computer and Information Sciences</mlo:isPartOf>
<dc:description>MSc Digital Health Systems - Become a leader in the field of health and care IT. Learn how to
     manage and analyse data collected from personal device and large-scale health and care systems. Develop
     software development and management skills to support planning and delivery of better care systems. Partial
     accreditation by the British Computer Society.</dc:description>
<dc:description xsi:type="xcriTerms:specialFeature">Work with the multidisciplinary Digital Health and Wellness
     Research group based in computer and information science. This group has been involved in several major
     collaborative research and development projects and evaluations within the UK and internationally. The group
      were lead investigators in the evaluation of a 37GBP million Innovate UK dallas programme to deploy
     assistive digital health and wellness technologies at scale across the UK.</dc:description>
                        <dc:identifier>https://www.strath.ac.uk/courses/postgraduatetaught/digitalhealthsystems
                             /</dc:identifier>
                        <dc:identifier xsi:type="courseDataProgramme:internalID">PG064</dc:identifier>
                        <dc:subject xsi:type="courseDataProgramme:JACS3" identifier="I110">Computer architectures
                             and operating systems</dc:subject>
                        <dc:subject xsi:type="courseDataProgramme:JACS3" identifier="I500">Health informatics</dc
                             :subject>
                <dc:subject xsi:type="courseDataProgramme:JACS3" identifier="I510">Health technologies</dc:
                     subject>
                <dc:title>Digital Health Systems</dc:title>
                        <dc:type xsi:type="courseDataProgramme:courseTypeGeneral" courseDataProgramme:identifier
                             ="PG">Postgraduate</dc:type>
                        <dc:type xsi:type="mlo:RTCourseTypeFlag" mlo:RT-identifier="T">Taught</dc:type>
                        <mlo:url>https://www.strath.ac.uk/courses/postgraduatetaught/digitalhealthsystems/</mlo:
                             url>
<abstract>This professional masters degree in business and marketing is an exciting route for anyone working in
     any field.</abstract>
                        <applicationProcedure href="http://www.poppleton.ac.uk/postgraduate/courses/how-to-apply
                             /"/>
<mlo:assessment>Taught modules are assessed using a combination of individual projects, group projects and final
     exams. The project is assessed on the quality of the project report (ie Master thesis). An overall minimum
     of 50% across all assessed classes and report is required in order to be awarded the Master in Digital
     Health Systems.</mlo:assessment>
<learningOutcome>Students will learn about the lifecycle of designing, developing and evaluating health
     technologies from mhealth and novel personal health and wellness devices (eg mobile apps, wearables) to
     ehealth and larger scale hospital and community based IT systems (eg electronic health records). Students
     will understand agile participatory and co-design approaches for delivering health and care IT solution. </
     learningOutcome>
                <mlo:objective>Successful award of MSc</mlo:objective>
                <mlo:prerequisite>First degree in any subject.</mlo:prerequisite>
                <regulations href="www.strath.ac.uk/sees/educationenhancement/qualityassurance/
                     universityregulations/"/>
                <mlo:qualification>
                    <dc:identifier>MBA001</dc:identifier>
                    <dc:title>MSc Digital Health Systems</dc:title>
                    <abbr>MBA</abbr>
                    <dc:description>Master of Science</dc:description>
                    <dcterms:educationLevel>Postgraduate</dcterms:educationLevel>
                    <mlo:url>https://www.strath.ac.uk/courses/postgraduatetaught/digitalhealthsystems/</mlo:url>
                    <awardedBy>University of Strathclyde</awardedBy>
                </mlo:qualification>

<!-- example truncated -->
```

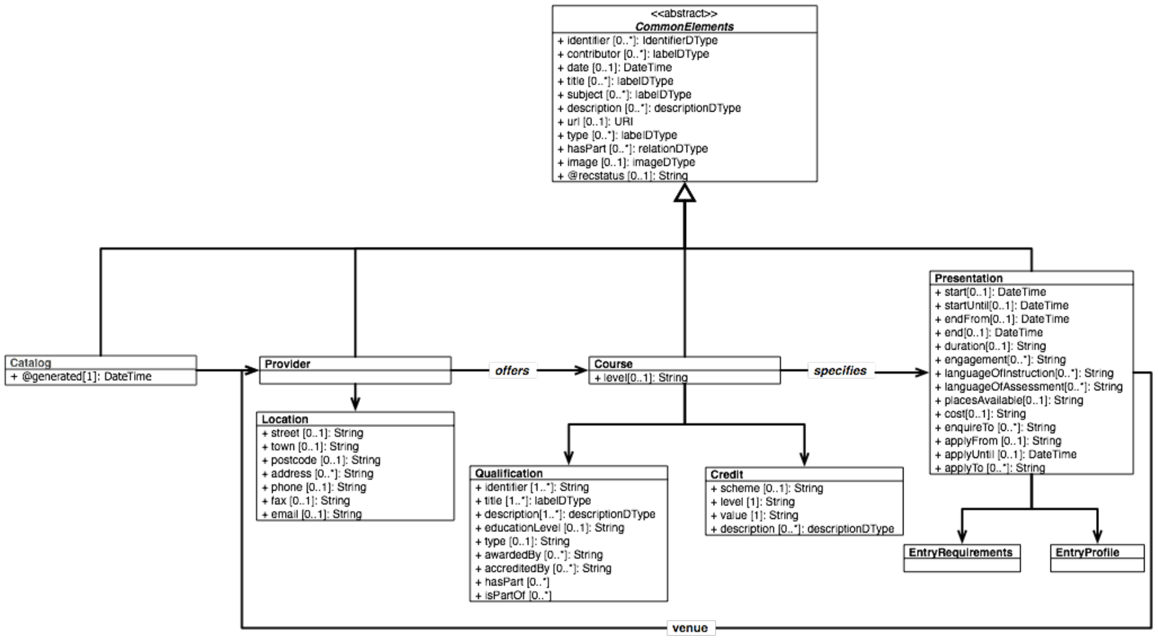
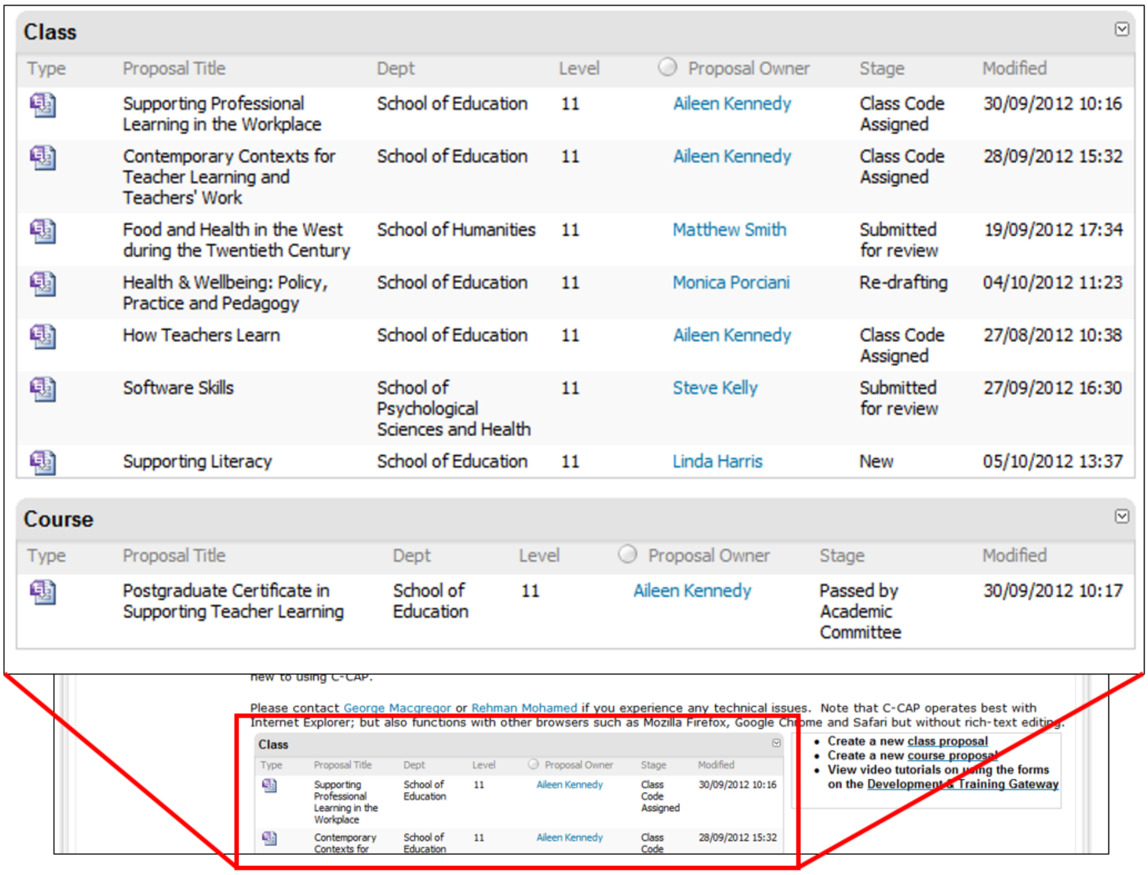

<<abstract>>
CommonElements
+ identifier [0..*]: IdentifierDType
+ contributor [0..*]: labelDType
+ date [0..1]: DateTime
+ title [0..*]: labelDType
+ subject [0..*]: labelDType
+ description [0..*]: descriptionDType
+ url [0..1]: URI
+ type [0..*]: labelDType
+ hasPart [0..*]: relationDType
+ image [0..1]: imageDType
+ @recstatus [0..1]: String
Catalog
+ @generated[1]: DateTime
Provider
offers
Course
+ level[0..1]: String
specifies
Presentation
+ start[0..1]: DateTime
+ startUntil[0..1]: DateTime
+ endFrom[0..1]: DateTime
+ end[0..1]: DateTime
+ duration[0..1]: String
+ engagement[0..*]: String
+ languageOfInstruction[0..*]: String
+ languageOfAssessment[0..*]: String
+ placesAvailable[0..1]: String
+ cost[0..1]: String
+ enquireTo [0..*]: String
+ applyFrom [0..1]: String
+ applyUntil [0..1]: DateTime
+ applyTo [0..*]: String
Location
+ street [0..1]: String
+ town [0..1]: String
+ postcode [0..1]: String
+ address [0..*]: String
+ phone [0..1]: String
+ fax [0..1]: String
+ email [0..1]: String
Qualification
+ identifier [1..*]: String
+ title [1..*]: labelDType
+ description[1..*]: descriptionDType
+ educationLevel [0..1]: String
+ type [0..1]: String
+ awardedBy [0..*]: String
+ accreditedBy [0..*]: String
+ hasPart [0..*]
+ isPartOf [0..*]
Credit
+ scheme [0..1]: String
+ level [1]: String
+ value [1]: String
+ description [0..*]: descriptionDType
EntryRequirements
EntryProfile
venue

Figure 5.2: XCRI-CAP, UML model [2].

As innovative as these systems are (including the prototype described in PW9), they are curious insofar as many of the data elements comprising a data model such as XCRI have to be created by the academic user in order for the design to be accurately captured. The system itself may capture technical aspects of the design automatically, such as technical metadata or inferred data properties based on users' academic context (e.g. subject/discipline, institutional affiliation, teaching remit, etc.). Previous designs may also be cloned for reuse. But metadata pertaining to the teaching delivery, learning outcomes, assessment strategy, curriculum structure and so forth are created by the intended teacher of that design [248]. In other words, academic users are engaging with a familiar task albeit demonstrating high levels of 'intrinsic' cognitive load [261, 262], while also being simultaneously exposed to high levels of 'extraneous' cognitive load [263, 264] as the user attempts to complete their task using unfamiliar or novel technology. This highlights the system interaction issues which can arise when users demonstrate high levels of 'domain expertise' but lower 'task-based expertise', or vice versa.

HCI experiments have revealed that in certain conditions domain expertise can be a determinant of whether related information tasks are completed with satisfactory efficacy; however, that level of efficacy is directly influenced by the precise level of domain expertise and system knowledge (i.e. task-based expertise). For example, studies exploring these variables across information searching and interaction behaviour have concluded that the effect of domain expertise can be limited [265, 266]. In



Class

Type	Proposal Title	Dept	Level	Proposal Owner	Stage	Modified
	Supporting Professional Learning in the Workplace	School of Education	11	Aileen Kennedy	Class Code Assigned	30/09/2012 10:16
	Contemporary Contexts for Teacher Learning and Teachers' Work	School of Education	11	Aileen Kennedy	Class Code Assigned	28/09/2012 15:32
	Food and Health in the West during the Twentieth Century	School of Humanities	11	Matthew Smith	Submitted for review	19/09/2012 17:34
	Health & Wellbeing: Policy, Practice and Pedagogy	School of Education	11	Monica Porciani	Re-drafting	04/10/2012 11:23
	How Teachers Learn	School of Education	11	Aileen Kennedy	Class Code Assigned	27/08/2012 10:38
	Software Skills	School of Psychological Sciences and Health	11	Steve Kelly	Submitted for review	27/09/2012 16:30
	Supporting Literacy	School of Education	11	Linda Harris	New	05/10/2012 13:37

Course

Type	Proposal Title	Dept	Level	Proposal Owner	Stage	Modified
	Postgraduate Certificate in Supporting Teacher Learning	School of Education	11	Aileen Kennedy	Passed by Academic Committee	30/09/2012 10:17

new to using C-CAP.

Please contact George Macgregor or Rehman Mohamed if you experience any technical issues. Note that C-CAP operates best with Internet Explorer; but also functions with other browsers such as Mozilla Firefox, Google Chrome and Safari but without rich-text editing.

Class

Type	Proposal Title	Dept	Level	Proposal Owner	Stage	Modified
	Supporting Professional Learning in the Workplace	School of Education	11	Aileen Kennedy	Class Code Assigned	30/09/2012 10:16
	Contemporary Contexts for	School of Education	11	Aileen Kennedy	Class Code	28/09/2012 15:32

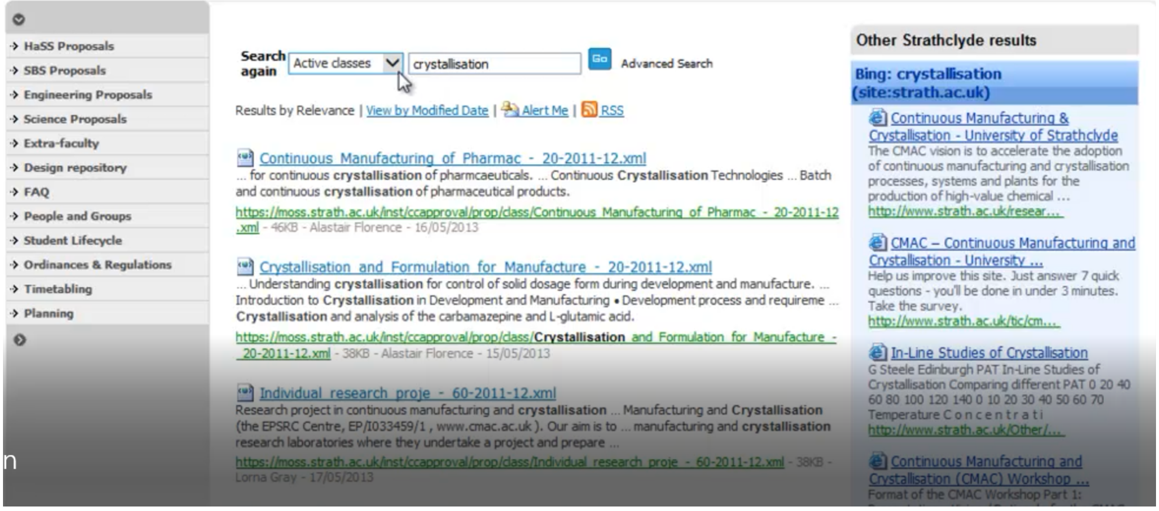
- Create a new class proposal
- Create a new course proposal
- View video tutorials on using the forms on the Development & Training Gateway

Figure 5.3: Creating and managing curriculum designs within the prototype design repository.

others, domain expertise has been found to influence web searching task completion. White et al. have, for instance, demonstrated models for characterizing and predicting expertise levels thereby allowing system interactions that are more responsive to user search needs [267]. But outside of information retrieval or information seeking behaviour, a more problematic relationship between domain and task knowledge is observable (e.g. [268, 269].

In addition, systems such as the prototype in PW9 are seeking to generate more detailed curriculum designs and, in essence, creating the kind of metadata which might normally be expected of a data professional. As PW9 explains in section 2.2, such systems therefore seek to minimize extraneous cognitive load since, 'Systems that expose users to high levels of extraneous cognitive load as a result of poor system design and usability have been shown to erode human cognitive processing. This generally manifests itself in a measurable decline in task performance, inefficiency in task completion, increased error rates and user frustration' [**21**].

The present author was responsible for the evaluation of the prototype system,



Continuous Manufacturing of Pharmac - 20-2011-12.xml
... for continuous **crystallisation** of pharmcaeuticals. ... Continuous **Crystallisation** Technologies ... Batch and continuous **crystallisation** of pharmaceutical products.
https://moss.strath.ac.uk/inst/ccapproval/prop/class/Continuous_Manufacturing_of_Pharmac_-_20-2011-12.xml - 46KB - Alastair Florence - 16/05/2013

Crystallisation and Formulation for Manufacture - 20-2011-12.xml
... Understanding **crystallisation** for control of solid dosage form during development and manufacture. ... Introduction to **Crystallisation** in Development and Manufacturing • Development process and requireme ... **Crystallisation** and analysis of the carbamazepine and L-glutamic acid.
https://moss.strath.ac.uk/inst/ccapproval/prop/class/**Crystallisation**_and_Formulation_for_Manufacture_-_20-2011-12.xml - 38KB - Alastair Florence - 15/05/2013

Individual research proje - 60-2011-12.xml
Research project in continuous manufacturing and **crystallisation** ... Manufacturing and **Crystallisation** (the EPSRC Centre, EP/I033459/1 , www.cmac.ac.uk). Our aim is to ... manufacturing and **crystallisation** research laboratories where they undertake a project and prepare ...
https://moss.strath.ac.uk/inst/ccapproval/prop/class/Individual_research_proje_-_60-2011-12.xml - 38KB - Lorna Gray - 17/05/2013

Bing: crystallisation (site:strath.ac.uk)

Continuous Manufacturing & Crystallisation - University of Strathclyde
The CMAC vision is to accelerate the adoption of continuous manufacturing and crystallisation processes, systems and plants for the production of high-value chemical ...
http://www.strath.ac.uk/resear...

CMAC – Continuous Manufacturing and Crystallisation - University ...
Help us improve this site. Just answer 7 quick questions - you'll be done in under 3 minutes. Take the survey.
http://www.strath.ac.uk/tic/cm...

In-Line Studies of Crystallisation
G Steele Edinburgh PAT In-Line Studies of Crystallisation Comparing different PAT 0 20 40 60 80 100 120 140 0 10 20 30 40 50 60 70 Temperature Concentrati
http://www.strath.ac.uk/Other/...

Continuous Manufacturing and Crystallisation (CMAC) Workshop ...
Format of the CMAC Workshop Part 1:

Figure 5.4: Discovering designs for possible reuse (or 'cloning') in the creation of new curriculum designs.

which resulted in an extensive programme of disparate evaluative threads, the foci of which included HCI, data modelling, metadata interoperability, business process re-engineering and curriculum design [**228, 229**]. **PW9** documents a detailed user study of the prototype tech-supported curriculum design system [**21**] and represents the outcome of one of these evaluative threads. The study was essentially designed to evaluate the system within a 'real user' context and measure the system's capacity to support participants in the creation of curriculum designs. The work consequently documents the intersection between:

1. tech-supported curriculum design and the bureaucratic processes underpinning it,

2. the metadata generated to drive the system and the designs deposited in the repository; and, most notably,

3. aspects of HCI, specifically human-centred design factors; since for (1) and (2) to be successful, it is necessary to ensure the minimization of users' cognitive load.

Of the works assembled at this point for the thesis, PW9 displays the highest level of methodological sophistication and rigour. The procedure employs a mix of quantitative and qualitative methods and demonstrates a level of enquiry unseen in prior works. Protocol analysis, or the 'think aloud' technique [270, 271], was utilized with study participants and productively combined with stimulated recall [272, 270, 273] to generate a rich foundation of qualitative data. The nature of this data included screen capture recordings of both participants' system interactions (visual data) and audio recordings of their 'think aloud' protocols (audio data), as well as audio recordings of the stimulated recall sessions. All of this data was subjected to exhaustive content analysis, coding and further enquiry.

Participant interactions with the prototype system were inserted between especially customized questionnaire instruments, deployed pre- and post- the interaction session. The pre-session instrument used features of Murphy et al.'s Computer Self-Efficacy (CSE) scale [274], which incorporated more recent modifications proposed by the literature in order to benchmark the IT efficacy of study participants [275]. A customized version of the Brooke's System Usability Scale (SUS) was the post-session instrument [276, 277]. The SUS was modified as per the findings of Bangor et al. [278] and by Finstad [279], and was supplemented by the Adjective Rating Statement (ARS).

The SUS is an instrument which has — and continues to be -– developed, deployed and validated by HCI research, ranging from topics such as information retrieval and resource discovery (e.g. [280, 281, 282, 283]), to more generic aspects of HCI (e.g. [278, 279, 284, 285, 286, 287, 288]. SUS has also been successfully translated into languages other than English [289, 290].

The lack of prior work in this area, or even any conceptual understanding about users' interactions with systems like or similar to the prototype, therefore necessitated rich data gathering via numerous research instruments in order to support the triangulation of findings. Although PW9 represents a robust evaluative study underpinned by methodological rigour, it is evident that insufficient control over — or measurement of — extraneous cognitive load was attempted, despite this variable representing an important motivation in assessing the efficacy of the prototype system. The influence of load on the participants was instead determined almost entirely through qualitative data, collected from the protocol analyses, both visual and audio data (e.g. heuristic behaviour, participant comments during the protocol, etc.). Such qualitative data are not without merit and can provide insights about cognition that quantitative techniques cannot [291, 292, 293, 294]; but the protocol analyses were not supplemented by any quantitative instruments, nor was this limitation highlighted in the study caveats. The post-session SUS instrument gathered metrics from which aspects of extraneous cognitive load were inferred but at an insufficient level of specificity to draw conclusions about this aspect of participants' experiences.

Upon reflection the decision not to include a quantitative measure appears odd in retrospect because such measures are widely documented in the literature [295, 296, 270, 282] and some research instruments are considered 'standard' within any HCI research (e.g. [297]). It is nevertheless the case that the quantitative measurement of cognitive load during any HCI tasks can be problematic. Most standard techniques are questionnaire instruments [270], which can be clumsy and intrusive to participants' task performance since to gather optimal data they ideally need to be administered during task performance which, for obvious reasons, is too disruptive. As such they are normally administered post-task, at which point participants' recollections of mental exertion may be incorrectly recalled and ergo incorrectly reported, giving rise to data validity concerns [282].

The subjective nature of self-reporting instruments over those based on directly observable and quantifiable characteristics remains a contested issue in the literature, with more recent works exploring how cognitive load can be quantified during task completion in unobtrusive ways; for example via speech-based load measurement, derived from protocol data, and used to map speech features (e.g. rate of pauses, voice pitch, etc.) to users' mental exertion [298] -– and the mapping of users' eye movements via eye-tracking techniques to quantify levels of cognitive load relative to website complexity [299]. The use of mixed methods, as in PW9, is an important way of combating the subjective reporting of cognitive load, especially in lieu of speech analysis or eye-tracking — and given the myriad of instruments already deployed in PW9, there were opportunities to further triangulate data gathered by using, for instance, the seminal NASA Task Load Index (TLI) instrument, alongside the protocol analyses used [300, 297]. This is not to state that there were no quantitative measures — recall that a modified version of Brooke's SUS instrument [276, 277] alongside the ARS [278] was administered — more that greater attention could have been paid to this aspect of the study had dedicated measures been deployed in tandem.

Despite the aforementioned potential shortcomings in methodological design, PW9 nevertheless demonstrates a generally robust contribution to the research area, especially in its exhaustive use of qualitative data to generate a hierarchical coding framework capable of eliciting significant observations about participants' acceptance of the prototype system. The conclusion that the prototype was 'positively received' by participants — as triangulated across data gathered from a number of instruments — was negated by the study's failure to model participants' 'real world' tasks. This shortcoming resulted in a failure of the prototype system to deliver on one of its core objectives — or at least the methodology failed to detect it: reflection or inventiveness in participants' design creation process, something which was hypothesized would result in the creation of superior designs capable of greater learning impact among students. PW9 reports that the failure here was the artificial nature of the task participants were set and proposes solutions in any future research (see Section 4 — Conclusion). However, it is difficult to contemplate the recruitment of participants willing to generate 'design diaries' for designs created from scratch, something which engenders high levels of intrinsic cognitive load; or to employ user captured data as proposed by some in the literature [301] for such an involved task as curriculum design. More generally, designing evaluative tasks which accurately model users' real world information tasks and which demonstrate satisfactory external — as well as internal — validity remain problematic because a level of artificiality is inevitably introduced into a lab setting [270]. To this extent, it is possible to state that the artificiality of the user task in PW9 was also due to the lab setting and was no different to most other user studies.

As an example of tech-supported curriculum design to support the capture of designs for discovery, management and reuse, the prototype system demonstrated success. Systematizing designs, their content and their metadata immediately adds to their value by rendering them more useful to others [302, 303, 304], a familiar knowledge management ethos which often has inconsistent adoption in general HEI operations [232, 305, 306, 307].

The exposure of designs via XCRI compliant metadata for potential aggregation through platforms such as XXP reinforces a thematic link pertaining to interoperability and its relationship to discovery, as in previous chapters of this thesis; although, an important aspect which sits outside the scope of PW9 and its associated works could be said to be the failure of XCRI to migrate to an RDF data model. As a limitation this has become more obvious since the publication of PW9, and calls into question the scalability of XCRI data aggregation. Schema objects within XCRI which are capable of being referenced by URI are instead described as literals and obvious weaknesses in semantic interoperability are therefore likely to result. This may explain recent experiments with XCRI mappings to schema.org in lieu of a comprehensive RDF XCRI vocabulary specification [308].

The methodological approach adopted in PW9 was necessarily multi-pronged owing to the complexity but also the novelty of the research area. Tech-supported curriculum design remains an embryonic area of study [309, 310], of which the use of technology to support the design and capture of curricula — of the type akin to the prototype system in PW9 — is an even greater subset. For example, since PW9 was published in 2012, there have been just four works which have evaluated the success of similar systems, or sought to refine understanding on how designs can be managed for the purposes of discovery or reuse [242, 311, 312, 313].

It is also significant to note that despite conceptual works in the late 1970s on the potential application of 'systems thinking' and its technological application within the curriculum design process [314, 315, 316], only a few examples of (what became known as) 'tech-supported curriculum design' have been observed in the literature prior to circa 2010 (at which point activity started to grow). See, for example: [317, 239, 318, 236, 319]. To this extent, PW9 is both a novel and significant contribution to our understanding of how users interact with tech-supported curriculum design tools since the prototype system was more mature than those described in the extant literature and included many innovative features, such as the storage of designs in a repository for discovery and cloning, the metadata modelling and information management capabilities, and the academic quality management features available [**21, 226, 320**]. Given the relative paucity of published literature in this area, it is regrettable that PW9 was not developed or repurposed for publication in the academic literature where it may have enjoyed greater impact. Instead, PW9 remains one of many technical outputs from the Principles in Patterns (PiP) project, most of which were published in the open scholarly commons as reports or discussion papers [**320, 228, 231, 227, 21, 230**].

Chapter 6

Open science: resource discovery & open repositories

The final works assembled for this thesis, and for consideration in this chapter, are **PW10** and **PW11**. These related works explore the evolution of a particular research concept within the setting of open science, specifically the evaluation of system optimizations designed to effect improved resource discovery. These works are a crystallization of various research themes already explored in Chapters 3, 4 and 5 and, as published works, these publications unify knowledge and understanding from previous phases of the present author's career to deliver a distinctive contribution to open science and the operation of open scholarly infrastructure.

The works focus on the efficacy of open repositories as nodes within open scholarly communications infrastructure and as discovery mechanisms for open content, particularly scholarly content belonging to the knowledge commons. This chapter will present the research motivations surrounding PW10 and PW11 but also explore the context [**321, 322**], which necessitates consideration of existing repository support for resource discovery as well as delineating 'open repositories' for the purposes of this chapter.

6.1 Open repositories, open science & the knowledge commons

As examples of software used to manage and deliver digital content, definitions of 'repositories' in this thesis have thus far been generic. However, open repositories -– as a component in open scholarly communications infrastructure — assume a more precise meaning, a meaning which has gained currency outside of pure scholarship

[323, 324, 325, 326]. Definitions of repositories can vary in the literature (e.g. [327, 328, 329, 330, 331] but most agree that they typically deliver and support:

1. Heterogeneous and open digital content, often of a scholarly nature, such as (non)peer-reviewed research texts, research datasets, theses and grey literature [328, 331], **[332]**; although increasingly delivering multimedia assets, digitized collections and so forth [333].

2. The management of digital assets over time, normally using open source technologies, in order to ensure the identification, persistence, digital preservation and curation of digital objects. Such management is essential to the maintenance of unique digital collections [334] but an increasingly important instrument in maintaining the 'digital scholarly record', an issue which is being confronted in instances where less stable publishing technologies have been deployed [335, 336, 337].

3. Community-driven or community-focused management of digital content. Repositories will typically serve a community of users or be operated by a specific community of practice (e.g. arXiv.org[1], for mathematical sciences). This community, whether subject-based or institutional, determines what should be deposited in the repository but is often simultaneously a contributor to the content deposited and exposed by that repository [338]. In other words, community members are frequently both authors and copyright owners, especially within the context of scholarly open repositories.

4. The improved exposure, visibility and discovery potential of open digital content. Perhaps most importantly given the scope of this chapter, repositories are designed to expose digital content and promote discovery of that content [339, 340, 341], thereby also generating scholarly impact for open research content [342, 343, 344]. This requires system support for a variety of technical standards and protocols designed to promote interoperability with search agents and to facilitate participation in a distributed global repository network, or 'cooperative' [79]. These technical expectations are, as in the discussions of Chapter 3, a condition of participating meaningfully the global repository cooperative, which stipulates the technologies, syntax and semantic expectations of the participating services. As well shall see, such networks are central to content aggregation,

[1] arXiv.org: `https://arxiv.org/`

> data mining, the creation of new discovery tools, and the production of new or unexpected knowledge derived (often computationally) from repository content.

The content that open repositories in their various permutations provide is a contribution to the burgeoning 'knowledge commons' [345, 346], or in some circumstances the 'digital commons' — although it should be noted that characterizations of the latter are frequently confined to free and open-source software (FOSS) and remain the subject of continued intellectual debate [347].

The concept of the 'commons' within the realm of information and computing is not new (e.g. [348]), but since the emergence of the Web obvious opportunities have arisen, most notably through the Creative Commons licensing project[2], upon which much of the content stored in repositories depends. As a contribution to the commons, information, data, and digital content demonstrates a degree of collective or community ownership (as per 3 above), with its reuse and dissemination encouraged. The peculiar aspect of this arrangement is that digital content is 'non-subtractible' insofar as multiple users can access the same content with zero effect on their quantity or quality [349].

The umbrella concept of 'open science', or 'open research', can be a broad one. It includes many related sub-concepts surrounding open scholarly communications infrastructure, Open Access to research content, open data, open peer-review, open annotation and so forth and is well documented in the literature [350, 351, 352, 353]. Suffice to state, open science is a logical extension of the knowledge commons, by ensuring scholarly findings are disseminated more rapidly thereby accelerating scientific achievement for the benefit of humanity [354]. With openness new opportunities are presented to reuse and add value to existing findings, for example through replication or reuse of datasets for unexpected applications. With openness in content and scholarly infrastructure, digital research content can be text and data mined (TDM), enabling the extraction of implicit knowledge contained in a growing corpus of tens of millions of full-text documents [355, 356, 357, 358], well beyond the corpora used by Swanson's pioneering TDM work in the late 1980s and early 1990s [359, 360].

[2] Creative Commons licensing project: `https://creativecommons.org/`

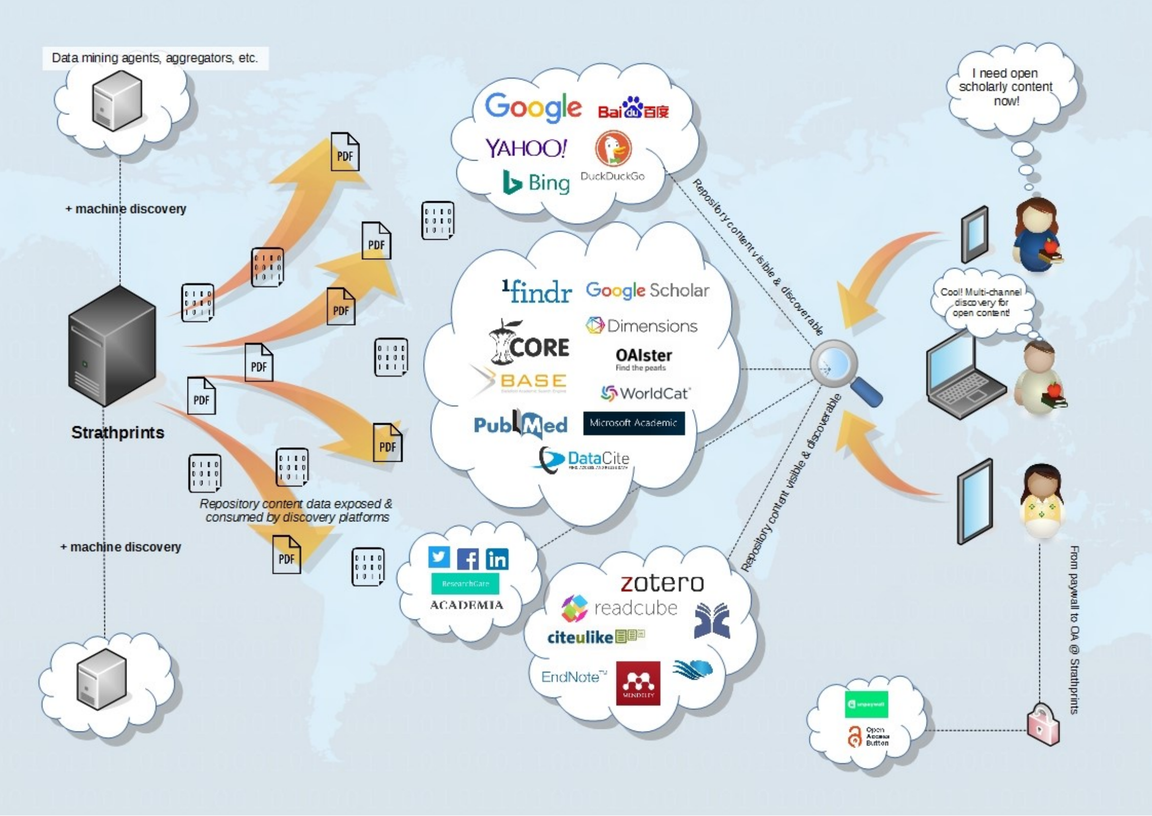

Data mining agents, aggregators, etc.
+ machine discovery
Strathprints
+ machine discovery
PDF
Repository content data exposed & consumed by discovery platforms
Google
Baidu百度
YAHOO!
Bing
DuckDuckGo
1findr
Google Scholar
CORE
Dimensions
OAIster
Find the pearls
BASE
WorldCat
PubMed
Microsoft Academic
DataCite
ResearchGate
ACADEMIA
zotero
readcube
citeulike
EndNote
MENDELEY
unpaywall
Open Access Button
Repository content visible & discoverable
Repository content visible & discoverable
I need open scholarly content now!
Cool! Multi-channel discovery for open content!
From paywall to OA @ Strathprints

Figure 6.1: The technical expectations of participating in the 'repository cooperative' means that repositories demonstrate high levels of interoperability with a disparate array of discovery tools and systems, as diagrammed here.

6.2 Technical expectations of the repository cooperative

As per 4 above, participation in the global open repository network, or cooperative, necessitates the fulfilment of certain technical expectations, the majority of which promote interoperability between distributed repositories but also repository interoperability with search agents and content aggregators, thereby supporting the exposure of repository content. Most open source repository platforms fulfil these expectations and, taken together, fulfilment means that repository content is by default well-placed to be exposed and consumed by a wide variety of resource discovery services, as diagrammed in Fig. 6.1.

6.2.1 OAI-PMH

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) is a principal building block of the open repository network and provides a machine interface to repository content [361, 362]. Repository data are provided as XML according to any XML-based schema (e.g. MODS, MARCXML, METS) or metadata application profiles (e.g. EThOS, OpenAIRE), although at a minimum DC XML must be provided. Data can be queried according to a set of OAI-PMH specified queries (or 'verbs') and allows data exchange across repositories but also the harvesting of repository content by aggregation engines, such as CORE[3] or BASE[4]. Aggregation and exposure of content by services such as CORE not only facilitates additional visibility for repository content, but also allows novel tools to be constructed on top of this data and content. For example, APIs to enable TDM across aggregated content [363, 358], alternative bibliometric approaches [364], scholarly recommendation engines [365], or paywall circumvention widgets such as the CORE Discovery browser plugin [366] or the Unpaywall plugin [367]. An example of a typical `GetRecord` response is provided in Fig.6.2.

Though OAI-PMH is the principal machine interface to repositories, and will likely remain so for the foreseeable future, it is expected to be superseded by ResourceSync, a de facto update to OAI-PMH. ResourceSync better aligns with current and future scholarly infrastructure technical requirements [368, 369] and is an essential component of the 'Next Generation Repositories' framework [370]; although issues with its performance (e.g. [92]) are such that efforts continue to improve the efficacy of OAI-PMH based metadata application profiles, such as RIOXX, the governance group for which the present author is currently chair [371].

6.2.2 Embedded metadata

Open repositories will support the embedding of rich page metadata, typically associated with the digital deposits of a repository [372]. By way of example, a default installation of EPrints provides rich metadata in a wide variety of serializations to ensure optimum interoperability with an unanticipated number of resource discovery agents. Data are available according to the EPrints Schema and DC, but also as RDF/XML, RDF N-Triples, RDF+N3, JSON, MODS, METS, and many more. This ensures inclusion requirements for academic services such as Google Scholar are met [373] while also satisfying general search engine interoperability. Interoperability with reference management and sharing platforms (e.g. Zotero, Mendeley, etc.) is also delivered via embedded `.ris`, `.bib`, `.enw`, etc., thereby — through user adoption of associated browser plugins — enabling detection of in-page metadata and the automatic importation of metadata and digital content into personal reference collections

[3] CORE: `https://core.ac.uk/`

[4] BASE: `https://www.base-search.net/`

Figure 6.2: An example of an OAI-PMH response to a `GetRecord` *query, used to retrieve an individual metadata record comforming to the RIOXX metadata application profile.*

```
<?xml version='1.0' encoding='UTF-8'?>
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:
     schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2019-11-28T14:39:26Z</responseDate>
  <request verb="GetRecord" identifier="oai:strathprints.strath.ac.uk:69192" metadataPrefix="rioxx">https://
       strathprints.strath.ac.uk/cgi/oai2</request>
  <GetRecord>
    <record>
    <header>
      <identifier>oai:strathprints.strath.ac.uk:69192</identifier>
      <datestamp>2019-11-27T05:19:18Z</datestamp>
      <setSpec>7374617475733D707562</setSpec>
      <setSpec>7375626A656374733D51:5144</setSpec>
      <setSpec>7375626A656374733D54:5441:5441313634</setSpec>
      <setSpec>74797065733D61727469636C65</setSpec></header>
    <metadata>
      <rioxx xmlns="http://www.rioxx.net/schema/v2.0/rioxx/"
      xmlns:ali="http://ali.niso.org/2014/ali/1.0" xmlns:dc="http://purl.org/dc/elements/1.1/"  xmlns:dcterms="
           http://purl.org/dc/terms/"
      xmlns:rioxxterms="http://docs.rioxx.net/schema/v2.0/rioxxterms/" xsi:schemaLocation="http://www.rioxx.net/
           schema/v2.0/rioxx/ http://www.rioxx.net/schema/v2.0/rioxx/rioxx.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
     <ali:free_to_read></ali:free_to_read>
     <ali:license_ref start_date="2019-09-13">http://creativecommons.org/licenses/by/4.0</ali:license_ref>
     <dc:description>The potential to bioprint and study 3D bacterial biofilm constructs could have great
          clinical significance at a time when antimicrobial resistance is... </dc:description>
     <dc:format>application/pdf</dc:format>
<dc:identifier>https://strathprints.strath.ac.uk/69192/7/
     Ning_etal_2019_3D_bioprinting_of_mature_bacterial_biofilms.pdf</dc:identifier>
<dc:language>en</dc:language>
<dc:source>1758-5082</dc:source>
<dc:subject>QD</dc:subject>
<dc:subject>TA164</dc:subject>
<dc:title>3D Bioprinting of mature bacterial biofilms for antimicrobial resistance drug testing</dc:title>
<rioxxterms:author>Ning, Evita</rioxxterms:author>
<rioxxterms:author>Turnbull, Gareth</rioxxterms:author>
<rioxxterms:author>Clarke, Jon</rioxxterms:author>
<rioxxterms:author>Picard, Frederic</rioxxterms:author>
<rioxxterms:author id="https://orcid.org/0000-0002-7708-4607">Riches, Philip</rioxxterms:author>
<rioxxterms:author>Vendrell, Marc</rioxxterms:author>
<rioxxterms:author id="https://orcid.org/0000-0002-6079-2105">Graham, Duncan</rioxxterms:author>
<rioxxterms:author id="https://orcid.org/0000-0001-8736-7566">Wark, Alastair W.</rioxxterms:author>
<rioxxterms:author id="https://orcid.org/0000-0002-5567-7399">Faulds, Karen</rioxxterms:author>
<rioxxterms:author>Shu, Wenmiao</rioxxterms:author>
<rioxxterms:project funder_name="EPSRC (Engineering and Physical Sciences Research Council)">EP/L016559/1</
     rioxxterms:project>
<rioxxterms:project funder_name="EPSRC (Engineering and Physical Sciences Research Council)">EP/N010914/1</
     rioxxterms:project>
<rioxxterms:publication_date>2019-09-13</rioxxterms:publication_date>
<rioxxterms:type>Journal Article/Review</rioxxterms:type>
<rioxxterms:version>VoR</rioxxterms:version>
<rioxxterms:version_of_record>https://doi.org/10.1088/1758-5090/ab37a0</rioxxterms:version_of_record>
</rioxx>
</metadata>
</record>
</GetRecord>
</OAI-PMH>
```

for organization and future reference, as well as sharing in the cloud with other users [374, 375].

6.2.3 Browse interfaces & link architecture

The way in which content within repositories is structured and organized will tend to follow faceted browsing conventions, with content delivered as human-readable equivalents of an XML sitemap (e.g. by year, author, subject, collection, etc.), as favoured by crawlers [373] and reflected in the 'sets' offered in OAI-PMH interfaces [361]. Most repository platforms also support XML sitemaps by default but guidance from leading search engines, Google for example, suggests that they tend not to be used if efficient access to site content can be achieved through well-organized pages demonstrating effective link architecture [376].

6.2.4 SWORD

The Simple Web-service Offering Repository Deposit (SWORD) protocol is another interoperability standard supported by repositories, allowing them to ingest deposits from multiple remote clients in multiple formats using a standardized M2M protocol. Typical use case scenarios are described by Lewis et al. [377] and can include the deposit of digital content into multiple repositories simultaneously, automatic deposit of content directly from desktop clients [378], repository-to-repository deposit, and so forth. SWORD is a profile of the Atom Publishing Profile (AtomPub) [379] and allows (authenticated) users to route content to recipient services. Recipient repositories may exercise discretion of what content is accepted (e.g. lack of authentication, file type not supported by repository, corrupted MD5 checksum, etc.). SWORD therefore provides an efficient mechanism for moving content across repositories, sharing content and exposing it. It is noteworthy that, based on SWORD, scholarly infrastructure services have emerged, such as the OpenAIRE Literature Broker [380] and the Jisc Publications Router [381], which seek to 'route' open digital content (from other participating repositories, publishers, learned societies, etc.) to recipient repositories thereby ensuring content which otherwise may be missed by those communities are captured, better exposed and digitally preserved.

6.3 Enhancing & evaluating repository discoverability: PW10 & PW11

Despite the out-of-the-box optimizations which benefit the discovery potential of repositories, there remains wide variation in the visibility of content delivered by some repositories. For example, Arlitsch and O'Brien have reported inconsistencies in the indexing of repositories by Google Scholar [382], something confirmed by Google Scholar engineers [373]. Askey and Arlitsch have also noted the influence of other signals in promoting search engine visibility [383].

Inconsistent configuration of OAI-PMH endpoints and a failure to model metadata according to established schema or application profiles within OAI-PMH responses continues to be an issue for aggregation services too. This is especially the case for repositories exposing only minimal DC (via OAI-PMH) and where content referencing is inconsistently applied; for instance `dc:identifier` failing to reference digital file(s), or use of `dc:relation` to reference content which is better suited to `dc:identifier` [384, 385].

Variation in OAI-PMH metadata modelling, and a failure to implement schema with greater specificity, can result in lower levels of harvesting by aggregation services and ergo lower visibility. Recent experiments comparing the performance of OAI-PMH and ResourceSync found wide variation in the average number of OAI-PMH requests it took to accurately identify content on certain repository platforms (where DC was exposed). For example, accurate identification of file content on EPrints took an average of circa 9 OAI-PMH requests, while on Digital Commons it required an extraordinary 13,286 requests [386], suggesting that variation was in some cases hardwired into repository software and necessitated intervention from repository administrators to correct. Such variation in metadata modelling is typically replicated in any embedded metadata exposed by repositories too, resulting in inconsistent indexing by search agents.

Open science is at a crossroads, something which potentially limits the discovery of repository content. The influence of national research assessment frameworks and increased research management at HEIs has stimulated the growth of Current Research Information Systems (CRIS). CRIS software, though not new (e.g. [387, 388]), has emerged as a solution to the information management tasks associated with research assessments; however, this focus on CRIS software is commonly to the detriment of open science and the knowledge commons. This is because some institutions elect to decommission or deprioritize repository development in favour of their CRIS software in order to satisfy their research management remit [389, 390] thereby neglecting commitment to the principles of open science or the scholarly knowledge commons. Most CRIS software is proprietary and operated by organizations openly hostile to open science (e.g. RELX) [391]), is ill designed for participating in the knowledge commons, and generally demonstrates poor support for the technical expectations of the repository cooperative [**392, 322**], [393]. In fact, the dichotomy between open science and research management is at the core of much debate in the literature about the future of scholarly publishing and academia, a debate which though interesting is outside of the scope of this thesis (see instead for example: [394, 395, 396, 397]). Suffice to state that it is necessary to re-articulate the importance of repositories as nodes within open scholarly communications infrastructure and their superior capabilities in supporting resource discovery.

Despite numerous contributions to the literature about the importance of repositories in supporting discovery, work by only a single group of researchers has sought to investigate the nature of certain variables [398, 399, 383, 382]. Further still, these works have not sought to codify variables or to measure the impact of repository optimizations on discoverability. **PW10** and **PW11** [**321, 322**] respond to this gap in our understanding and use preliminary results reported by the present author elsewhere [**400**] as a starting point for a more detailed research narrative. Among other things, the works seek to explore the potential correlations between visibility / discoverability (i.e. the independent variables) and COUNTER usage / web impact (i.e. the dependent variables).

Both works study and codify specific technical adjustments and improvements which are hypothesized to enhance repository discoverability, with PW11 continuing the 'further research' narrative established by PW10. In fact, PW11 could be described as a direct continuation of PW10 because it:

- involves observations and evaluation using an overlapping but larger longitudinal dataset in order to confirm the indicative results from PW10, and;
- employs additional data and analytical techniques to corroborate conclusions.

The published works assembled for this chapter are both journal articles (PW10 and PW11), the second of which was originally a peer-reviewed conference paper invited for journal publication in an enhanced form (after undergoing an additional round of peer-review). Both works are presented within the context outlined in previous sections and both seek to better understand the multitude of variables which influence repository visibility and discoverability. The works adopt a 'before-after' repeated measure experimental set-up, entailing the capture of relevant data prior to the implementation of technical changes and the monitoring of those data in the months and years afterwards. This approach enabled the effects of change to be observed during the temporal periods selected for study which, for reasons of refutation, are different in each work.

Research in an area dependent upon third party systems (e.g. search engine search and indexing) and data analytics immediately confront experimental compromises. To secure the most reliable results such a study should adopt a controlled experimental set-up, with two identical repositories — one representing the controlled repository and the other the experimental (or 'treatment') repository. This methodological approach is common within resource discovery related research, most notably experiments typifying TREC-based information retrieval, where the comparative performance between different retrieval techniques can be measured against controlled document collections [401, 402]. However, because there is a dependence upon third party systems in the topics studied as part of PW10 and PW11, the suggestion of a controlled experimental set-up is hypothetical. It is impossible to effect change or exert control on third party systems and, owing to the unknown nature of how certain aspects of how some third party systems function, it is impossible to control for all variables hypothesized to influence the visibility of repositories. Use of a repeated measure set-up in PW10 and PW11 was therefore a necessary compromise in order to study this topic area. The reliance on data analyses which are largely dependent on correlation and inference, as opposed to direct causation, are also a consequence of the reduced level of control the present author had over the experimental process.

Across both published works, the present author nevertheless considers the necessary compromise in study design to have been compensated for by the following:

1. Capture of data from multiple data sources to ensure a level of triangulation in findings, including use of COUNTER[5] usage statistics [403] via IRUS-UK[6], a variety of web analytics derived from Google Analytics [404], and retrieval data extracted from Google Search Console [405]. Transparency surrounding the use of Google Search Console data are addressed in both works and again

[5] COUNTER: https://www.projectcounter.org/

[6] IRUS-UK (Institutional Repository Usage Statistics UK): https://irus.jisc.ac.uk/

are a result of Google's current dominance in search and the lack of comparable tools from competitors.

2. Deployment of inferential statistical analyses, in addition to descriptive statistical analyses, to deliver data insights which might be expected within conventional controlled experiments.

3. Replication of the study (in PW11) using a larger longitudinal dataset, segmented over different temporal periods, to control for any cyclical patterns in repository usage.

4. Inclusion of additional statistical analyses in PW11 to obviate shortcomings in analyses reported in PW10, most notably the use of exponential regression as a way of discounting a possible exponential relationship between the volume of content deposited and the repository usage generated.

These four steps counterbalance the underlying compromise to deliver findings that are as robust as can be expected given the methodological constraints.

Possible criticisms of the works assembled for this chapter are that they are insufficiently holistic in their interpretation of visibility and discoverability. A more holistic evaluation could have included greater consideration of alternative discovery mechanisms (e.g. OAI-PMH) as opposed to being restricted to search engine discovery. While the motivation for the present author's particular focus in PW10 and PW11 is articulated clearly, and data within the works support a focus on web search, the motivation for considering other aspects of discovery are similarly strong given the limited evaluative work which has been undertaken over the years. As with the discovery potential of repositories to search agents, scientific investigation of the efficacy of other mechanisms, such as OAI-PMH, SWORD, etc. has been limited. OAI-PMH compliant repositories and so-called 'static' OAI repositories [406, 407] have all enjoyed thorough treatment; but much of this treatment has been either theoretical, based on the self-evident benefits of exposing content via OAI-PMH [408]. Or they have been entirely anecdotal and based on the logical corollary of open content exposed via the knowledge commons [409]. Instead significantly more research time has been spent on ancillary topics such as the influence of metadata quality in repositories [410, 411, 412, 91]. Though it has implications for content aggregation and is undoubtedly important for systems interoperability, metadata quality is rarely considered in relation to resource discovery. Only McCown et al. [413] and Allison [414] attempt to measure OAI-PMH and its implications for discovery, with McCown et al. providing an early exploration into the search engine coverage of OAI-PMH compliant copora and Allison evaluating users' engagement with OAI-PMH harvested content within a wider digital library context. Measurement of to what extent SWORD content deposits — via any of the use cases mapped out by Lewis et al. [377] -– can contribute to discoverability or the visibility of open content also remains under-researched. Most literature explains SWORD's potential [415] or describes novel system implementations rather than performing any evaluation (e.g. [380, 416, 417]).

The wider digital library context highlighted by Allison [414] is relevant when due consideration is given to the wider open commons and the extraordinary volume of content now being aggregated by services such as CORE which, at time of writing, is in excess of 136 million papers [418]. The use case for CLM — which the present author introduced in Chapter 3 -– within OAI-PMH interfaces remains an unexplored one too. Its potential for supporting improved discoverability has nevertheless been recognised by researchers [419, 420, 421]. This recognition has partly been a consequence of the OAI-PMH protocol itself, which natively supports collection-level metadata within an `<identify>` response [361]; but also its potential for exposing far richer CLM using DC based schema, such as the DC Collection Description Terms [44]. Only Foulonneau et al. [27] have attempted integrating CLM within OAI-PMH item-level metadata as part of a prototype system. Foulonneau et al.'s work is especially noteworthy since it also reports on preliminary findings of its retrieval potential, about which they conclude:

> Collection-level [metadata] can be used as a way to preserve or restore context otherwise lost when item-level metadata are harvested from disparate and heterogeneous repositories and can also provide an additional level of descriptive granularity that may be better suited for some user queries.

Despite these positive findings no systematic programme of research ever emerged from Foulonneau et al. or others active in metadata, knowledge organization or resource discovery research. This is surprising because the intended CLM use case is now more applicable within the rapidly expanding knowledge commons, yet none of this was a research consideration for the present author when planning and conducting the research necessary for PW10 and PW11. Upon reflection this appears to reinforce a lack of cognisance surrounding the present author's research history and how this history can apply to new or emerging research themes.

It nevertheless remains noteworthy that not all aspects of PW10 and PW11 lack this cognisance. For example, the increasing relevance of semantically aware structured data to search tools, especially through the recent incorporation of schema.org into algorithms such as PageRank [422], is factored into the research — providing a conceptual link to the research themes examined in Chapter 4. This is achieved through the native support for RDF-based schema demonstrated by the case study repository software but, secondly, through the use of the Google Data Highlighter tool [423] as one of the techniques to optimize repository indexing [**322, 321**]. By deploying this pattern matching tool it was possible to replicate schema.org data within the technical constraints of the case study repository system.

This is not to diminish the research and findings published in PW10 and PW11 as insignificant. They are a unique contribution to community understanding of how open repositories — as significant nodes within open scholarly communications infrastructure and the wider the global knowledge commons -– can be optimized to deliver demonstrable improvements in discovery potential.

Chapter 7

Methodological evolution

7.1 Addressing resource discovery actors

This thesis began by considering a unifying theme of resource discovery and the present author's motivation to explore a series of interlinked research topics within the broader research theme of resource discovery. As we have noted, there are some who have proposed taxonomic analyses of resource discovery in order to aid understanding and investigation, including by Vanthournout et al. [20]. They have noted that resource discovery typically involves three principal actors: *resource providers*, *resource users*, and the *resource discovery service* itself. As the present author has demonstrated, the published works assembled — discussed and critiqued in previous chapters — have examined issues involving all three of these taxonomic 'actors', with some works examining several simultaneously:

- **Resource providers**: PW5, PW6, PW7, PW9, PW10, PW11
- **Resource users**: PW9
- **Resource discovery service**: PW1, PW2, PW3, PW4, PW6, PW7, PW8

The works presented have therefore addressed all salient actors considered to operate within typical resource discovery scenarios. Moreover, the conceptual and methodological approaches adopted throughout these works were diverse, reflecting the nature of the topics under investigation or consideration, especially the varied nature of the digital scenarios presented; but, taken chronologically, these works also demonstrate an incremental evolution in methodological sophistication and a greater confidence in the scientific method. While the approaches adopted for the published works have been studied in prior chapters, it is worth summarizing all the techniques used. They include: theoretical / conceptual, literature based, questionnaire (including bespoke and established instruments), protocol analysis, stimulated recall, data analytics, statistical analyses, focus groups, qualitative content analysis, and comparative analysis. In most cases the works were multimethod (i.e. involving multiple data collection methods) but also included several instances of 'mixed methods research' (MMR) in order to triangulate findings, with numerous techniques often deployed within the same work (e.g. PW2, PW9) [424, 425].

7.2 Transition to pragmatism

As noted in the critical commentary accompanying prior chapters, works from earlier in the present author's career tend to demonstrate higher levels of methodological immaturity, with some methodological limitations being especially apparent in the light of reflective criticism. This is to be expected given that the works chart the present author's research career, from inexperienced researcher to experienced. Earlier works, including those not included within the thesis, often subscribed to an ad hoc research approach, with less coordination in the use of research instruments or consideration of their efficacy. However, it can be observed that MMR — as distinct from multimethod — began to be deployed more frequently as the present author's research career developed. This is an observation which becomes especially observable when published works outside those assembled for this thesis are also considered (e.g. [**426, 231, 227, 427, 229, 189**]; additional methodological approaches and alternative research instruments are introduced (e.g. 1-2-1 interviews, group interviews, Most Significant Change (MSC), etc.). Furthermore, as the present author has evolved, greater theoretical consideration is given to mixing methods, with improved use of triangulation across data gathering techniques to ensure the highest levels of validity. It could therefore be suggested that this development towards MMR was a tacit acknowledgement that hitherto methodological approaches had been too simplistic given the complex human-computer interaction issues under examination; that optimum understanding of the research phenomena could only be achieved through a combination of quantitative and qualitative data [428, 270]. Of course, it should also be highlighted these techniques were often underpinned by technical development work on the part of the present author.

Given the commitment to a diverse suite of research methods and data collection techniques, as well as the heterogeneous nature in which they were deployed, it may also be suggested that the present author's methodological journey has concluded in the creation of a *methodological pragmatist.* A pragmatist, in this context, refers to the philosophy of *pragmatism*, as originally posited by philosophers such as William James [429] and John Dewey [430], and which is frequently linked to MMR approaches to research enquiry [431, 432, 433, 434]. Pragmatism is not a research paradigm that the present author explicitly incorporated into his research practice, but rather one that evolved naturally and tacitly as the research phenomena under investigation grew in their complexity. Pragmatism goes beyond reaching for what, for example, 'simply works' or is 'pragmatic' in any given MMR scenario [433]. It presents a coherent philosophy to underpin research enquiry, emphasizing the power of analysed experience for practical-minded researchers [435]. More than this, it is an acknowledgement that the metaphysical and epistemic questions which can occupy some research paradigms can distract from the application of research methods. Instead pragmatism focusses itself towards solving practical problems in real world situations [434].

Solving practical problems in real world situations is certainly a common theme within the assembled works, as well as those not included within this thesis; but it can also be stated that the assembled works represent a de facto rejection of alternative research philosophies, while emphasizing empiricism. Empiricism is normally associated with positivism [436] but nevertheless remains important to pragmatism [434], to such an extent that it is occasionally criticised in some disciplines as a 'Trojan horse' for positivism, e.g. [437]. This is not to state that diverse research strategies are unused by the present author, or that theoretical analyses or understanding are absent from the works; PW1, PW2, and PW9, for example, demonstrate conceptual understanding, analytic induction and thematic analyses. However, it is about the adoption of a paradigm that seeks to combine the strengths of qualitative and quantitative methods in practicable or actionable ways (i.e. MMR), rather than valuing metaphysical or epistemic questions.

Nor is it to state that pragmatism 'lacks philosophy' [438]. In fact, in conjunction with MMR, pragmatism provides a useful – and some argue necessary [439] – framework for understanding or approaching enquiry within the present author's research areas. For example, Dalsgaard & Dindler [440] have introduced the notion of 'bridging concepts' as an intermediary form of knowledge, situated somewhere in-between theory and practice. Their approach to human-computer interaction research is based entirely on Deweyan pragmatist concepts. The summation of the present author's methodological evolution as one of *pragmatism* is therefore a natural evolution and it parallels recent trends within wider information science scholarship [428, 270, 441, 442], itself the outcome of the overtly positivist approaches which have historically typified the information and computing sciences. The dangers of positivism have been identified by noted information seeking and knowledge organization scholar, Birger Hjørland, as obsessing about "correlations between variables while drawing no conclusions about causes" [436]. Maintaining cognisance of the core pragmatist principles will therefore be fundamental for the present author's research going forward. It will be essential to ensure research questions are studied holistically and in line with emerging expectations within the author's discipline [439, 442], but also to avoid so-called Trojan horse criticisms or obsessions about correlation.

Though methodological limitations have been noted in our collective appraisal of the assembled works, numerous instances of novelty and contribution to knowledge have also been noted. These contributions will be described more substantively in Chapter 9; but the works can also be said to suggest the emergence of a unifying theory of resource discovery, through the concept of *(meta)data alignment*. This concept will be proposed and discussed in more detail in the next chapter (Chapter 8), alongside a consideration of the present author's future research agenda.

Chapter 8

(Meta)data alignment & future research

8.1 (Meta)data alignment as a unifying theory of resource discovery

Returning to the notion of a unifying theme of resource discovery, it could be stated that the interlinked research topics surrounding resource discovery, which have been considered and assembled as part of this thesis, are unified in their exploration of *(meta)data alignment*. The works have studied metadata (in their variety of permutations) but primarily structured open data (e.g. applications of RDF/XML, XML, etc.), the expression of KOS as structured data, the syntactic and semantic interoperability issues that arise in distributed resource discovery contexts, the influence of data optimization in promoting the discovery of open content, and the implications of all of the aforementioned on questions of human-computer interaction.

As a concept 'metadata alignment' remains ill defined. Its usage is loose in the literature, emerging in the mid-2000s as disparate data corpora grew in size and number thereby raising challenges in integration and interoperability. For example, it is most frequently used to describe miscellaneous approaches to merging metadata schema, augmenting or optimizing them, or 'aligning' knowledge structures such as taxonomies, thesauri, and ontologies, with considerable experimentation noted within digital heritage and digital libraries, e.g. [443, 444, 445, 446, 447, 448, 449, 450]. Even with such fertile experimentation it is disappointing to note that no scholars have attempted to formalize what 'metadata alignment' means. In this conceptual vacuum it is possible to propose a definition based on the published works assembled for this thesis.

Utilizing the concept of ‘alignment’ and applying it is unsurprisingly popular within a number of research disciplines. ‘Alignment’ could be said to describe the arrangement of things, whatever they may be, in a straight line or perhaps in parallel, or to correct their relative positioning. From Merriam-Webster it is clear that the English definition finds application more readily within engineering, as per “[Alignment] : the act of aligning or state of being aligned, especially: the proper positioning or state of adjustment of parts (as of a mechanical or electronic device) in relation to each other”[1]. However, despite its engineering connotations — or perhaps because of them — the concept has been extended and applied within disciplines such as cognitive sciences, data science and educational psychology. ‘Conceptual alignment’ is an important idea within cognitive sciences and social cognition as it describes how two persons achieve mutual understanding by using ‘the same computational procedures, implemented in the same neuronal substrate, and operating over temporal scales independent from the signals’ occurrences’ [451]. In other words, these two persons achieve mutual understanding because they are conceptually aligned [452].

These ideas of conceptual alignment have been adopted elsewhere, most notably within aspects of data mining and machine learning in order to simplify understanding of the complex data problems associated with domain modelling. Here conceptual alignment is principally concerned with “preserving partial isomorphism which maps formal concepts of one concept lattice onto formal concepts of another concept lattice. Two concept lattices are in total conceptual alignment if a total order-preserving isomorphism exists between them” [453]. Achieving direct alignment of conceptual lattices is central to thinking here, whereas in educational psychology and pedagogical domains, the alternative notion of ‘constructive alignment’ is well-established and has become a key tenet of educational theory and psychology. Pioneered by educational psychologist, J. Biggs [233, 454], and explored or tested by countless educational scholars (e.g. [455, 456, 457, 458]), constructive alignment describes the optimum conceptualization of a curriculum design: one in which there is a synergy between the stated aims of the learning programme, the intended learning outcomes for learners and the design of assessments and their evaluation criteria. Harmonization between these layers of the pedagogical process promotes deeper levels of learning as students are forced to draw upon higher cognitive resources. The greater the harmonization, the greater the expected learning impact on the student.

Based on the work assembled for this thesis it could be proposed that *(meta)data alignment*, at a conceptual level at least, encapsulates aspects of the philosophies used

[1] Merrium-Webster, ‘alignment’: `https://www.merriam-webster.com/dictionary/alignment`

in both the conceptual and constructive alignment approaches described above. Ergo, in our context, *(meta)data alignment* could be said to be principally concerned with the arrangement of data structures across horizontal or vertical layers, in such a way as to promote ever increasing levels of harmonization and mutual understanding between systems, thereby promoting commensurate improvements in discovery impact for users. The 'meta' in *(meta)data alignment* is placed within parentheses because within this thinking we are concerned with not just the descriptive, structural, administrative or technical metadata normally associated with many present day metadata schema [459], but also alternative data structures, whether these be KOS vocabulary specifications or metadata application profiles, and the data preferences of the discovery services that may consume this data and present it to users.

The model diagram below (Fig. 8.1) provides a visual representation of this concept to aid interpretation.

- Collection-based services are situated at the top of the diagram. Such services are those powered by CLM and offer information landscaping functionality, or could even be services powered by collaboratively generated collection descriptions [36]. This represents the broadest step in the resource discovery process as users' discovery is initiated at a general level, involving the identification and elimination of entire corpora.

- The distributed service layer encompasses services such as open content aggregators, distributed digital libraries, federated search tools, digital services offering clumping, and so forth. Some of these services may have been identified as a result of interactions with the collection-based services as possible routes to discovering content at the item level, as per the information landscaping techniques described in Chapter 3.

- The semantic interoperability layer includes terminology services, concept linking tools delivered via LOD or semantically aware middleware. This layer is principally concerned with aligning user queries entered at the distributed service layer with the service layer below.

- The discovery layer represents the conclusion of *meta(data) alignment* as conceived in this model, insofar as the alignment in all prior layers of the model have promoted optimum syntactic and semantic interoperability at all levels, and within all data structures used by participating systems, thereby delivering superior discovery experiences for end users.

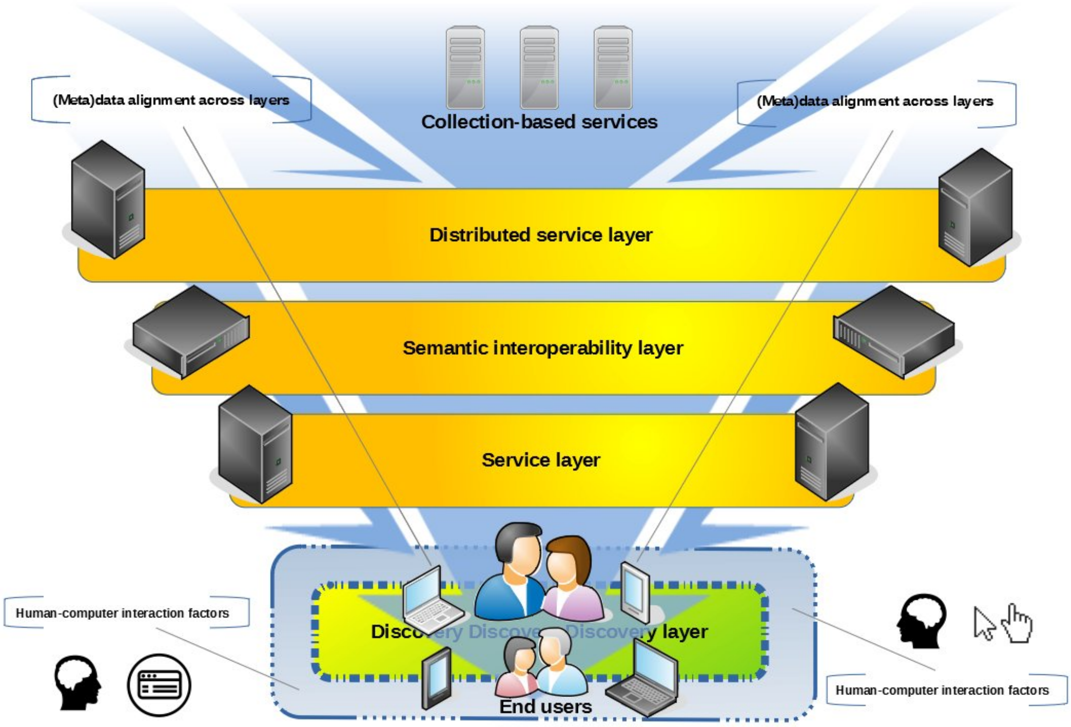

(Meta)data alignment across layers
Collection-based services
(Meta)data alignment across layers
Distributed service layer
Semantic interoperability layer
Service layer
Human-computer interaction factors
Discovery layer
End users
Human-computer interaction factors

Figure 8.1: Model diagram of (meta)data alignment.

- By time we reach the discovery layer, factors relating to HCI are introduced which can influence end users' interaction with any discovered resources and associated metadata, can impact on users' personal information management, users' ordering or arrangement of results and so forth. This is diagrammed as subsuming the discovery layer but it does not represent a layer in itself. No *(meta)data alignment* need take place here as this has been performed in all preceding layers. Instead, the principal considerations here are how structured data are harnessed and put to use by resource discovery services. In other words, how data are presented to users so as to aid their interactions with the resources and its metadata, while simultaneously minimizing users' cognitive load.

The model presented in Fig. 8.1 not only goes some way to describing the works in this thesis, their coherence when considered collectively, and their relationship to each other; but could also be said to have applications in other controlled information environments, such as those described in previous chapters (e.g. digital libraries, open repositories, etc.), since alignment across layers is necessary to deliver high-quality discovery services to users. The greater the integration between systems and layers, the greater (meta)data alignment is necessary. For this reason, the model is not generally applicable outside controlled information environments, such as generic web search tools. For example, there are fewer layers in these contexts as integration between collection-based services and distributed service layers does not exist, nor does the same level of rich metadata exist to power discovery. Similarly, the semantic interoperability layer cannot be said to exist in the open web in any meaningful way. There are well documented initiatives to ameliorate this issue [460, 461, 462] but its implementation by resource providers remains inadequate.

The model assumes local resource providers and resource discovery services to have a level of control over their systems. An emerging but significant development in the provision of proprietary software solutions has been a gradual move towards software as a service (SaaS) [463]. Digital heritage institutions, libraries, archives and so forth are among the bodies which, for some services, have migrated from local delivery systems to those based on a SaaS approach [464]. Though SaaS software tends to benefit from the economies of scale secured through multiple customers sharing the same software and infrastructure, the consequence is rigidly standardized software products. These products will typically offer serviceable functionality in order to meet the base requirements of the maximum number of customers [463]. This presents potential difficulties for longer term meta(data) alignment in some layers of the model – or in some future scenarios, since alignment is predicated upon an ability to control, customize and optimize systems, thereby enabling *(meta)data alignment* across layers.

8.2 Future research

8.2.1 Equivalence match types, interactive QE & automatic mapping via 'terminological dataset triangulation'

The research trajectory and the unifying narrative of the assembled published works has been explained. However, analyses and criticism of the works in the preceding chapters highlighted both lost opportunities for further research and also prompted areas for future work, all within the continuing context of (meta)data alignment within heterogeneous discovery contexts.

Among the most notable areas is further exploration of possible ways the equivalence match types proposed in **PW8** should be modelled within the RDF data model for deployment alongside the wider SKOS specification. Not pursuing this line of enquiry was highlighted as a shortcoming of the works presented in Chapter 4; yet, little has changed since PW8 was released into the public domain. SKOS now incorporates within its `skos:mappingRelation` a series of conceptual matches: `skos:closeMatch`, `skos:exactMatch`, `skos:broadMatch`, `skos:narrowMatch` and `skos:relatedMatch` – all of which are included within the main vocabulary instead of a separate RDF vocabulary, as was initially the case with the MVS [465]. A level symmetry between SKOS and OWL is maintained. The limitations of the existing SKOS mapping equivalences are noted in the literature [466] and emanate largely from its focus on conceptual equivalence at the expense of semantic equivalence [467]. There is, therefore, clear merit in proposing a more detailed, separate RDF vocabulary using the findings of PW8 as a foundation and evaluating their efficacy within terminology services and/or alternative Semantic Web or LOD applications.

A motivation also exists in evaluating the mapped concepts and their associated match types with users, but within the context of interactive QE functionality (e.g. within digital repositories, digital libraries, retrieval systems, etc.). Recall that a principal function of the terminology service presented in **PW6** and **PW7** was also to support the integration of terminological data within local services thereby enabling them to harness the hierarchical or syntactic relationships for browsing or interactive QE. Further exploration of this research agenda has been attempted by others [468] but it seems self-evident that any alternative approach to mapping necessitates re-evaluation of interactive QE efficacy, especially if the breadth and quality of mappings improves.

Such research also needs to accommodate matters pertaining to cognitive load in relation to match types, most likely via user evaluation in a lab setting. Match types are intended to convey meaning about the nature of an equivalence mapping and therefore function as an indicator of relevance to end users; particular types of mapping match type therefore infer a level of relevance more than others and the prospect of retrieving more relevant resources. Notwithstanding the absence of any user-centric evaluations of this nature, the prospect of introducing a suite of potentially more complex match types has the potential to cause user confusion during resource discovery and consequently demands rigorous testing.

Following the notion that mapping breadth and quality needs to improve, work in KOS mapping could be advanced through alternative automatic approaches. Though pure automatic approaches were not explicitly addressed in the works presented in Chapter 4, research investigating automatic KOS-to-KOS mapping has been ongoing

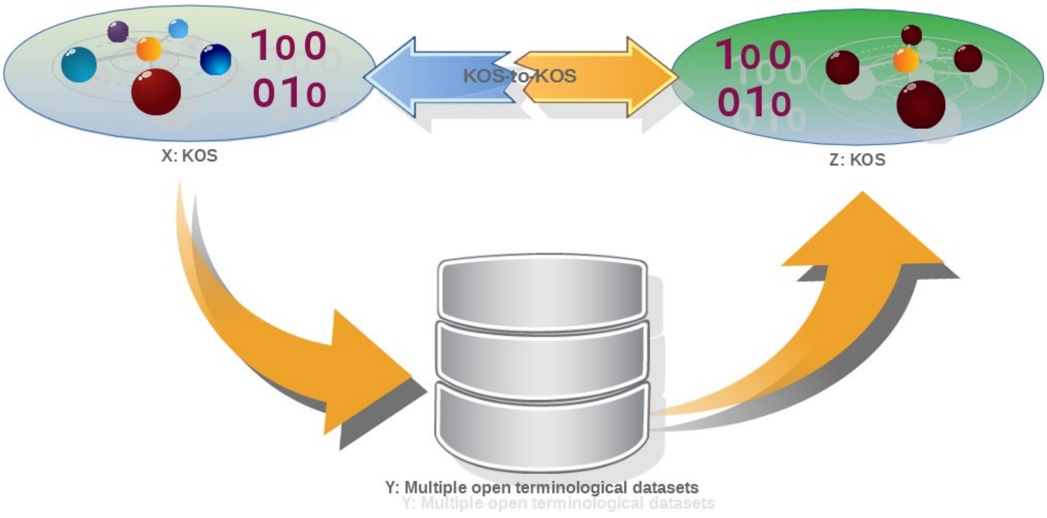


Figure 8.2: A proposed automatic concept mapping approach based on 'dataset triangulation'.

since the late 1990s (e.g. [469]) with exploratory work undertaken within medical informatics as far back as 1991 [470]. Its lack of success was why the assembled works in Chapter 4 concentrated on human intervention with the possibility of harnessing a collaborative, wiki-inspired approach to scaling up and maintaining mappings [**133, 185**]. Others have proposed 'interactive' mapping models whereby the process is essentially a semi-automatic one, with automatic mappings verified and, where necessary, modified via human intervention. This helps to maintain the highest data quality, something deemed especially necessary within digital heritage and digital library contexts [471, 472, 473].

Though the 'state-of-the-art' for pure automatic mapping research remains limited [474] there have been successes reported recently (e.g. [475, 476]. Successes have sought to improve the calculation of concept closeness [474]. The existence of large terminological datasets is now more common and their open availability is generally assured. A potentially fruitful line of enquiry would therefore be to explore automatic KOS-to-KOS mapping -– or via a switching KOS or spine -– the using such datasets to triangulate mappings. For instance, Ballatore et al. [477] evaluated WordNet as a semantic hub to increase the success of KOS integrations. They developed

Voc2WordNet, an unsupervised mapping technique designed for mapping geographic terms, which employs WordNet as a 'semantic support tool' to assist in the discovery of 'implicit semantic relations between features, such as subsumption or meronomy...'.

It therefore seems apposite to explore new automatic mapping approaches using what could be described as 'dataset triangulation'. That is to say, using multiple open terminological datasets to aid in the computation of KOS mappings (either directly or via switching or spine) (see Fig. 8.2). In such a scenario it would theoretically be possible to interrogate certain large, open, terminological datasets as a 'sense checker'; establishing more accurate machine-based mappings by verifying conceptual and semantic similarities or closeness based on the features of other KOS within the terminological datasets. This creates triangulation by delivering KOS-to-KOS mappings that are potentially more accurate, because mappings from X to Z is mediated by sense checking by Y.

The research strands presented in this section detail a series of exciting lines of academic enquiry, all of which have the potential to influence the design or application of (meta)data within digital libraries, repositories and information retrieval systems. But there are also research strands arising which are more practice-based in nature.

8.2.2 CLM, resource discovery & the discovery of open commons content

The use case for CLM in resource discovery has been well established as part of this thesis, especially within Chapter 3. We also noted in Chapters 3 and 6 that an untapped use case for CLM exists within resource discovery research and that promising results found by Foulonneau et al. when deploying CLM within repositories resulted in no programme of further research [27]. This may have been because — at the time of publication — the number of potential real world applications for their results was limited by the immaturity of global repository infrastructure at that time. However, this line of enquiry is arguably more relevant today given the growth which has subsequently been observed. Massive growth in the number of repositories has been observed over the past 10 years, all of which have been cultivating an increasing volume of digital content . When Foulonneau et al. performed their research a mere 85 repositories were operational globally (see Fig. 8.3 [2]), and it seems reasonable to assume that a large proportion of these would have been prototypical

[2] As per Open Directory of Open Access Repositories (OpenDOAR): https://v2.sherpa.ac.uk/view/repository_visualisations/1.html (CC-BY-NC-ND). Graph captured 08/02/2020).

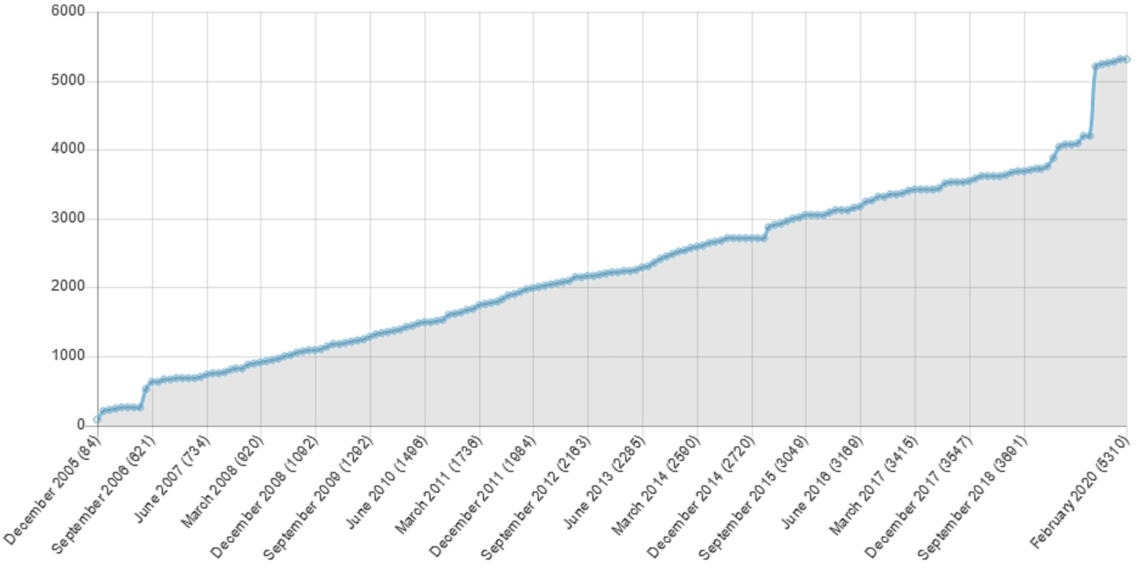


Figure 8.3: Global growth in the number of repositories registered with OpenDOAR since late 2005 to February 2020.

or experimental systems and therefore unsuitable to end users and, ergo, as a viable testbed of research.

Given how much more conducive the current global repository context is to testing such a line of enquiry, it would be highly productive to revisit the trajectory first established by Foulonneau et al. and evaluate it within today's environment; one where superior resource discovery support for users is needed and one in which a more useful role for CLM can be envisaged to improve item-level retrieval for users. It could be envisaged that repositories could expose CLMs within OAI-PMH `<identify>` responses to aggregators but according to a new, dedicated CLM schema and, in particular, one which better accommodates concept URIs. This would enable increased accuracy in subject-based information landscaping and therefore, theoretically, increased accuracy in item-level retrieval -– all of which is worthy of evaluation and necessary to better understand the discovery potential CLMs within the knowledge commons.

Of course, item-level retrieval through search remains a dominant form of discovery for content held within the knowledge commons. **PW10** and **PW11** documented techniques for optimizing open repositories for discovery agents, through a combination of what we are now terming (meta)data alignment approaches. The impact of these techniques were evaluated and positive results were reported. However, a number of caveats and limitations were noted, some of which are explicitly enumerated in PW11 [**322**] and form an additional research agenda for the present author:

- *Traffic latency*: Monitoring and measuring traffic latency as a possible factor on improved Google Scholar indexing of repositories;

- *Coversheets*: Better measuring the influence, if any, of coversheets on repository indexing, a factor which is purported to inhibit crawling and indexing by services such as Google Scholar and others [373, 478] but the validity of which was questioned by PW11.

- *Combating the correlation fallacy*: Establishing a more robust study design for evaluating the relative visibility and discovery potential of open repositories. Such a study design would seek to improve the control of extraneous variables to obviate the 'correlation fallacy' described in PW11. Better control would dictate a study design involving collaboration across a number of participating repositories, thereby enabling present usage and web analytics to be benchmarked across a larger number of search agents, with specific repositories used as a control. This would be a large project, involving numerous participant repositories and institutions, and may be outside the scope of what is normally considered routine 'practice-based research', necessitating dedicated project funding.

- *Enhanced analytics = Enhanced insights*: Introducing additional, sophisticated analytics into the evaluation methodologies used deserves further attention. There continue to be numerous analytics and metrics which can be studied to glean new perspectives on repository visibility and discoverability. In addition to those described in PW10 and PW11, there are opportunities to combine data on search queries, web traffic and COUNTER usage with larger, related datasets, such as those from Kibana[3]. Kibana — which is embedded within the present author's repository infrastructure — facilitates the extraction and visualization of Elasticsearch data and navigation of the Elastic Stack, enabling novel facets of the research problem to be analysed with alternative units of measurement. For instance, Kibana can facilitate the gathering of diverse data which can then be used track query load, better understand how users flow through repositories, and then harnessed to perform graph analyses which could potentially uncover hidden relationships between the content users access.

Each of the above research agenda items has potential to contribute significant real world impact within the knowledge commons by improving community understanding

[3] Kibana: https://www.elastic.co/kibana/. Kibana code available from GitHub: https://github.com/elastic/kibana

of how repositories and their digital content interacts with discovery systems and users.

All of the research strands presented in this Future Research section also crystallizes a research career transition for the present author. From one in which the research agenda has been predominantly 'academic' to another that is predominantly, though not exclusively, 'practice-based'. Indeed, the research trajectory described in section 8.2.2 is conducted within a practice-based environment, using technology, data and information services that engage real world users while simultaneously reporting the findings of this research, not only to academic audiences, but to practice-based audiences too.

Chapter 9

Contribution to knowledge

Collectively, the works assembled for this thesis contribute to the body of knowledge on resource discovery within heterogeneous digital environments and have improved our understanding of how resource providers, resource users and resource services can influence the overall efficacy of the resource discovery experience. Throughout Chapters 3-8 instances of novelty have been noted, as well as the various contributions to knowledge that the assembled works have made. This Chapter addresses these contributions more substantively, highlighting their academic contribution, and, in later sections, discusses the implications of the assembled works for practitioners.

9.1 Contribution of individual works

From the published works presented as part of this thesis, it is possible to summarize the following contributions to academic knowledge:

1. *Advanced our understanding of the role of CLM in resource discovery and its use in information landscaping approaches within distributed digital library scenarios.*

   This contribution emerges from **PW1** and to a lesser extent **PW2**. As **Chapter 3** reported, at the time of publication the use of CLM in resource discovery tools remained embryonic. There were few documented examples and little exploration of the role CLM could assume in the information landscaping of large digital corpora. Instead investigations into improved resource discovery tools tended to focus on item-level retrieval. Perhaps more significantly in this case, theoretical and conceptual work surrounding definitions of functional granularity were lacking. **PW1**'s contribution to knowledge was to address these gaps and to deliver additional conceptual and theoretical work such that CLM could be better integrated into digital library resource discovery tools. Such a contribution to community understanding appears to be acknowledged by its influence in a large number of subsequent works which explore and prototype approaches to CLM and CLM-based systems, e.g. [479, 480, 51, 52, 37, 54, 53, 34, 35, 481, 482, 37, 483].

2. *Established national metadata interoperability strategies for distributed digital libraries which were adopted by national services.*

   The concept of 'clumping' was introduced in **Chapter 3** as enabling federated searching of multiple digital libraries and repositories (or 'targets'). Clumping was noted as significant for the research contexts of PW1, PW2 and PW3. The research context for **PW2** was the poor adherence of cooperative members with the cooperative's technical expectations, resulting in the retrieval of inconsistent or even low quality results for users. But the research context also concerned resolving questions on the efficacy of distributed or centralized cooperative approaches. **PW2**'s contribution was to improve community understanding of the management of distributed digital libraries and the interoperability problems therein. It was the first – and remains the only – study of its kind to conduct an evaluation of distributed and centralized approaches. It also confirmed a negative finding surrounding the unsatisfactory level of interoperability between distributed digital libraries and proposed the creation of national metadata strategies and transferable recommendations on Z-server management. These were formalized in a separate report [**89**], which national services such as Copac[1], but also regional services, then implemented [66]. PW3 was also awarded 'highly commended' as part of the 2005 Emerald Literati Awards, in recognition of its contribution.

3. *Improved community understanding of Z-server performance issues in digital library , where 'clumping' approaches are being used.*

   Recall that **PW3** was an exploration of the concept of 'transparency' within distributed digital libraries and, in particular, the transparency of services offering federated search functionality (clumping). Its contribution was to refute prevailing thinking that the specific technical protocol (known as Z39.50) underpinning the federated searching functionality was inherently sluggish. Like

[1] Library Hub Discover, formerly known as Copac (the Consortium of Online Public Access Catalogues): `https://discover.libraryhub.jisc.ac.uk/`

PW2, **PW3** exposed further evidence that targets were deviating from the expectations of the cooperative, resulting in a series of performance issues when conducting federated searches. As the only large-scale study of its kind, this work contributed to the wider research agenda surrounding distributed digital libraries. In particular, it clearly reinforced the viability of Z39.50 based approaches to distributed digital library item-level retrieval. As **Chapter 3** reported, deployment of the protocol persists today and **PW3** has influenced community thinking about its viability in newer digital library applications, e.g. [98]. The work also informed the establishment of the aforementioned national metadata and Z-server management guidelines **[89].**

4. *Advanced our conceptual understanding of the resource discovery potential of KOS with specific reference to collaborative tagging systems.*

   At the emergence of collaborative tagging systems ('social bookmarking') there was a deficit in scholarly understanding of the potential efficacy of such approaches; yet, as **PW4** demonstrates, existing understanding of KOS could instead be used to assess the limitations of tagging by harnessing logic and theory. As a conceptual exercise to measure collaborative tagging as an effective knowledge organization mechanism, **PW4** is the most cited published work presented as part of this thesis, acquiring in excess of 440 citations and has ergo been highly influential on the study of tagging and in the evolution of tagging-based systems[2], even motivating and inspiring the creation of conceptual frameworks, e.g. [484].

5. *Influenced international approaches to terminology services for serving KOS-based data using Semantic Web standards and demonstrated uses for this data within a diverse range of resource discovery applications.*

   The influence of the present author on international approaches to terminology services within semantically aware information environments can be directly traced to **PW6**. This work, the first in a suite of related works discussed in **Chapter 4** (**PW7** & **PW8**), helped to develop emerging thinking on how semantically aware terminology services should behave in a web service context, how semantic data could be exposed and re-used by client services, and incorporated by clients to improve resource discovery for end users. The proposed system's design was elaborated in **PW7** and directly influenced developments in

[2] Citations for PW4, according to Google Scholar: `https://scholar.google.co.uk/scholar?oi=bibs&hl=en&cites=8504780136253658297`

similar projects, particularly in Germany (e.g. [485, 206, 486]) and in prototypes developed within domains as diverse as agriculture [221, 487] and biomedicine [488, 489] – the latter domain of which, as the present author reported in **Chapter 4**, is fertile ground for experimentation with terminology services. Both PW6 and PW7 formed the basis of a W3C Semantic Web Deployment use case study [193], used to inform emerging W3C Semantic Web standards, particularly contributing to the development of SKOS [194]. Furthermore, the technical outcome of this research (a prototype terminology service) was successfully incorporated by several academic search portals, including the Jisc 'intute'[3] and EDINA 'GeGeo'[4] services.

PW8 extended this further by providing a substantive contribution to the study of equivalence matching in KOS mapping. Its significance was a consequence of its scope, which examined mapping of *multiple* KOSs across an intermediary terminology 'spine' using a functioning prototype. Prior literature focussed on the mapping issues encountered when equivalence between two single KOSs was being established (e.g. [223, 224], so the research documented in **PW8** was unique. By testing mapping quality across a number of disparate KOS types the work has been cited in the evolution of terminology service requirements [218], mapping approaches [219, 220, 198, 221] and ontology mapping research [222].

Together, **PW6**, **PW7** and **PW8** constituted a phase of publishing which culminated in the present author's invitation to guest edit of a journal special issue[5] of 'Global Knowledge, Memory and Communication', therefore constituting a de facto contribution. The issue invited research articles exploring applications of Semantic Web technologies within digital library contexts [**490**].

6. *Evaluated a novel prototype tech-supported curriculum design system to facilitate the generation, reuse and discovery of curriculum designs and their associated data, as well as improved our understanding of the academic quality potential of such technologies.*

   Commentary included as part of **Chapter 5** highlighted the novelty of **PW9** and the wider innovation in tech-supported curriculum design tools or systems.

[3] intute: http://www.intute.ac.uk/

[4] EDINA: https://edina.ac.uk/

[5] Global Knowledge, Memory and Communication (formerly Library Review), *Special Issue: Digital libraries and the Semantic Web: context, applications and research*: `https://www.emerald.com/insight/publication/issn/0024-2535/vol/57/iss/3`

In particular, it was noted that **PW9** provided a significant contribution to discipline understanding of how users interacted with such tools, especially the interactions necessary to generate structured designs about curricula they might later deliver. This included exploration of issues surrounding the cognitive load exposed to academic users, engaging with an unfamiliar type of system in order to perform a cognitively onerous task.

The contribution of **PW9** is one primarily of uniqueness. No similar examples remain reported in the literature, and those prototype systems that are reported tend to be less sophisticated or remain unevaluated. The prototype system considered within **PW9** was more advanced and included innovative features, such as the storage of curriculum designs in a repository for discovery, cloning or re-use, the ability to model curriculum design metadata, wider information management capabilities to support academic administration, as well as academic quality management features. It is interesting to note that, following the work reported in **PW9** and its related work [**226, 227, 228, 229, 230, 231**], the prototype system was eventually adopted within the University of Strathclyde and reportedly remains in use. Documentation produced by the University of Strathclyde as part of the University's Quality Assurance Agency for Higher Education (Scotland)(QAA) *Enhancement-led Institutional Review 2019* noted a degree of institutional impact, reporting success in "capturing structured curriculum information and data" using the system [491].

7. *Delivered a series of unique contributions to open science community understanding in the discovery of open content and knowledge commons, the optimization of open repositories – and reasserted the need for open scholarly communications infrastructure.*

   Despite open repositories being important publishers of scholarly content within the global knowledge commons, **Chapter 6** established a surprising lack of community understanding about how repositories should be optimized in order to deliver demonstrable improvements in discovery potential. **PW10** and **PW11** established this lack of prior work in their research motivation, once again highlighting novelty as their principal contribution.

   However, both **PW10** and **PW11** have a relevance beyond our understanding of how open scholarly communications infrastructure operates and how repositories can feed content to the knowledge commons to also impinge on questions of research impact. As was noted earlier in **Chapter 6**, system support for academic research management and assessment has emerged in tandem with the proliferation of CRIS software. The concept of capturing 'impact' within the CERIF data model deployed by most CRIS platforms emphasizes its importance in research management thinking [492]; yet, these platforms remain ill equipped to deliver the visibility necessary to drive discovery and ergo citation or 'alternative' impact [493]. **PW10** and **PW11** therefore represents a significant contribution to discussions surrounding the software ecosystem surrounding open science (e.g. repositories) and research management (e.g. CRIS) — and how re-balancing is necessary to minimize system tension and ensure the objectives of each are fulfilled.

Both **PW10** and **PW11** are recent publications, making their contribution difficult to assess, but it can nevertheless be noted that both works have been influential within the relevant stakeholder communities, including within the blogosphere [494] and Twittersphere[6]. Even within the newly published literature, Walker [495] has cited the research of **PW10** as evidence that discoverability needs to be considered as part of strategic thinking about research impact. Arlitsch et al. [496] – the only other research group actively investigating questions of repository visibility and discoverability – have acknowledged findings from **PW10** in their methodological justifications. Given the recent publication date of PW11 (early 2020), it is perhaps too early to expect citations to be accrued.

9.2 Collective contribution to knowledge & the unifying theory

A contribution to knowledge beyond that presented in the assembled published works can also be identified as part of this thesis. This contribution was presented in **Chapter 8** and is the collective outcome of assessing all the assembled works:

- *Proposal of a unifying theory of resource discovery based on the concept of '(meta)data alignment' and an accompanying visual conceptual model.*

Fig. 9.1 diagrams the relationship between the contributions made by the works (as outlined above in **section 9.1**) and the unifying theory. In particular, the 'contributions to knowledge' node in Fig. 9.1 notes the chronological, self-seeding nature of the individual contributions but also their circular, recursive trajectory. As described in **section 8.2.2**, specific research concepts have travelled 'full circle' during

[6] Altmetric report for PW11: `https://www.altmetric.com/details/75112702`

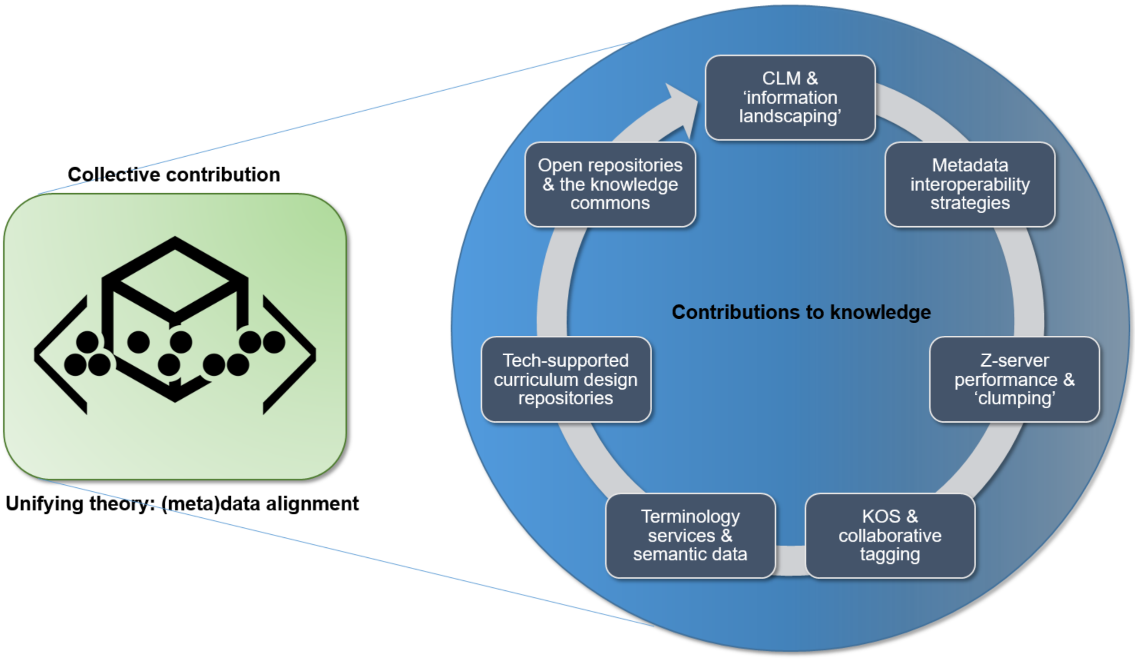

Collective contribution
Unifying theory: (meta)data alignment
Contributions to knowledge
CLM & 'information landscaping'
Metadata interoperability strategies
Z-server performance & 'clumping'
KOS & collaborative tagging
Terminology services & semantic data
Tech-supported curriculum design repositories
Open repositories & the knowledge commons

Figure 9.1: The present author's holistic contribution to knowledge, incorporating the unifying theory of resource discovery [(Meta)data alignment] as a collective outcome of the assembled works. 'Metadata' icon by Pascal Conil-lacoste from the Noun Project, `https://thenounproject.com` - CC-BY.

the present author's research journey, with topics surrounding CLM, for example – research topics from the very beginning of the present author's journey – re-emerging as an area worthy of new additional research but within a different scholarly context. In other words, the works assembled for this thesis, when critically assessed and their individual findings understood collectively, also provide a holistic contribution to knowledge.

This analysis assists in the identification of a number of research topics worthy of further research; but it also highlights an incremental transition by the present author, from using research to inform the development of technologies designed to support or facilitate resource discovery, particularly at a 'meta' level, to the application of specific technologies to address resource discovery issues in a local context. Despite this variation the research narrative has remained focussed on topics surrounding resource discovery in heterogeneous digital content environments and is noted as having generated a coherent body of work.

The unifying theory of research discovery presented in **section 8.2.2** and visualized in **Fig. 8.1** itself delivers a collective contribution to knowledge by:

1. Describing the inter-linked nature of the published works assembled for this thesis and establishing their coherence as a collective body of work.

2. Providing a model, based on the concept of *(meta)data alignment*, which can be transferred by others to conceptualize common resource discovery interactions in terms of layers, and thereby;
3. Enabling better understanding of the alignment necessary to ensure optimum syntactic and semantic interoperability across all discovery layers and within all data structures used by all participating system which, in turn, can deliver superior resource discovery experiences for end users.

9.3 Contribution to practice

When reflecting on the assembled works and their collective contribution to knowledge, it is possible to observe that their impact was, and is, not merely restricted to abstract scholarship or academia. Instead a distinct, ongoing and parallel contribution to professional practice can also be identified. This has been an important strand in the present author's research career and, as noted in the consideration of future research (Chapter 8), continues to influence the present author's research agenda. Suffice to state that the works have contributed to improved practice-based understanding of metadata interoperability issues within local digital libraries and repositories, especially via PW2 and PW3. The role of PW2 and PW3 in defining national interoperability and Z-server management recommendations was highlighted in section 9.1, impacting the work of practitioners in institutions that chose to adopt them but ensuring improved service provision for end users. But it could also be suggested that the works provided a key 'teachable moment' for practice, drawing attention to much needed improvements in digital library interoperability and promoting the concept of 'interoperability' as central to metadata management within increasingly distributed digital library environments.

There are works included within this thesis that, though contributing to the present author's body of research, are published in destinations designed to communicate to a predominantly practice-based community. This is especially true of PW4 and PW5. PW5, in particular, communicates theoretical concepts about potentially disruptive changes to resource discovery in the form of RDF, LOD and the Semantic Web; but is a chapter published within the 'The E-Resources Management Handbook' expressly to communicate with digital librarians and information science practitioners – thereby potentially influencing their future approaches to solving imminent resource discovery problems or expanding thinking around metadata modelling.

Perhaps more significantly, works published later in the chronology (PW10 and PW11) have contributed to discourse surrounding the management of open digital collections, particularly those contributing to the knowledge commons, such as open repositories. In other words, these works are examples of practice-based research designed expressly to inform those involved in the generation or management of the open knowledge commons, with the intended audience encompassing practitioner stakeholders such as system administrators, repository developers and scholarly communications librarians. Indeed, there remains huge potential for repository developments and system administrators to apply the research of PW10 and PW11 in their local context in order to improve users' resource discovery outcomes; but also to influence the technical development and future software releases of some of the most important digital library or repository platforms, including EPrints, DSpace, Samvera, and Invenio. PW10, in particular, describes a series of technical 'improvements' and 'adjustments' made to a test-case repository but which could easily be incorporated into most repository platforms so that they display the necessary behaviours out-of-the-box (OOTB). It also speaks to policy-makers responsible for steering the course of open science research infrastructure, including important standard setting bodies such as OpenAIRE[7] and COAR[8]; as well research funding bodies, many of which participate in the formulation of policy frameworks. Of relevance here are cOAlition S[9], the Wellcome Trust and UK Research & Innovation (UKRI).

[7] OpenAIRE: `https://www.openaire.eu/`

[8] Confederation of Open Access Repositories (COAR): `https://www.coar-repositories.org/`

[9] cOAlition S: `https://www.coalition-s.org/`

Chapter 10

Conclusion & preamble to published works

10.1 Concluding remarks

The research presented within the portfolio of published works assembled for this thesis chart an evolution in the research career of the present author. Chapters 3-6 have sought to critically comment upon and contextualize this body of work. In so doing the intellectual linkages within the presented works and across the extant literature have been critiqued and limitations of the works exposed. These chapters – and, of course, the works themselves – present from the earliest periods of the author's career, in which research was being undertaken with limited oversight or mentoring from more experienced senior colleagues. During this period methodological confidence was lacking and an immature academic writing style is clearly evident (e.g. PW2 & PW3). But, from PW4 onwards, it is possible to observe a gradual transition during which the reverse not only becomes evident, but is accompanied by superior conceptualization, data analysis and reasoning. This methodological evolution was examined closely in Chapter 7 and revealed a definite philosophical transition to pragmatism and a commitment to mixed-methods where practicable.

When examining the chronology of the works it is possible to observe a notable feature of the present author's research journey: an incremental transition from using research to inform the development or building of technologies to support or facilitate resource discovery, particularly at a 'meta' level (e.g. distributed or federated solutions), to the application of specific technologies to address resource discovery issues in a local context. To this extent it could be suggested that the present author's research journey has been characterized by a journey from 'abstract research' to 'practice-based research'. This partly reflects the career path of the present author:

from research assistant, research fellow and then lecturer, to a practitioner operating within the knowledge commons. Despite this the research narrative has remained focussed on topics surrounding resource discovery in heterogeneous digital content environments and has generated a coherent body of work. As was reflected in the model of *(meta)data alignment* presented in Fig. 8.1 – and which this coherent body of work has facilitated – the assembled works chart a research narrative across a variety of resource discovery service layers, with earlier published works principally concerned with distributed service issues within resource discovery contexts (PW1, PW2 & PW3) and addressing problems of semantic interoperability (PW4, PW5, PW6, PW7 & PW8); while later works continue the narrative but within service and discovery layers (PW9, PW10 & P11).

Although Chapter 9 highlights the numerous contributions to academic knowledge and professional practice that the assembled works have made, Chapter 8 also demonstrates that there remains a series of research trajectories of both an academic and practice-based nature that deserve future investigation, such is the ongoing nature of research enquiry. The model of *(meta)data alignment* provides a meaningful conceptual model onto which the present author's future research ambitions can be attached and maintain theoretical coherence.

Disconnection between research and practice is an important and significant concluding observation to be made. Throughout this thesis – and in particular the critical commentary of the assembled works – there have been examples of a failure to implement the knowledge findings from research in practice, or to transfer lessons from one context to another. This has resulted in a repetition of failure in the quality of the resource discovery experiences delivered to end-users and poor service outcomes. This is especially evident in the management of distributed digital libraries and repositories which are members of system cooperatives, where recurring inadequacies in the configuration of machine endpoints and the semantic interoperability of metadata exposed by these endpoints remains evident.

10.2 Preamble to Published Works

The published works assembled for this thesis reflect the research evolution of the present author. This evolution has not only explored various aspects of resource discovery vis-à-vis *(meta)data alignment* while demonstrating a collective coherence; but has also revealed a research progression which has matured over time and displayed increased levels of methodological sophistication.

All of the **11 published works** have been reproduced in **Appendix B**, prefaced by a numbered legend to aid readers locate specific works.

References

[1] G. Dunsire and G. Macgregor, "Clumps and collection description in the information environment in the UK with particular reference to Scotland," *Program*, vol. 37, pp. 218–225, 2003. [Online]. Available: https://doi.org/10.1108/00330330310500694

[2] Cetis, "XCRI Github – Cetis LLP Standards Support," 2014. [Online]. Available: http://cetis.github.io/xcri/

[3] T. F. Berestova, "The concept of information resources and other components of the theory of information-resource science," *Scientific and Technical Information Processing*, vol. 43, no. 2, pp. 83–87, Apr. 2016. [Online]. Available: https://doi.org/10.3103/S0147688216020027

[4] M. Courtney, "Discovery tools," in *Reimagining Reference in the 21st Century*. Purdue University Press, pp. 121–131.

[5] C. A. Lynch, "Networked information resource discovery: an overview of current issues," vol. 13, no. 8, pp. 1505–1522, conference Name: IEEE Journal on Selected Areas in Communications. [Online]. Available: https://doi.org/10.1109/49.464719

[6] M. Bowman, P. B. Danzig, U. Manber, and M. F. Schwartz, "Scalable internet resource discovery: Research problems and approaches ; CU-CS-679-93." [Online]. Available: https://scholar.colorado.edu/concern/reports/rx913q821

[7] M. D. Smucker, "Information representation," in *Interactive Information Seeking, Behaviour and Retrieval*, I. Ruthven and D. Kelly, Eds. London: Facet, 2011, pp. 77–93.

[8] J. J. Eaton and D. Bawden, "What kind of resource is information?" *International Journal of Information Management*, vol. 11, no. 2, pp. 156–165, Jun. 1991. [Online]. Available: https://doi.org/10.1016/0268-4012(91)90006-X

[9] W. M. Beyene, “Resource Discovery and Universal Access: Understanding Enablers and Barriers from the User Perspective,” *Studies in Health Technology and Informatics*, vol. 229, pp. 556–566, 2016. [Online]. Available: http://hdl.handle.net/10642/4604

[10] R. Bruce and A. McGregor, “Resource discovery,” in *Trends, Discovery, and People in the Digital Age*, ser. Chandos Digital Information Review, D. Baker and W. Evans, Eds. Chandos Publishing, Jan. 2013, pp. 105–122. [Online]. Available: https://doi.org/10.1016/B978-1-84334-723-1.50007-3

[11] J. Walker, “New resource discovery mechanisms (2),” in *The E-Resources Management Handbook*. UKSG, 2006, pp. 1–11. [Online]. Available: https://doi.org/10.1629/9552448-0-3.8.2

[12] T. Wilson, “On user studies and information needs,” *Journal of Documentation*, vol. 37, no. 1, pp. 3–15, Jan. 1981. [Online]. Available: https://doi.org/10.1108/eb026702

[13] A. Rakotonirainy and G. Groves, “Resource Discovery for Pervasive Environments,” in *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, ser. Lecture Notes in Computer Science, R. Meersman and Z. Tari, Eds. Springer Berlin Heidelberg, 2002, pp. 866–883. [Online]. Available: https://doi.org/10.1007/3-540-36124-3_57

[14] A. Jaimes, D. Gatica-Perez, N. Sebe, and T. S. Huang, “Introduction: Human-Centered Computing–Toward a Human Revolution,” *Computer*, vol. 40, no. 5, pp. 30–34, May 2007.

[15] L. Bowler, H. Julien, and L. Haddon, “Exploring youth information-seeking behaviour and mobile technologies through a secondary analysis of qualitative data,” *Journal of Librarianship and Information Science*, vol. 50, pp. 322–331, Sep. 2018. [Online]. Available: https://doi.org/10.1177/0961000618769967

[16] E. Greifeneder, “Trends in information behaviour research,” *Information Research*, vol. 19, no. 4, Dec. 2014, proceedings of ISIC: the information behaviour conference, Leeds, 2-5 September, 2014: Part 1. [Online]. Available: https://curis.ku.dk/ws/files/137513587/Trends_in_information_behaviour_research.htm

[17] A. Spink and C. Cole, "Human Information Behavior: Integrating Diverse Approaches and Information Use," *J. Am. Soc. Inf. Sci. Technol.*, vol. 57, no. 1, pp. 25–35, Jan. 2006. [Online]. Available: https://doi.org/10.1002/asi.v57:1

[18] H. Weber, D. Becker, and S. Hillmert, "Information-seeking behaviour and academic success in higher education: Which search strategies matter for grade differences among university students and how does this relevance differ by field of study?" *Higher Education*, vol. 77, no. 4, pp. 657–678, Apr. 2019. [Online]. Available: https://doi.org/10.1007/s10734-018-0296-4

[19] I. Rowlands, D. Nicholas, P. Williams, P. Huntington, M. Fieldhouse, B. Gunter, R. Withey, H. R. Jamali, T. Dobrowolski, and C. Tenopir, "The Google generation: the information behaviour of the researcher of the future," *Aslib Proceedings*, vol. 60, no. 4, pp. 290–310, Jul. 2008. [Online]. Available: https://doi.org/10.1108/00012530810887953

[20] K. Vanthournout, G. Deconinck, and R. Belmans, "A Taxonomy for Resource Discovery," *Personal Ubiquitous Comput.*, vol. 9, no. 2, pp. 81–89, Mar. 2005. [Online]. Available: https://doi.org/10.1007/s00779-004-0312-9

[21] G. Macgregor, "Principles in Patterns (PiP) : User Acceptance Testing of Course and Class Approval Online Pilot (C-CAP)," Strathprints, University of Strathclyde, Glasgow, Report, Feb. 2012. [Online]. Available: https://strathprints.strath.ac.uk/46510/

[22] R. Mandala, T. Tokunaga, and H. Tanaka, "Query expansion using heterogeneous thesauri," *Information Processing & Management*, vol. 36, no. 3, pp. 361–378, May 2000. [Online]. Available: https://doi.org/10.1016/S0306-4573(99)00068-0

[23] G. Macgregor, "Collection-level description: metadata of the future?" *Library Review*, vol. 52, pp. 247–250, 2003. [Online]. Available: https://doi.org/10.1108/00242530310482015

[24] M. Brenner, T. Larsen, and C. Weston, "Digital Collection Management through the Library Catalog," *Information Technology and Libraries*, Jun. 2006. [Online]. Available: https://pdxscholar.library.pdx.edu/ulib_fac/27

[25] G. Dunsire, “Landscaping the future for collaborative collection management,” in *World Library and Information Congress: 73rd IFLA General Conference and Council.* Durban: IFLA, 2007. [Online]. Available: https://strathprints.strath.ac.uk/6027/

[26] H.-L. Lee, “What is a collection?” *Journal of the American Society for Information Science*, vol. 51, no. 12, pp. 1106–1113, 2000. [Online]. Available: https://doi.org/10.1002/1097-4571(2000)9999:9999⟨::AID-ASI1018⟩3.0.CO;2-T

[27] M. Foulonneau, T. W. Cole, T. G. Habing, and S. L. Shreeves, “Using collection descriptions to enhance an aggregation of harvested item-level metadata,” in *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL ’05).* New York, NY, USA: ACM, Jun. 2005, pp. 32–41. [Online]. Available: https://doi.org/10.1145/1065385.1065393

[28] K. Fenlon, P. Organisciak, J. Jett, and M. Efron, “Semi-automated collection evaluation for large-scale aggregations,” *Proceedings of the American Society for Information Science and Technology*, vol. 48, no. 1, pp. 1–3, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/meet.2011.14504801319

[29] I. Lourdi, C. Papatheodorou, and M. Doerr, “Semantic Integration of Collection Description: Combining CIDOC/CRM and Dublin Core Collections Application Profile,” *D-Lib Magazine*, vol. 15, no. 7/8, Jul. 2009. [Online]. Available: https://doi.org/10.1045/july2009-papatheodorou

[30] I. Lourdi and C. Papatheodorou, “A metadata application profile for collection-level description of digital folklore resources,” in *Proceedings. 15th International Workshop on Database and Expert Systems Applications, 2004.*, Sep. 2004, pp. 90–94.

[31] M. Note, “Metadata and information management,” in *Managing Image Collections*, ser. Chandos Information Professional Series, M. Note, Ed. Chandos Publishing, Jan. 2011, pp. 107–133. [Online]. Available: https://doi.org/10.1016/B978-1-84334-599-2.50005-2

[32] J.-r. Park, “Semantic Interoperability across Digital Image Collections: Evaluation of Metadata Mapping for Resource Discovery and Sharing,”

Proceedings of the Annual Conference of CAIS / Actes du congrès annuel de l'ACSI, vol. 0, no. 0, Oct. 2013. [Online]. Available: https://journals.library.ualberta.ca/ojs.cais-acsi.ca/index.php/cais-asci/article/view/303

[33] K. M. Wickett, A. Isaac, M. Doerr, K. Fenlon, C. Meghini, and C. Palmer, "Representing Cultural Collections in Digital Aggregation and Exchange Environments," *D-Lib Magazine*, vol. 20, no. 5/6, May 2014. [Online]. Available: https://doi.org/10.1045/may2014-wickett

[34] O. L. Zavalina, "Collection-Level Subject Access in Aggregations of Digital Collections: Metadata Application and Use," Ph.D. dissertation, University of Illinois at Urbana-Champaign, Urbana, Illinois, 2010. [Online]. Available: http://hdl.handle.net/2142/16620

[35] ——, "Contextual Metadata in Digital Aggregations: Application of Collection-Level Subject Metadata and Its Role in User Interactions and Information Retrieval," *Journal of Library Metadata*, vol. 11, no. 3-4, pp. 104–128, Jul. 2011. [Online]. Available: https://doi.org/10.1080/19386389.2011.629957

[36] B. Stvilia and C. Jörgensen, "User-generated collection-level metadata in an online photo-sharing system," *Library & Information Science Research*, vol. 31, no. 1, pp. 54–65, Jan. 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0740818808001370

[37] G. Therrell, "More product, more process: metadata in digital image collections," *Digital Library Perspectives*, vol. 35, no. 1, pp. 2–14, Feb. 2019. [Online]. Available: https://doi.org/10.1108/DLP-06-2018-0018

[38] A. Apps, "IESR: A Registry of Collections and Services." The Hague: MIMAS, Apr. 2006. [Online]. Available: http://epub.mimas.ac.uk/papers/srug2006/apps-srug2006_summary.html

[39] ——, "Using an Application Profile Based Service Registry," *International Conference on Dublin Core and Metadata Applications*, vol. 0, no. 0, pp. 63–73, Aug. 2007. [Online]. Available: http://dcpapers.dublincore.org/pubs/article/view/867

[40] L. L. Hill, G. Janée, R. Dolin, J. Frew, and M. Larsgaard, "Collection metadata solutions for digital library applications," *Journal of the American Society for Information Science*, vol. 50, no. 13, pp. 1169–1181, 1999. [Online]. Available:

https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%281999%2950%3A13%3C1169%3A%3AAID-ASI3%3E3.0.CO%3B2-3

[41] M. Heaney, “An Analytical Model of Collections and their Catalogues,” University of Bath, Bath, Tech. Rep., 2000. [Online]. Available: http://www.ukoln.ac.uk/metadata/rslp/model/

[42] ——, “An Extension of the Analytical Model of Collections and their Catalogues into Usage and Transactions,” University of Bath, Bath, Text, Jun. 2001. [Online]. Available: http://www.ukoln.ac.uk/cd-focus/model-ext/

[43] A. Powell, M. Heaney, and L. Dempsey, “RSLP Collection Description,” *D-Lib Magazine*, vol. 6, no. 9, Sep. 2000. [Online]. Available: http://www.dlib.org/dlib/september00/powell/09powell.html

[44] D. C. C. D. T. Group, “DCMI: Dublin Core Collection Description Terms,” 2013. [Online]. Available: https://www.dublincore.org/specifications/dublin-core/collection-description/collection-terms/

[45] A. Chapman, “Collection-level description: Joining up the domains,” *Journal of the Society of Archivists*, vol. 25, no. 2, pp. 149–155, Oct. 2004. [Online]. Available: https://doi.org/10.1080/0037981042000271475

[46] G. Dunsire, “Extending the SCONE Collection Descriptions Database for CC-interop : Report for Work Package B of the CC-interop JISC Project,” University of Strathclyde, Glasgow, Report, Oct. 2002. [Online]. Available: https://strathprints.strath.ac.uk/68642/

[47] ——, “The Collection Description Schema Forum,” *Ariadne*, no. 39, 2004. [Online]. Available: http://www.ariadne.ac.uk/issue/39/cdfocus-schema-rpt/

[48] ——, “Development of a relational database schema for collection-level descriptions in SCONE, the Scottish Collections Network,” 2004. [Online]. Available: https://strathprints.strath.ac.uk/3174/

[49] ——, “Collection-level descriptions in the Scottish Collections Network (SCONE),” University of Strathclyde, Glasgow, Tech. Rep., 2004. [Online]. Available: http://hdl.handle.net/10760/5890

[50] Europeana Foundation, "Europeana Data Model," 2019. [Online]. Available: https://pro.europeana.eu/resources/standardization-tools/edm-documentation

[51] H.-h. Chen, C.-m. Tsai, and Y.-c. Ho, "Landscaping Taiwan's Cultural Heritages – The Implementation of the TELDAP Collection-Level Description," in *The Role of Digital Libraries in a Time of Global Change*, ser. Lecture Notes in Computer Science, G. Chowdhury, C. Koo, and J. Hunter, Eds. Springer Berlin Heidelberg, 2010, pp. 130–139.

[52] K. Friday, "Learning from e-family history: online research behaviour and strategies of family historians and implications for local studies collections." Ph.D. dissertation, Robert Gordon University, Aberdeen, May 2012. [Online]. Available: http://hdl.handle.net/10059/734

[53] M. Zani, "Granularità: un percorso di analisi," *DigItalia*, vol. 2, no. 0, pp. 60–128, Apr. 2006. [Online]. Available: http://digitalia.sbn.it/article/view/302

[54] H.-W. Lee, "A Study on the Model of Collection-Level Description based on Ontology for Resources Sharing," *Journal of the Korean Society for information Management*, vol. 25, no. 3, pp. 209–230, 2008. [Online]. Available: http://www.koreascience.or.kr/article/JAKO200831852745546.page

[55] G. Dunsire, "Collection landscaping in the common information environment: a case study using the Scottish Collections Network (SCONE) : report for work package B of the JISC CC-interop project - E-LIS repository," University of Strathclyde, Glasgow, Tech. Rep., 2004. [Online]. Available: http://hdl.handle.net/10760/5887

[56] ——, "Conspectus and the Scottish Collections Network: landscaping the Scottish common information environment," *Signum*, vol. 3, pp. 20–27, 2006. [Online]. Available: http://pro.tsv.fi/stks/signum/200603/4.pdf

[57] D. Nicholson, G. Dunsire, and G. Macgregor, "SPEIR: developing a common information environment in Scotland," *Electronic Library*, vol. 24, pp. 94–107, 2006. [Online]. Available: https://doi.org/10.1108/02640470610649272

[58] C. L. Palmer, E. M. Knutson, M. Twidale, and O. Zavalina, "Collection Definition in Federated Digital Resource Development," *Proceedings of the*

American Society for Information Science and Technology, vol. 43, no. 1, pp. 1–16, 2006. [Online]. Available: https://doi.org/10.1002/meet.14504301161

[59] C. L. Palmer, O. L. Zavalina, and M. Mustafoff, "Trends in Metadata Practices: A Longitudinal Study of Collection Federation," in *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL '07. New York, NY, USA: ACM, 2007, pp. 386–395, event-place: Vancouver, BC, Canada. [Online]. Available: http://doi.acm.org/10.1145/1255175.1255251

[60] K. Wickett, "A logic-based framework for collection/item metadata relationships," *Journal of Documentation*, Oct. 2018. [Online]. Available: https://doi.org/10.1108/JD-01-2018-0017

[61] C. L. Borgman, "The Digital Future is Now: A Call to Action for the Humanities," *Digital Humanities Quarterly*, vol. 3, no. 4, Jan. 2010. [Online]. Available: https://escholarship.org/uc/item/0fp9n05s

[62] Y.-N. Chen, "A RDF-based approach to metadata crosswalk for semantic interoperability at the data element level," *Library Hi Tech*, vol. 33, no. 2, pp. 175–194, Jun. 2015. [Online]. Available: https://www.emeraldinsight.com/doi/full/10.1108/LHT-08-2014-0078

[63] N. Piedra, J. Chicaiza, J. Lopez-Vargas, and E. T. Caro, "Guidelines to producing structured interoperable data from Open Access Repositories," in *2016 IEEE Frontiers in Education Conference (FIE)*, Oct. 2016, pp. 1–9.

[64] S. G. Roy, B. Sutradhar, and P. P. Das, "Large-scale Metadata Harvesting—Tools, Techniques and Challenges: A Case Study of National Digital Library (NDL)," *World Digital Libraries - An international journal*, vol. 10, no. 1, pp. 1–10, Jan. 2017. [Online]. Available: https://content.iospress.com/articles/world-digital-libraries-an-international-journal/wdl10101

[65] H. Suleman, "The design abd architecture of digital libraries," in *Digital Libraries and Information Access: Research Perspectives*, G. G. Chowdhury and S. Foo, Eds. London: Facet Publishing, Sep. 2012.

[66] J. Gilby, "Hyper Clumps, Mini Clumps and National Catalogues: Resource Discovery for the 21st Century," *Ariadne*, no. 42, 2005. [Online]. Available: http://www.ariadne.ac.uk/issue/42/cc-interops-rpt/

[67] ANSI/NISO, *ANSI/NISO Z39.50-2003 (S2014) Information Retrieval: Application Service Definition & Protocol Specification.* Baltimore, MD.: National Information Standards Organization, 2015. [Online]. Available: https://www.niso.org/publications/ansiniso-z3950-2003-s2014

[68] Library of Congress, "Z39.50 Gateway," 2018. [Online]. Available: https://www.loc.gov/z3950/

[69] C. Pautasso, O. Zimmermann, and F. Leymann, "Restful Web Services vs. "Big"' Web Services: Making the Right Architectural Decision," in *Proceedings of the 17th International Conference on World Wide Web*, ser. WWW '08. New York, NY, USA: ACM, 2008, pp. 805–814, event-place: Beijing, China. [Online]. Available: http://doi.acm.org/10.1145/1367497.1367606

[70] Index Data, "YAZ [toolkit supporting the development of Z39.50/SRW/SRU clients and servers]," 2019. [Online]. Available: https://perma.cc/3YRH-DUCJ

[71] B. Bradley, "Collaborating for access: Implementing aeon's openurl in our discovery system," May 2018. [Online]. Available: http://hdl.handle.net/1903/20588

[72] E. M. Corrado, "Discovery Products and the Open Archives Initiative Protocol for Metadata Harvesting," *International Information & Library Review*, vol. 50, no. 1, pp. 47–53, Jan. 2018. [Online]. Available: https://doi.org/10.1080/10572317.2017.1422905

[73] M. D'Ambrosio, "The IRIS Consortium (Florence, Italy) and RDA: Perspectives and Possibilities from a Loyal (Non!) Opposition," *JLIS.it*, vol. 9, no. 1, Jan. 2018. [Online]. Available: https://www.jlis.it/article/view/61-65

[74] D. B. Krstićev, "Information Retrieval Using a Middleware Approach," *Information Technology and Libraries*, vol. 32, no. 1, pp. 54–69, Mar. 2013. [Online]. Available: https://ejournals.bc.edu/index.php/ital/article/view/1941

[75] A. Rego Máñez, L. García-García, M. Llopis, and J. Lloret, "A new Z39.50 protocol client to search in libraries and improve research collaboration," in *Network Protocols and Algorithms*, vol. 8. Macrothink Institute, Oct. 2016, pp. 29–54. [Online]. Available: https://riunet.upv.es/handle/10251/82477

[76] Y. Guo, Z. Yu, Y. Men, and X. Xu, "Data Sharing of Power Grid Meteorological Disaster based on Metadata," *IOP Conference Series: Materials Science and Engineering*, vol. 466, p. 012028, Dec. 2018. [Online]. Available: https://doi.org/10.1088%2F1757-899x%2F466%2F1%2F012028

[77] S. B. Shirude and S. R. Kolhe, "Agent-Based Architecture for Developing Recommender System in Libraries," in *Knowledge Computing and its Applications: Knowledge Computing in Specific Domains: Volume II*, S. Margret Anouncia and U. K. Wiil, Eds. Singapore: Springer Singapore, 2018, pp. 157–181. [Online]. Available: https://doi.org/10.1007/978-981-10-8258-0_8

[78] G. Macgregor and F. Nicolaides, "Towards improved performance and interoperability in distributed and physical union catalogues," *Program*, vol. 39, pp. 227–247, 2005. [Online]. Available: https://doi.org/10.1108/00330330510610573

[79] A. S. Tanenbaum, *Distributed systems: principles and paradigms*, 2nd ed. Upper Saddle RIiver, NJ: Pearson Prentice Hall, 2007.

[80] OCLC, "Bath Profile compliance checklist," Aug. 2018. [Online]. Available: https://help.oclc.org/Discovery_and_Reference/FirstSearch/Z3950_access/Bath_Profile_compliance_checklist

[81] G. E. Gorman and P. Clayton, *Qualitative Research for the Information Professional*, 2nd ed. London: Facet Publishing, 2006. [Online]. Available: https://www.dawsonera.com/abstract/9781856047982

[82] J. Saldana, *The Coding Manual for Qualitative Researchers*, 2nd ed. SAGE, 2016, google-Books-ID: ZhxiCgAAQBAJ.

[83] V. Elliott, "Thinking about the Coding Process in Qualitative Data Analysis," *The Qualitative Report*, vol. 23, no. 11, pp. 2850–2861, Nov. 2018. [Online]. Available: https://nsuworks.nova.edu/tqr/vol23/iss11/14

[84] H. Guetzkow, "Unitizing and categorizing problems in coding qualitative data," *Journal of Clinical Psychology*, vol. 6, no. 1, pp. 47–58, 1950. [Online]. Available: https://doi.org/10.1002/1097-4679(195001)6:1⟨47::AID-JCLP2270060111⟩3.0.CO;2-I

[85] B. G. Glaser, *The discovery of grounded theory : strategies for qualitative research*, ser. Strategies for qualitative research, A. L. Strauss, Ed. New Brunswick, N.J.: New Brunswick, N.J. : Aldine Transaction, 1999.

[86] G. Dunsire, "Joined up indexes: interoperability issues in Z39.50 networks," *International Cataloguing and Bibliographic Control*, vol. 32, pp. 47–49, 2003. [Online]. Available: https://strathprints.strath.ac.uk/2329/

[87] P. Hider, "The bibliographic advantages of a centralised union catalogue for ILL and resource sharing," *Interlending & Document Supply*, vol. 32, no. 1, pp. 17–29, Mar. 2004. [Online]. Available: https://doi.org/10.1108/02641610410520224

[88] W. E. Moen, "Assessing Interoperability in the Networked Environment: Standards, Evaluation, and Testbeds in the Context of Z39.50," 2001. [Online]. Available: https://digital.library.unt.edu/ark:/67531/metadc102281/

[89] G. Dunsire and G. Macgregor, "Improving Interoperability in Distributed and Physical Union Catalogues through Co-ordination of Cataloguing and Indexing Policies : Report for Work Package B of the JISC CC-interop Project," University of Strathclyde, Glasgow, Report, May 2004. [Online]. Available: https://strathprints.strath.ac.uk/57653/

[90] S. J. Heron, B. Simpson, A. K. Weiss, and J. Phillips, "Merging Catalogs: Creating a Shared Bibliographic Environment for the State University Libraries of Florida," *Cataloging & Classification Quarterly*, vol. 51, no. 1-3, pp. 139–155, Jan. 2013. [Online]. Available: https://doi.org/10.1080/01639374.2012.722591

[91] A. S. Jackson, M.-J. Han, K. Groetsch, M. Mustafoff, and T. W. Cole, "Dublin Core Metadata Harvested Through OAI-PMH," *Journal of Library Metadata*, vol. 8, no. 1, pp. 5–21, Apr. 2008. [Online]. Available: https://doi.org/10.1300/J517v08n01_02

[92] P. Knoth, M. Cancellieri, M. Klein, and H. Van de Sompel, "Evaluating the Performance of OAI-PMH and ResourceSync," Sep. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1433822

[93] G. Macgregor, "Z39.50 broadcast searching and Z-server response times: perspectives from CC-interop," *Online Information Review*, vol. 29, pp. 90–106, 2005. [Online]. Available: https://doi.org/10.1108/14684520510583963

[94] ISO, "ISO/IEC 10746-3:2009," 2009. [Online]. Available: http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/57/55724.html

[95] P. Stubley, R. Bull, and T. Kidd, "Feasibility Study for a National Union Catalogue," Jisc, London, Tech. Rep., 2001. [Online]. Available: https://www.suncat.ac.uk/description/SUNCAT-NUCrep.pdf

[96] H. Booth and R. J. Hartley, "User behaviour in the searching of union catalogues: an investigation for work package C of CC-interop," CERLIM, Manchester, Report, Jan. 2004. [Online]. Available: https://e-space.mmu.ac.uk/1374/

[97] R. J. Hartley and H. Booth, "Users and union catalogues," *Journal of Librarianship and Information Science*, vol. 38, no. 1, pp. 7–20, Mar. 2006. [Online]. Available: https://doi.org/10.1177/0961000606060956

[98] S. Kapidakis and M. Sfakakis, "Eliminating query failures in a work-centric library meta-search environment," *Library Hi Tech*, vol. 27, no. 2, pp. 286–307, Jun. 2009. [Online]. Available: https://doi.org/10.1108/07378830910968236

[99] M. L. Zeng and L. M. Chan, "Trends and issues in establishing interoperability among knowledge organization systems," *Journal of the American Society for Information Science and Technology*, vol. 55, no. 5, pp. 377–395, 2004. [Online]. Available: https://doi.org/10.1002/asi.10387

[100] F. W. Lancaster, *Vocabulary control for information retrieval*, 2nd ed. Arlington, Va: Information Resources Press, 1986.

[101] S. R. Ranganathan, *Prolegomena to library classification.* London: Madras Library Association, Madras, 1937. [Online]. Available: http://hdl.handle.net/10973/19232

[102] B. C. Vickery, "The structure of information retrieval systems," in *Proceedings of the International Conference on Scientific Information*, vol. 2. USA: National Academy of Sciences, 1959, pp. 1275–1290.

[103] G. Macgregor and E. McCulloch, "Collaborative tagging as a knowledge organisation and resource discovery tool," *Library Review*, vol. 55, pp. 291–300, 2006. [Online]. Available: https://doi.org/10.1108/00242530610667558

[104] M. L. Zeng and G. Hodge, “Developing a Dublin Core Application Profile for the knowledge organization systems (KOS) resources,” *Bulletin of the American Society for Information Science and Technology*, vol. 37, no. 4, pp. 30–34, 2011. [Online]. Available: https://doi.org/10.1002/bult.2011.1720370409

[105] M. Zeng, M. Hlava, J. Qin, G. Hodge, and D. Bedford, “Knowledge organization systems (KOS) standards,” *Proceedings of the American Society for Information Science and Technology*, vol. 44, no. 1, pp. 1–3, 2007. [Online]. Available: https://doi.org/10.1002/meet.145044019

[106] R. R. Souza, D. Tudhope, and a. M. B. Almeida, “Towards a Taxonomy of KOS: Dimensions for Classifying Knowledge Organization Systems,” *KNOWLEDGE ORGANIZATION*, vol. 39, no. 3, pp. 179–192, 2012. [Online]. Available: https://doi.org/10.5771/0943-7444-2012-3-179

[107] C. Binding and D. Tudhope, “KOS at your Service: Programmatic Access to Knowledge Organisation Systems,” *Journal of Digital Information*, vol. 4, no. 4, Feb. 2006. [Online]. Available: https://journals.tdl.org/jodi/index.php/jodi/article/view/110

[108] P. A. Gaona-García, D. Martín-Moncunill, E. E. Gaona-García, A. Gómez-Acosta, and C. Monenegro-Marin, “Usability of Big Data Resources in Visual Search Interfaces of Repositories Based on KOS,” in *Proceedings of the 2018 2Nd International Conference on Cloud and Big Data Computing*, ser. ICCBDC’18. New York, NY, USA: ACM, 2018, pp. 33–37, event-place: Barcelona, Spain. [Online]. Available: http://doi.acm.org/10.1145/3264560.3264567

[109] J. Greenberg, “Introduction: Knowledge organization innovation: Design and frameworks,” *Bulletin of the American Society for Information Science and Technology*, vol. 37, no. 4, pp. 12–14, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/bult.2011.1720370404

[110] L. L. Hill, B. Olha, G. Janée, and Z. Lei, “Integration of Knowledge Organization Systems into Digital Library Architectures,” *Data Analysis and Knowledge Discovery*, vol. 20, no. 1, pp. 4–8, Jan. 2004. [Online]. Available: https://doi..org/10.11925/infotech.1003-3513.2004.01.02

[111] D. Nicholson and E. McCulloch, “HILT Phase III : design requirements of an SRW-compliant terminologies mapping pilot,” in *5th European Networked*

Knowledge Organization Systems (NKOS) Workshop. Alicante: ECDL, Sep. 2006. [Online]. Available: https://strathprints.strath.ac.uk/2320/

[112] A. Shiri and S. Chase-Kruszewski, "Knowledge organisation systems in North American digital library collections," *Program*, Apr. 2009. [Online]. Available: https://doi.org/10.1108/00330330910954352

[113] A. Shiri and K. Molberg, "Interfaces to knowledge organization systems in Canadian digital library collections," *Online Information Review*, Dec. 2005. [Online]. Available: https://doi.org/10.1108/14684520510638061

[114] D. Soergel, "Digital Libraries and Knowledge Organization," in *Semantic Digital Libraries*, S. R. Kruk and B. McDaniel, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 9–39. [Online]. Available: https://doi.org/10.1007/978-3-540-85434-0_2

[115] R. Szostak, "The importance of knowledge organization," *Bulletin of the Association for Information Science and Technology*, vol. 40, no. 4, pp. 37–42, 2014. [Online]. Available: https://doi.org/10.1002/bult.2014.1720400414

[116] M.-C. Tang, "Browsing and searching in a faceted information space: A naturalistic study of PubMed users' interaction with a display tool," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 13, pp. 1998–2006, 2007. [Online]. Available: https://doi.org/10.1002/asi.20689

[117] D. Tudhope, C. Binding, S. Jeffrey, K. May, and A. Vlachidis, "A STELLAR role for knowledge organization systems in digital archaeology," *Bulletin of the American Society for Information Science and Technology*, vol. 37, no. 4, pp. 15–18, 2011. [Online]. Available: https://doi.org/10.1002/bult.2011.1720370405

[118] J. Walsh, "The use of Library of Congress Subject Headings in digital collections," *Library Review*, Apr. 2011. [Online]. Available: https://doi.org/10.1108/00242531111127875

[119] M. L. Zeng, M. Hlava, J. A. Busch, O. Buchel, and M. Žumer, "If You Build It, Will They Come?: A Discussion of Use Cases and Barriers of Using the Knowledge Organization Systems (KOS) Available As Linked Open Data (LOD)," in *Proceedings of the 78th ASIS&T Annual Meeting: Information Science with Impact: Research in and for the Community*, ser. ASIST '15. Silver Springs, MD, USA: American Society for Information

Science, 2015, pp. 3:1–3:3, event-place: St. Louis, Missouri. [Online]. Available: http://dl.acm.org/citation.cfm?id=2857070.2857073

[120] D. Nicholas, P. Huntington, P. Williams, and T. Dobrowolski, "Re-appraising information seeking behaviour in a digital environment," *Journal of Documentation*, Feb. 2004. [Online]. Available: https://doi.org/10.1108/00220410410516635

[121] D. Nicholas, P. Huntington, H. R. Jamali, and T. Dobrowolski, "Characterising and evaluating information seeking behaviour in a digital environment: Spotlight on the 'bouncer'," *Information Processing & Management*, vol. 43, no. 4, pp. 1085–1102, Jul. 2007. [Online]. Available: https://doi.org/10.1016/j.ipm.2006.08.007

[122] H. R. Jamali and D. Nicholas, "Information-seeking behaviour of physicists and astronomers," *Aslib Proceedings*, Sep. 2008. [Online]. Available: https://doi.org/10.1108/00012530810908184

[123] R. W. White and S. M. Drucker, "Investigating Behavioral Variability in Web Search," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 21–30, event-place: Banff, Alberta, Canada. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242576

[124] E. N. Efthimiadis, "Query Expansion," in *Annual Review of Information Science and Technology (ARIST)*, 1996, vol. 31, pp. 121–187.

[125] Y. Gavel and P.-O. Andersson, "Multilingual query expansion in the SveMed+ bibliographic database: A case study," *Journal of Information Science*, vol. 40, no. 3, pp. 269–280, Jun. 2014. [Online]. Available: https://doi.org/10.1177/0165551514524685

[126] N. Griffon, W. Chebil, L. Rollin, G. Kerdelhue, B. Thirion, J.-F. Gehanno, and S. J. Darmoni, "Performance evaluation of unified medical language system®'s synonyms expansion to query PubMed," *BMC Medical Informatics and Decision Making*, vol. 12, no. 1, p. 12, Feb. 2012. [Online]. Available: https://doi.org/10.1186/1472-6947-12-12

[127] Z. Lu, W. Kim, and W. J. Wilbur, "Evaluation of Query Expansion Using MeSH in PubMed," *Information Retrieval*, vol. 12, no. 1, pp. 69–80, 2009. [Online]. Available: https://doi.org/10.1007/s10791-008-9074-8

[128] K. Kapoor, B. R. Green, Y. Ye, and Y. Li, "Systems and methods for automated query expansion," US Patent US20 190 155 929A1, May, 2019. [Online]. Available: https://patents.google.com/patent/US20190155929A1/en

[129] W. Selmi, H. Kammoun, and I. Amous, "MeSH-Based Semantic Query Expansion," in *2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, Oct. 2018, pp. 1–8.

[130] A. A. Shiri and C. Revie, "The effects of topic complexity and familiarity on cognitive and physical moves in a thesaurus-enhanced search environment," *Journal of Information Science*, vol. 29, no. 6, pp. 517–526, Nov. 2003. [Online]. Available: https://doi.org/10.1177/0165551503296008

[131] A. Shiri and C. Revie, "Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 4, pp. 462–478, 2006. [Online]. Available: https://doi.org/10.1002/asi.20319

[132] M. L. Zeng and L. M. Chan, "Trends and issues in establishing interoperability among knowledge organization systems," *Journal of the American Society for Information Science and Technology*, vol. 55, no. 5, pp. 377–395, 2004. [Online]. Available: https://doi.org/10.1002/asi.10387

[133] G. Macgregor, A. Joseph, and D. Nicholson, "A SKOS Core approach to implementing an M2M terminology mapping server," in *International Conference on Semantic Web and Digital Libraries (ICSD-2007)*, Feb. 2007, pp. 109–120. [Online]. Available: https://strathprints.strath.ac.uk/2970/

[134] C. Cole, "A theory of information need for information retrieval that connects information to knowledge," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 7, pp. 1216–1231, 2011. [Online]. Available: https://doi.org/10.1002/asi.21541

[135] J. C. dos Reis, C. Pruski, M. Da Silveira, and C. Reynaud-Delaître, "Understanding semantic mapping evolution by observing changes in

biomedical ontologies," *Journal of Biomedical Informatics*, vol. 47, pp. 71–82, Feb. 2014. [Online]. Available: https://doi.org/10.1016/j.jbi.2013.09.006

[136] J. D. Tenenbaum, P. L. Whetzel, K. Anderson, C. D. Borromeo, I. D. Dinov, D. Gabriel, B. Kirschner, B. Mirel, T. Morris, N. Noy, C. Nyulas, D. Rubenson, P. R. Saxman, H. Singh, N. Whelan, Z. Wright, B. D. Athey, M. J. Becich, G. S. Ginsburg, M. A. Musen, K. A. Smith, A. F. Tarantal, D. L. Rubin, and P. Lyster, "The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research," *Journal of Biomedical Informatics*, vol. 44, no. 1, pp. 137–145, Feb. 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046410001553

[137] M. Torjmen-Khemakhem and K. Gasmi, "Document/query expansion based on selecting significant concepts for context based retrieval of medical images," *Journal of Biomedical Informatics*, vol. 95, p. 103210, Jul. 2019. [Online]. Available: https://doi.org/10.1016/j.jbi.2019.103210

[138] T. O'Reilly, *What is Web 2.0.* "O'Reilly Media, Inc.", Sep. 2009, google-Books-ID: NpEk_WFCMdIC.

[139] C. Shirky, "Ontology is Overrated – Categories, Links, and Tags," 2005. [Online]. Available: https://perma.cc/9ESH-V2YE

[140] ——, "The Semantic Web, Syllogism, and Worldview," 2003. [Online]. Available: http://www.shirky.com/writings/herecomeseverybody/semantic_syllogism.html

[141] Z. Xu, Y. Fu, J. Mao, and D. Su, "Towards the Semantic Web: Collaborative Tag Suggestions," in *Collaborative web tagging workshop at WWW2006*. Edinburgh: ACM, 2006. [Online]. Available: http://ra.ethz.ch/CDstore/www2006/www.rawsugar.com/www2006/13.pdf

[142] A. Ankolekar, M. Krötzsch, T. Tran, and D. Vrandecic, "The Two Cultures: Mashing Up Web 2.0 and the Semantic Web," in *Proceedings of the 16th International Conference on World Wide Web*, ser. WWW '07. New York, NY, USA: ACM, 2007, pp. 825–834, event-place: Banff, Alberta, Canada. [Online]. Available: http://doi.acm.org/10.1145/1242572.1242684

[143] F. Dotsika and K. Patrick, "Towards the new generation of web knowledge," *VINE Journal of Information and Knowledge Management Systems*, vol. 36, no. 4, pp. 406–422, Oct. 2006. [Online]. Available: https://doi.org/10.1108/03055720610716665

[144] J. Hendler and J. Golbeck, "Metcalfe's Law, Web 2.0, and the Semantic Web," *Web Semant.*, vol. 6, no. 1, pp. 14–20, Feb. 2008. [Online]. Available: https://doi.org/10.1016/j.websem.2007.11.008

[145] N. Shadbolt, T. Berners-Lee, J. Hendler, C. Hart, and R. Benjamins, "The Next Wave of the Web," in *Proceedings of the 15th International Conference on World Wide Web*, ser. WWW '06. New York, NY, USA: ACM, 2006, pp. 750–750, event-place: Edinburgh, Scotland. [Online]. Available: http://doi.acm.org/10.1145/1135777.1135889

[146] L. Specia and E. Motta, "Integrating Folksonomies with the Semantic Web," in *The Semantic Web: Research and Applications*, ser. Lecture Notes in Computer Science, E. Franconi, M. Kifer, and W. May, Eds. Springer Berlin Heidelberg, 2007, pp. 624–639. [Online]. Available: https://doi.org/10.1007/978-3-540-72667-8_44

[147] H. Dong, W. Wang, and H. Liang, "Learning Structured Knowledge from Social Tagging Data: A Critical Review of Methods and Techniques," in *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, Dec. 2015, pp. 307–314.

[148] J. Lorince, K. Joseph, and P. M. Todd, "Analysis of Music Tagging and Listening Patterns: Do Tags Really Function as Retrieval Aids?" in *Social Computing, Behavioral-Cultural Modeling, and Prediction*, ser. Lecture Notes in Computer Science, N. Agarwal, K. Xu, and N. Osgood, Eds. Springer International Publishing, 2015, pp. 141–152. [Online]. Available: https://doi.org/10.1007/978-3-319-16268-3_15

[149] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme, "Evaluating Similarity Measures for Emergent Semantics of Social Tagging," in *Proceedings of the 18th International Conference on World Wide Web*, ser. WWW '09. New York, NY, USA: ACM, 2009, pp. 641–650, event-place: Madrid, Spain. [Online]. Available: http://doi.acm.org/10.1145/1526709.1526796

[150] A. Passant, "Laublet P.: Meaning Of A Tag: A collaborative approach to bridge the gap between tagging and Linked Data," in *Proceedings of the Linked Data on the Web (LDOW2008) workshop at WWW2008*, 2008. [Online]. Available: http://events.linkeddata.org/ldow2008/papers/22-passant-laublet-meaning-of-a-tag.pdf

[151] A. X. Zhang and J. Cranshaw, "Making Sense of Group Chat Through Collaborative Tagging and Summarization," *Proc. ACM Hum.-Comput. Interact.*, vol. 2, no. CSCW, pp. 196:1–196:27, Nov. 2018. [Online]. Available: https://doi.org/10.1145/3274465

[152] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, and D. Vrandečić, "Introducing Wikidata to the Linked Data Web," in *The Semantic Web – ISWC 2014*, ser. Lecture Notes in Computer Science, P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. Knoblock, D. Vrandečić, P. Groth, N. Noy, K. Janowicz, and C. Goble, Eds. Springer International Publishing, 2014, pp. 50–65. [Online]. Available: https://doi.org/10.1007/978-3-319-11964-9_4

[153] T. v. Veen, "Wikidata:," *Information Technology and Libraries*, vol. 38, no. 2, pp. 72–81, Jun. 2019. [Online]. Available: https://doi.org/10.6017/ital.v38i2.10886

[154] C. J. Godby and K. Smith-Yoshimura, "From Records to Things: Managing the Transition from Legacy Library Metadata to Linked Data," *Bulletin of the Association for Information Science and Technology*, vol. 43, no. 2, pp. 18–23, Jan. 2017. [Online]. Available: https://doi.org/10.1002/bul2.2017.1720430209

[155] M. Klein and A. Kyrios, "VIAFbot and the Integration of Library Data on Wikipedia," *The Code4Lib Journal*, no. 22, Oct. 2013. [Online]. Available: http://journal.code4lib.org/articles/8964

[156] S. Allison-Cassin and D. Scott, "Wikidata: a platform for your library's linked open data," *The Code4Lib Journal*, no. 40, May 2018. [Online]. Available: http://journal.code4lib.org/articles/13424

[157] J. Voß, "Classification of Knowledge Organization Systems with Wikidata," in *Proceedings of the 15th European Networked Knowledge Organization Systems Workshop (NKOS 2016)*. Hannover, Germany: CEUR Workshop Proceedings, 2016, pp. 15–22. [Online]. Available: http://ceur-ws.org/Vol-1676/paper2.pdf

[158] P. Bartley, “Book tagging on LibraryThing: How, why, and what are in the tags?” *Proceedings of the American Society for Information Science and Technology*, vol. 46, no. 1, pp. 1–22, Jan. 2009. [Online]. Available: https://doi.org/10.1002/meet.2009.1450460228

[159] M. Harvey, I. Ruthven, and M. Carman, “Ranking social bookmarks using topic models,” in *19th ACM international conference on Information and knowledge management*, Oct. 2010, pp. 1401–1404. [Online]. Available: https://doi.org/10.1145/1871437.1871632

[160] H.-J. Lee and D. Neal, “A new model for semantic photograph description combining basic levels and user-assigned descriptors,” *Journal of Information Science*, vol. 36, no. 5, pp. 547–565, Oct. 2010. [Online]. Available: https://doi.org/10.1177/0165551510374930

[161] W. G. Stock, “Folksonomies and science communication: A mash-up of professional science databases and Web 2.0 services,” *Information Services & Use*, vol. 27, no. 3, pp. 97–103, Jul. 2007. [Online]. Available: https://doi.org/10.3233/ISU-2007-27303

[162] I. Peters and W. G. Stock, ““Power tags” in information retrieval,” *Library Hi Tech*, Mar. 2010. [Online]. Available: https://doi.org/10.1108/07378831011026706

[163] F. M. Suchanek, M. Vojnovic, and D. Gunawardena, “Social Tags: Meaning and Suggestions,” in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ser. CIKM ’08. New York, NY, USA: ACM, 2008, pp. 223–232, event-place: Napa Valley, California, USA. [Online]. Available: http://doi.acm.org/10.1145/1458082.1458114

[164] M. Guy and E. Tonkin, “Folksonomies: Tidying up Tags?” *D-Lib Magazine*, vol. 12, no. 1, Jan. 2006. [Online]. Available: https://doi.org/10.1045/january2006-guy

[165] I. Peters, *Folksonomies. Indexing and Retrieval in Web 2.0.* Berlin, Boston: De Gruyter Saur, 2009. [Online]. Available: https://www.degruyter.com/view/product/42362

[166] I. Peters, L. Schumann, J. Terliesner, and W. G. Stock, "Retrieval effectiveness of tagging systems," *Proceedings of the American Society for Information Science and Technology*, vol. 48, no. 1, pp. 1–4, 2011. [Online]. Available: https://doi.org/10.1002/meet.2011.14504801338

[167] P. J. Rolla, "User Tags versus Subject Headings," *Library Resources & Technical Services*, vol. 53, no. 3, pp. 174–184, Apr. 2011. [Online]. Available: https://journals.ala.org/index.php/lrts/article/view/5281

[168] R. G. Henzler, "Free or controlled vocabularies," *KNOWLEDGE ORGANIZATION*, vol. 5, no. 1, pp. 21–26, 1978. [Online]. Available: https://doi.org/10.5771/0943-7444-1978-1-21

[169] J. Rowley, "The controlled versus natural indexing languages debate revisited: a perspective on information retrieval practice and research," *Journal of Information Science*, vol. 20, no. 2, pp. 108–118, Apr. 1994. [Online]. Available: https://doi.org/10.1177/016555159402000204

[170] F. Mazzochhi, "Knowledge organization system (KOS)," Kent, OH., 2019. [Online]. Available: https://perma.cc/VT89-76LE

[171] M. Doerr and D. Iorizzo, "The Dream of a Global Knowledge Network—A New Approach," *J. Comput. Cult. Herit.*, vol. 1, no. 1, pp. 5:1–5:23, Jun. 2008. [Online]. Available: http://doi.acm.org/10.1145/1367080.1367085

[172] M. L. Zeng and P. Mayr, "Knowledge Organization Systems (KOS) in the Semantic Web: a multi-dimensional review," *International Journal on Digital Libraries*, May 2018. [Online]. Available: https://doi.org/10.1007/s00799-018-0241-2

[173] D. Tudhope, T. Koch, and R. Heery, "Terminology Services and Technology. JISC state of the art review," UKOLN, Bath, Tech. Rep., 2006. [Online]. Available: http://www.ukoln.ac.uk/terminology/JISC-review2006.html

[174] K. Golub, D. Tudhope, M. L. Zeng, and M. Žumer, "Terminology registries for knowledge organization systems: Functionality, use, and attributes," *Journal of the Association for Information Science and Technology*, vol. 65, no. 9, pp. 1901–1916, 2014. [Online]. Available: https://doi.org/10.1002/asi.23090

[175] R. Cornet and A. K. Prins, “An Architecture for Standardized Terminology Services by Wrapping and Integration of Existing Applications,” *AMIA Annual Symposium Proceedings*, vol. 2003, p. 180, 2003. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480276/

[176] D. Gardner, D. H. Goldberg, B. Grafstein, A. Robert, and E. P. Gardner, “Terminology for Neuroscience Data Discovery: Multi-tree Syntax and Investigator-Derived Semantics,” *Neuroinformatics*, vol. 6, no. 3, pp. 161–174, Sep. 2008. [Online]. Available: https://doi.org/10.1007/s12021-008-9029-7

[177] K. Kawamoto and D. F. Lobach, “Design, Implementation, Use, and Preliminary Evaluation of an UMLS-Enabled Terminology Web Service for Clinical Decision Support,” *AMIA Annual Symposium Proceedings*, vol. 2006, p. 979, 2006. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1839661/

[178] H. Kittler, A. A. Marghoob, G. Argenziano, C. Carrera, C. Curiel-Lewandrowski, R. Hofmann-Wellenhof, J. Malvehy, S. Menzies, S. Puig, H. Rabinovitz, W. Stolz, T. Saida, H. P. Soyer, E. Siegel, W. V. Stoecker, A. Scope, M. Tanaka, L. Thomas, P. Tschandl, I. Zalaudek, and A. Halpern, “Standardization of terminology in dermoscopy/dermatoscopy: Results of the third consensus conference of the International Society of Dermoscopy,” *Journal of the American Academy of Dermatology*, vol. 74, no. 6, pp. 1093–1106, Jun. 2016. [Online]. Available: https://doi.org/10.1016/j.jaad.2015.12.038

[179] A. Metke-Jimenez, J. Steel, D. Hansen, and M. Lawley, “Ontoserver: a syndicated terminology server,” *Journal of Biomedical Semantics*, vol. 9, no. 1, p. 24, Sep. 2018. [Online]. Available: https://doi.org/10.1186/s13326-018-0191-z

[180] J. Pathak, H. R. Solbrig, J. D. Buntrock, T. M. Johnson, and C. G. Chute, “LexGrid: A Framework for Representing, Storing, and Querying Biomedical Terminologies from Simple to Sublime,” *Journal of the American Medical Informatics Association*, vol. 16, no. 3, pp. 305–315, May 2009. [Online]. Available: https://doi.org/10.1197/jamia.M3006

[181] K. J. Peterson, G. Jiang, S. M. Brue, and H. Liu, “Leveraging Terminology Services for Extract-Transform-Load Processes: A User-Centered Approach,” *AMIA Annual Symposium Proceedings*, vol. 2016, pp. 1010–1019,

Feb. 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333225/

[182] F. G. B. d. Quirós, C. Otero, and D. Luna, “Terminology Services: Standard Terminologies to Control Health Vocabulary,” *Yearbook of Medical Informatics*, vol. 27, no. 1, pp. 227–233, Aug. 2018. [Online]. Available: https://doi.org/10.1055/s-0038-1641200

[183] C. Tao, J. Pathak, H. R. Solbrig, W.-Q. Wei, and C. G. Chute, “Terminology representation guidelines for biomedical ontologies in the semantic web notations,” *Journal of Biomedical Informatics*, vol. 46, no. 1, pp. 128–138, Feb. 2013. [Online]. Available: https://doi.org/10.1016/j.jbi.2012.09.003

[184] G. Macgregor, E. McCulloch, and D. Nicholson, “Terminology server for improved resource discovery: analysis of model and functions,” in *Second International Conference on Metadata and Semantics Research*, Oct. 2007. [Online]. Available: http://strathprints.strath.ac.uk/3435/

[185] E. McCulloch and G. Macgregor, “Analysis of equivalence mapping for terminology services,” *Journal of Information Science*, vol. 34, pp. 70–92, Feb. 2008. [Online]. Available: http://dx.doi.org/10.1177/0165551507079130

[186] A. Chapman and R. Russell, “JISC Shared Infrastructure Services Synthesis Study: A review of the shared infrastructure for the JISC Information Environment,” Jisc, London, Tech. Rep., Sep. 2006. [Online]. Available: https://researchportal.bath.ac.uk/files/455759/jisc-sis-report-final.pdf

[187] G. Macgregor, “High-Level Thesaurus (HILT) project, phase IV,” in *JISC Shared Infrastructure Services Workshop*, GBR, Jun. 2007. [Online]. Available: https://strathprints.strath.ac.uk/71129/

[188] A. Powell, “A ‘service oriented’ view of the JISC Information Environment,” UKOLN, Bath, Tech. Rep., 2005. [Online]. Available: http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/soa/jisc-ie-soa.pdf

[189] D. Nicholson, E. McCulloch, and G. Macgregor, “High-level Thesaurus (HILT) Phase III [Project] : Final Report,” University of Strathclyde, Glasgow, Report, Feb. 2007. [Online]. Available: https://strathprints.strath.ac.uk/71139/

[190] I. Data, “Zthes - The Zthes specifications for thesaurus representation, access and navigation,” 2006. [Online]. Available: http://zthes.z3950.org/

[191] A. Miles, B. Matthews, M. Wilson, and D. Brickley, “SKOS Core: Simple knowledge organisation for the Web,” *International Conference on Dublin Core and Metadata Applications*, vol. 0, no. 0, pp. 3–10, Sep. 2005. [Online]. Available: http://dcpapers.dublincore.org/pubs/article/view/798

[192] “W3C SKOS Use Cases and Requirements,” 2009. [Online]. Available: https://www.w3.org/TR/skos-ucr/

[193] G. Macgregor, E. McCulloch, and D. Nicholson, “RucHilt - W3C Semantic Web Deployment Wiki,” 2007. [Online]. Available: https://www.w3.org/2006/07/SWD/wiki/RucHilt.html

[194] “SKOS Simple Knowledge Organization System Reference,” 2009. [Online]. Available: https://www.w3.org/TR/2009/REC-skos-reference-20090818/

[195] N. Chergui, S. Chikhi, and T. Kechadi, “Semantic Grid Resource Discovery Based on SKOS Ontology,” *Int. J. Grid Util. Comput.*, vol. 8, no. 4, pp. 269–281, 2017. [Online]. Available: https://doi.org/10.1504/IJGUC.2017.088255

[196] S. J. D. Cox, J. Yu, and T. Rankine, “SISSVoc: A Linked Data API for access to SKOS vocabularies,” *Semantic Web*, vol. 7, no. 1, pp. 9–24, Jan. 2016. [Online]. Available: http://www.semantic-web-journal.net/system/files/swj880.pdf

[197] N. Freire, R. Voorburg, R. Cornelissen, S. de Valk, E. Meijers, and A. Isaac, “Aggregation of Linked Data in the Cultural Heritage Domain: A Case Study in the Europeana Network,” *Information*, vol. 10, no. 8, p. 252, Aug. 2019. [Online]. Available: https://doi.org/10.3390/info10080252

[198] A. Isaac, S. Schlobach, H. Matthezing, and C. Zinn, “Integrated access to cultural heritage resources through representation and alignment of controlled vocabularies,” *Library Review*, vol. 57, no. 3, pp. 187–199, Mar. 2008. [Online]. Available: https://doi.org/10.1108/00242530810865475

[199] A. Isaac and B. Haslhofer, “Europeana Linked Open Data – data.europeana.eu,” *Semantic Web*, vol. 4, no. 3, pp. 291–297, Jan. 2013. [Online]. Available: https://eprints.cs.univie.ac.at/3732/

[200] E. Summers, A. Isaac, C. Redding, and D. Krech, "LCSH, SKOS and Linked Data," *International Conference on Dublin Core and Metadata Applications*, vol. 0, no. 0, pp. 25–33, Sep. 2008. [Online]. Available: http://dcpapers.dublincore.org/pubs/article/view/916

[201] G. Macgregor, "E-resource management and the Semantic Web : applications of RDF for e-resource discovery," in *The E-Resources Management Handbook - UKSG*. Newbury: UKSG, 2009, pp. 1–20. [Online]. Available: https://doi.org/10.1629/9552448-0-3.20.1

[202] D. U. Board, "DCMI: Dublin Core Metadata Element Set, Version 1.1: Reference Description," Dublin Core Metadata Initiative, Ohio, Tech. Rep., 2012. [Online]. Available: https://www.dublincore.org/specifications/dublin-core/dces/

[203] D. Brickley and L. Miller, "FOAF Vocabulary Specification," xmlns.com, Bristol, Tech. Rep., 2014. [Online]. Available: http://xmlns.com/foaf/spec/

[204] B. Adida, M. Birbeck, S. McCarron, and S. Pemberton, "RDFa in XHTML: Syntax and Processing," W3C, Cambridge, Massachusetts, Tech. Rep., 2008. [Online]. Available: https://www.w3.org/TR/rdfa-syntax/

[205] W. O. W. Group, "OWL 2 Web Ontology Language Document Overview (Second Edition)," W3C, Cambridge, Massachusetts, Tech. Rep., 2012. [Online]. Available: https://www.w3.org/TR/owl2-overview/

[206] F. Boteram, ""Content architecture" : Semantic interoperability in an international comprehensive knowledge organisation system," *Aslib Proceedings*, vol. 62, no. 4/5, pp. 406–414, Jul. 2010. [Online]. Available: https://doi.org/10.1108/00012531011074654

[207] L. Eric Si, A. O'Brien, and S. Probets, "Integration of distributed terminology resources to facilitate subject cross-browsing for library portal systems," *Aslib Proceedings*, vol. 62, no. 4/5, pp. 415–427, Jan. 2010. [Online]. Available: https://doi.org/10.1108/00012531011074663

[208] W. Gödert, "Ontological spine, localization and multilingual access: some reflections and a proposal," in *New Perspectives on Subject Indexing and Classification: Essays in Honour of Magda Heiner-Freiling.*

Leipzig: Deutsche Nationalbibliothek, 2008, pp. 233–240. [Online]. Available: http://ixtrieve.fh-koeln.de/crisscross/publikationen/goedert_frankfurt08.pdf

[209] L. G. Svensson, "National libraries and the Semantic Web: requirements and applications," in *International Conference on Semantic Web and Digital Libraries (ICSD-2007)*. Bangalore, India: Indian Statistical Institute in Bangalore, 2007.

[210] Y. Zhang, J. Peng, D. Huang, and F. Li, "Design of Automatic Mapping System between DDC and CLC," in *Digital Libraries: For Cultural Heritage, Knowledge Dissemination, and Future Creation*, ser. Lecture Notes in Computer Science, C. Xing, F. Crestani, and A. Rauber, Eds. Berlin, Heidelberg: Springer, 2011, pp. 357–366.

[211] E. McCulloch, A. Shiri, and D. Nicholson, "Challenges and issues in terminology mapping : a digital library perspective," *Electronic Library*, vol. 23, pp. 671–677, 2005. [Online]. Available: https://doi.org/10.1108/02640470510635755

[212] M. Doerr, "Semantic Problems of Thesaurus Mapping," *Journal of Digital Information*, vol. 1, no. 8, Jan. 2006. [Online]. Available: https://journals.tdl.org/jodi/index.php/jodi/article/view/31

[213] J. P. Silvester and P. H. Klingbiel, "An operational system for subject switching between controlled vocabularies," *Information Processing & Management*, vol. 29, no. 1, pp. 47–59, Jan. 1993. [Online]. Available: https://doi.org/10.1016/0306-4573(93)90022-6

[214] M. Panzer, "Cool URIs for the DDC: Towards Web-scale Accessibility of a Large Classification System," *International Conference on Dublin Core and Metadata Applications*, vol. 0, no. 0, pp. 183–190, Sep. 2008. [Online]. Available: http://dcpapers.dublincore.org/pubs/article/view/932

[215] D. M. Nicholson, A. Dawson, and A. Shiri, "HILT : a terminology mapping service with a DDC spine," *Cataloging and Classification Quarterly*, vol. 42, pp. 187–200, Oct. 2006. [Online]. Available: https://doi.org/10.1300/J104v42n03_08

[216] J.-E. Mai, "The Future of General Classification," *Cataloging & Classification Quarterly*, vol. 37, no. 1-2, pp. 3–12, Jul. 2003. [Online]. Available: https://doi.org/10.1300/J104v37n01_02

[217] M. A. Chaplan, “Mapping ”Laborline Thesaurus” Terms to Library of Congress Subject Headings: Implications for Vocabulary Switching,” *The Library Quarterly*, vol. 65, no. 1, pp. 39–61, Jan. 1995. [Online]. Available: https://doi.org/10.1086/602752

[218] D. Nicholson, “A Common Research and Development Agenda for Subject Interoperability Services?” *Signum*, vol. 36, no. 5, 2008. [Online]. Available: https://journal.fi/signum/article/view/3484

[219] S.-j. Chen and H.-h. Chen, “Mapping multilingual lexical semantics for knowledge organization systems,” *The Electronic Library*, vol. 30, no. 2, pp. 278–294, Apr. 2012. [Online]. Available: https://doi.org/10.1108/02640471211221386

[220] A. Gray, N. Gray, A. Millar, and I. Ounis, “Semantically enabled vocabularies in astronomy,” University of Glasgow, Glasgow, Tech. Rep., 2007. [Online]. Available: http://www.academia.edu/download/30680839/vocabulariesInAstronomy.pdf

[221] T. Zschocke, “Resolving controlled vocabulary in DITA markup: a case example in agroforestry,” *Program*, vol. 46, no. 3, pp. 321–340, Jul. 2012. [Online]. Available: https://doi.org/10.1108/00330331211244869

[222] R. Hoekstra, “BestMap: context-aware SKOS vocabulary mappings in OWL 2,” *CEUR Workshop Proceedings*, vol. 529, 2009. [Online]. Available: https://hdl.handle.net/11245/1.315856

[223] A. C. Liang and M. Sini, “Mapping AGROVOC and the Chinese Agricultural Thesaurus: Definitions, tools, procedures,” *New Review of Hypermedia and Multimedia*, vol. 12, no. 1, pp. 51–62, Jun. 2006. [Online]. Available: https://doi.org/10.1080/13614560600774396

[224] A. Liang, M. Sini, C. B. Chun, L. Sijing, L. Wenlin, H. Chunpei, and J. Keizer, “The Mapping Schema from Chinese Agricultural Thesaurus to AGROVOC.” Rome: FAO, 2005. [Online]. Available: http://www.fao.org/3/a-af241e.pdf

[225] J. Everett, G. Macgregor, and R. Mohamed, “An incremental approach to technology-supported curriculum design and approval,” in *IADIS International Conference WWW/Internet 2012*, Oct. 2012. [Online]. Available: https://strathprints.strath.ac.uk/46292/

[226] G. Macgregor, “C-CAP : Managing Curriculum Designs as Knowledge Assets - Briefing Paper,” University of Strathclyde, Glasgow, Report, Oct. 2012. [Online]. Available: https://strathprints.strath.ac.uk/59851/

[227] ——, “Principles in Patterns (PiP) : Evaluation of Impact on Business Processes,” University of Strathclyde, Glasgow, Report, Apr. 2012. [Online]. Available: https://strathprints.strath.ac.uk/46512/

[228] ——, “Principles in Patterns (PiP) : Project Evaluation Synthesis,” University of Strathclyde, Glasgow, Report, Jul. 2012. [Online]. Available: https://strathprints.strath.ac.uk/46513/

[229] ——, “Principles in Patterns (PiP) : Evaluation Approach,” London, Apr. 2012. [Online]. Available: https://strathprints.strath.ac.uk/59936/

[230] ——, “Principles in Patterns (PiP) : Heuristic Evaluation of Course and Class Approval Online Pilot (C-CAP),” University of Strathclyde, Glasgow, Report, Dec. 2011. [Online]. Available: https://strathprints.strath.ac.uk/46509/

[231] ——, “Principles in Patterns (PiP) : Piloting of C-CAP - Evaluation of Impact and Implications for System and Process Development,” University of Strathclyde, Glasgow, Report, Jun. 2012. [Online]. Available: https://strathprints.strath.ac.uk/46511/

[232] S. Agrawal, P. B. Sharma, and M. Kumar, “Knowledge Management Framework for Improving Curriculum Development Processes in Technical Education,” in *2008 Third International Conference on Convergence and Hybrid Information Technology*, vol. 2, Nov. 2008, pp. 885–890. [Online]. Available: http://doi.org/10.1109/ICCIT.2008.339

[233] J. B. Biggs, *Teaching for quality learning at university.*, 3rd ed. Maidenhead: Open University Press, 2007.

[234] P. Bartholomew and J. Everett, “Socio-technical ramifications of a new approach to course design and approval.” London: Jisc, 2011. [Online]. Available: http://jiscdesignstudio.pbworks.com/w/page/49099346/Socio-technical%20ramifications%20of%20a%20new%20approach%20to%20course%20design%20and%20approval

[235] Jisc, "Enhancing Curriculum Design with Technology : Outcomes from the Jisc Institutional Approaches to Curriculum Design Programme," Jisc, London, Tech. Rep., 2013. [Online]. Available: https://www.jisc.ac.uk/sites/default/files/enhancing-curriculum-design.pdf

[236] J. F. Scholl, "Using Technology to Improve Curriculum Development," *The Technology Source*, no. 182, 2001. [Online]. Available: http://horizon.unc.edu/TS/editor/182.html

[237] M. Godsk, "Efficient learning design - concept, catalyst, and cases." Dunedin, New Zealand: Australasian Society for Computers in Learning in Tertiary Education, 2014. [Online]. Available: http://ascilite.org/conferences/dunedin2014/files/fullpapers/146-Godsk.pdf

[238] G. Conole and J. Culver, "Cloudworks: Social networking for learning design," *Australasian Journal of Educational Technology*, vol. 25, no. 5, Nov. 2009. [Online]. Available: https://doi.org/10.14742/ajet.1120

[239] L. Kolås and A. Staupe, "Implementing delivery methods by using pedagogical design patterns." Association for the Advancement of Computing in Education (AACE), 2004, pp. 5304–5309. [Online]. Available: https://www.learntechlib.org/primary/p/11834/

[240] H. Beetham, "Institutional Approaches to Curriculum Design : Final Synthesis Report," Jisc, London, Tech. Rep., 2012. [Online]. Available: http://jiscdesignstudio.pbworks.com/w/file/fetch/61216296/JISC%20Curriculum%20Design%20Final%20Synthesis%20i1.pdf

[241] I. Cameron and G. Birkett, "A curriculum design, modelling and visualization environment," *23rd Annual Conference of the Australasian Association for Engineering Education 2012: Profession of Engineering Education: Advancing Teaching, Research and Careers, The*, p. 301, 2012. [Online]. Available: http://search.informit.com.au/documentSummary;dn=235355615977200;res=IELENG

[242] S. Cross, R. Galley, A. Brasher, and M. Weller, "OULDI-JISC Project Evaluation Report: the impact of new curriulum design tools and approaches on institutional process and design cultures," Jul. 2012. [Online]. Available: http://oro.open.ac.uk/34140/

[243] G. Dafoulas, B. Barn, and Y. Zheng, "Improving student employability by utilising semantic analysis of course data," Nov. 2014. [Online]. Available: http://eprints.mdx.ac.uk/16061/

[244] P. M. Parker and S. Quinsee, "Facilitating Institutional Curriculum Change in Higher Education," *International Journal of Learning*, vol. 18, pp. 49–60, 2012. [Online]. Available: http://openaccess.city.ac.uk/id/eprint/1762/

[245] P. Range and M. Stubbs, "Service-oriented architecture and curriculum transformation at Manchester Metropolitan University," *Campus-Wide Information Systems*, vol. 28, no. 4, pp. 299–304, Aug. 2011. [Online]. Available: https://www.emeraldinsight.com/doi/full/10.1108/10650741111162770

[246] R. M. Branch, *Instructional Design: The ADDIE Approach.* Springer Science & Business Media, Sep. 2009, google-Books-ID: mHSwJPE099EC. [Online]. Available: https://doi.org/10.1007/978-0-387-09506-6

[247] J. J. G. v. Merriënboer, P. A. Kirschner, and P. A. Kirschner, *Ten Steps to Complex Learning : A Systematic Approach to Four-Component Instructional Design.* London: Routledge, Oct. 2017. [Online]. Available: https://doi.org/10.4324/9781315113210

[248] G. Conole, "Learning design – making practice explicit," Sydney, Australia, Jun. 2010. [Online]. Available: http://cloudworks.ac.uk/cloud/view/4001

[249] D. Nicol, D. McDonald, C. Owen, J. Everett, and D. Cullen, "The Curriculum Design and Approval Process at the University of Strathclyde : Baseline of Processes and Curriculum Design Activities," University of Strathclyde, Glasgow, Report, Sep. 2009. [Online]. Available: https://strathprints.strath.ac.uk/70296/

[250] G. Conole, "An overview of design representations," in *Proceedings of the 7th International Conference on Networked Learning 2010.* Aalborg: Lancaster University, 2010, pp. 482–489. [Online]. Available: https://www.lancaster.ac.uk/fss/organisations/netlc/past/nlc2010/

[251] M. Stubbs and S. Wilson, "eXchanging Course-Related Information: a UK service-oriented approach," Oct. 2006. [Online]. Available: https://core.ac.uk/download/pdf/55533856.pdf

[252] BSI, *BS EN 15982:2011 - Metadata for learning opportunities (MLO). Advertising.* London: BSI, 2011. [Online]. Available: https://shop.bsigroup.com/ProductDetail/?pid=000000000030204827

[253] E. C. for Standardization, "CWA 15903 : Metadata for Learning Opportunities (MLO) - Advertising," CEN, Brussels, Tech. Rep., 2008. [Online]. Available: https://www.immagic.com/eLibrary/ARCHIVES/TECH/CEN_EU/C081208O.pdf

[254] V. Bell, "Testing the XCRI-CAP Standard on Course Advertising Information at the University of Worcester," Santiago de Compostella, Jun. 2009. [Online]. Available: http://eprints.worc.ac.uk/649/

[255] G. Dafoulas, B. Barn, and Y. Zheng, "Curriculum design tools: Using information modelling for course transformation and mapping," in *2012 International Conference on Information Technology Based Higher Education and Training (ITHET)*, Jun. 2012, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ITHET.2012.6246064

[256] L. McGill, "eXchanging course related information – XCRI timeline," Mar. 2012. [Online]. Available: http://blogs.cetis.org.uk/othervoices/2012/03/13/exchanging-course-related-information-xcri-timeline/

[257] A. Paull, "Consuming XCRI-CAP II: XCRI eXchange Platform (XXP)," Feb. 2013. [Online]. Available: https://alanepaull.wordpress.com/2013/02/25/consuming-xcri-cap-ii-xcri-exchange-platform-xxp/

[258] T. Tiropanis, H. Davis, D. Millard, M. Weal, S. White, and G. Wills, "Semantic Technologies in Learning and Teaching (SemTech) - JISC Report," 2009. [Online]. Available: https://eprints.soton.ac.uk/267534/

[259] P. Barker, "A short project on linking course data," Aug. 2015. [Online]. Available: https://blogs.pjjk.net/phil/a-short-project-on-linking-course-data/

[260] S. Oussena and K. Hyeonsook, "Linked university course information [LUCI]." Birmingham: Jisc, 2013. [Online]. Available: https://issuu.com/jiscinfonet/docs/west-london-i1

[261] P. Ayres, "Using subjective measures to detect variations of intrinsic cognitive load within problems," *Learning and Instruction*, vol. 16, no. 5, pp. 389–400,

Oct. 2006. [Online]. Available: https://doi.org/10.1016/j.learninstruc.2006.09.001

[262] J. Sweller, “Cognitive load theory, learning difficulty, and instructional design,” *Learning and Instruction*, vol. 4, no. 4, pp. 295–312, Jan. 1994. [Online]. Available: https://doi.org/10.1016/0959-4752(94)90003-5

[263] S. Oviatt, “Human-centered Design Meets Cognitive Load Theory: Designing Interfaces That Help People Think,” in *Proceedings of the 14th ACM International Conference on Multimedia*, ser. MM ’06. New York, NY, USA: ACM, 2006, pp. 871–880, event-place: Santa Barbara, CA, USA. [Online]. Available: http://doi.acm.org/10.1145/1180639.1180831

[264] J. Sweller, “CHAPTER TWO - Cognitive Load Theory,” in *Psychology of Learning and Motivation*, J. P. Mestre and B. H. Ross, Eds. Academic Press, Jan. 2011, vol. 55, pp. 37–76. [Online]. Available: https://doi.org/10.1016/B978-0-12-387691-1.00002-8

[265] J.-F. Rouet, “What was I looking for? The influence of task specificity and prior knowledge on students’ search strategies in hypertext,” *Interacting with Computers*, vol. 15, no. 3, pp. 409–428, Jun. 2003. [Online]. Available: https://doi.org/10.1016/S0953-5438(02)00064-4

[266] M. Stokmans and J. Kamphuis, “End-users Searching the Online Catalogue: The Influence of Domain and System Knowledge on Search Patterns,” *Electronic Library*, vol. 12, no. 6, pp. 335–343, Jun. 1994. [Online]. Available: https://doi.org/10.1108/eb045321

[267] R. W. White, S. T. Dumais, and J. Teevan, “Characterizing the Influence of Domain Expertise on Web Search Behavior,” in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, ser. WSDM ’09. New York, NY, USA: ACM, 2009, pp. 132–141, event-place: Barcelona, Spain. [Online]. Available: http://doi.acm.org/10.1145/1498759.1498819

[268] T. J. F. Mitchell, S. Y. Chen, and R. D. Macredie, “Hypermedia learning and prior knowledge: domain expertise vs. system expertise,” *Journal of Computer Assisted Learning*, vol. 21, no. 1, pp. 53–64, 2005. [Online]. Available: https://doi.org/10.1111/j.1365-2729.2005.00113.x

[269] A. Sutcliffe, “Symbiosis and synergy? scenarios, task analysis and reuse of HCI knowledge,” *Interacting with Computers*, vol. 15, no. 2, pp. 245–263, Apr. 2003.

[270] D. Kelly, “Methods for Evaluating Interactive Information Retrieval Systems with Users,” *Foundations and Trends in Information Retrieval*, vol. 3, no. 1—2, pp. 1–224, Jan. 2009. [Online]. Available: https://doi.org/10.1561/1500000012

[271] M. W. van Someren, Y. F. Barnard, and J. a. C. Sandberg, *The think aloud method: a practical approach to modelling cognitive processes.* LondenAcademic Press, 1994. [Online]. Available: https://hdl.handle.net/11245/1.103289

[272] Z. Guan, S. Lee, E. Cuddihy, and J. Ramey, “The Validity of the Stimulated Retrospective Think-aloud Method As Measured by Eye Tracking,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’06. New York, NY, USA: ACM, 2006, pp. 1253–1262, event-place: Montréal, Québec, Canada. [Online]. Available: https://doi.acm.org/10.1145/1124772.1124961

[273] J. Nielsen and R. L. Mack, Eds., *Usability Inspection Methods.* Wiley, May 1994, google-Books-ID: cuRQAAAAMAAJ.

[274] C. A. Murphy, D. Coover, and S. V. Owen, “Development and Validation of the Computer Self-Efficacy Scale,” *Educational and Psychological Measurement*, vol. 49, no. 4, pp. 893–899, Dec. 1989. [Online]. Available: https://doi.org/10.1177/001316448904900412

[275] G. Torkzadeh, J. C.-J. Chang, and D. Demirhan, “A contingency model of computer and Internet self-efficacy,” *Information & Management*, vol. 43, no. 4, pp. 541–550, Jun. 2006. [Online]. Available: https://doi.org/10.1016/j.im.2006.02.001

[276] J. Brooke, “SUS: A Retrospective,” *J. Usability Studies*, vol. 8, no. 2, pp. 29–40, Feb. 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2817912.2817913

[277] ——, “SUS: A ’Quick and Dirty’ Usability Scale,” Jun. 1996. [Online]. Available: https://doi.org/10.1201/9781498710411-35

[278] A. Bangor, P. T. Kortum, and J. T. Miller, “An Empirical Evaluation of the System Usability Scale,” *International Journal of Human–Computer*

Interaction, vol. 24, no. 6, pp. 574–594, Jul. 2008. [Online]. Available: https://doi.org/10.1080/10447310802205776

[279] K. Finstad, "The Usability Metric for User Experience," *Interacting with Computers*, vol. 22, no. 5, pp. 323–327, Sep. 2010. [Online]. Available: https://doi.org/10.1016/j.intcom.2010.04.004

[280] N. Griffon, G. Kerdelhué, S. Hamek, S. Hassler, C. Boog, J.-B. Lamy, C. Duclos, A. Venot, and S. J. Darmoni, "Design and usability study of an iconic user interface to ease information retrieval of medical guidelines," *Journal of the American Medical Informatics Association*, vol. 21, no. e2, pp. e270–e277, Oct. 2014. [Online]. Available: https://doi.org/10.1136/amiajnl-2012-001548

[281] E. Kaufmann and A. Bernstein, "Evaluating the usability of natural language query languages and interfaces to Semantic Web knowledge bases," *Journal of Web Semantics*, vol. 8, no. 4, pp. 377–393, Nov. 2010. [Online]. Available: https://doi.org/10.1016/j.websem.2010.06.001

[282] H. L. O'Brien and L. McCay-Peet, "Asking "Good" Questions: Questionnaire Design and Analysis in Interactive Information Retrieval Research," in *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, ser. CHIIR '17. New York, NY, USA: ACM, 2017, pp. 27–36, event-place: Oslo, Norway. [Online]. Available: http://doi.acm.org/10.1145/3020165.3020167

[283] R. Ravendran, I. MacColl, and M. Docherty, "Usability Evaluation of a Tag-based Interface," *J. Usability Studies*, vol. 7, no. 4, pp. 143–160, Aug. 2012. [Online]. Available: http://dl.acm.org/citation.cfm?id=2835484.2835486

[284] J. R. Lewis and J. Sauro, "The Factor Structure of the System Usability Scale," in *Human Centered Design*, ser. Lecture Notes in Computer Science, M. Kurosu, Ed. Berlin, Heidelberg: Springer, 2009, pp. 94–103. [Online]. Available: https://doi.org/10.1007/978-3-642-02806-9_12

[285] J. R. Lewis, B. S. Utesch, and D. E. Maher, "Measuring Perceived Usability: The SUS, UMUX-LITE, and AltUsability," *International Journal of Human–Computer Interaction*, vol. 31, no. 8, pp. 496–505, Aug. 2015. [Online]. Available: https://doi.org/10.1080/10447318.2015.1064654

[286] D. McKay and S. Burriss, “Improving the Usability of Novel Web Software: An Industrial Case Study of an Institutional Repository,” in *Web Information Systems Engineering – WISE 2008 Workshops*, ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, Sep. 2008, pp. 102–111. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-540-85200-1_12

[287] S. C. Peres, T. Pham, and R. Phillips, “Validation of the System Usability Scale (SUS): SUS in the Wild,” *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 57, no. 1, pp. 192–196, Sep. 2013. [Online]. Available: https://doi.org/10.1177/1541931213571043

[288] J. Sauro and E. Kindlund, “A Method to Standardize Usability Metrics into a Single Score,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’05. New York, NY, USA: ACM, 2005, pp. 401–409, event-place: Portland, Oregon, USA. [Online]. Available: http://doi.acm.org/10.1145/1054972.1055028

[289] A. I. Martins, A. F. Rosa, A. Queirós, A. Silva, and N. P. Rocha, “European Portuguese Validation of the System Usability Scale (SUS),” *Procedia Computer Science*, vol. 67, pp. 293–300, Jan. 2015. [Online]. Available: https://doi.org/10.1016/j.procs.2015.09.273

[290] Z. Sharfina and H. B. Santoso, “An Indonesian adaptation of the System Usability Scale (SUS),” in *2016 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Oct. 2016, pp. 145–148. [Online]. Available: https://doi.org/10.1109/ICACSIS.2016.7872776

[291] S. Elkin-Frankston, B. K. Bracken, S. Irvin, and M. Jenkins, “Are Behavioral Measures Useful for Detecting Cognitive Workload During Human-Computer Interaction?” in *Advances in The Human Side of Service Engineering*, ser. Advances in Intelligent Systems and Computing, T. Z. Ahram and W. Karwowski, Eds. Cham: Springer International Publishing, 2017, pp. 127–137. [Online]. Available: https://doi.org/10.1007/978-3-319-41947-3_13

[292] J. Gwizdka, “Distribution of cognitive load in Web search,” *Journal of the American Society for Information Science and Technology*, vol. 61, no. 11, pp. 2167–2187, 2010. [Online]. Available: https://doi.org/10.1002/asi.21385

[293] A. W. Kushniruk and V. L. Patel, “Cognitive and usability engineering methods for the evaluation of clinical information systems,” *Journal of Biomedical Informatics*, vol. 37, no. 1, pp. 56–76, Feb. 2004. [Online]. Available: https://doi.org/10.1016/j.jbi.2004.01.003

[294] A. F. Rose, J. L. Schnipper, E. R. Park, E. G. Poon, Q. Li, and B. Middleton, “Using qualitative studies to improve the usability of an EMR,” *Journal of Biomedical Informatics*, vol. 38, no. 1, pp. 51–60, Feb. 2005. [Online]. Available: https://doi.org/10.1016/j.jbi.2004.11.006

[295] R. Agarwal and E. Karahanna, “Time Flies When You’re Having Fun: Cognitive Absorption and Beliefs about Information Technology Usage,” *MIS Quarterly*, vol. 24, no. 4, pp. 665–694, 2000. [Online]. Available: https://doi.org/10.2307/3250951

[296] J. Cegarra and A. Chevalier, “The use of Tholos software for combining measures of mental workload: Toward theoretical and methodological improvements,” *Behavior Research Methods*, vol. 40, no. 4, pp. 988–1000, Nov. 2008. [Online]. Available: https://doi.org/10.3758/BRM.40.4.988

[297] N. H. P. R. Group, *NASA Task Load Index (TLX), v.1.0.* California: NASA Ames Research Center, 1980. [Online]. Available: https://humansystems.arc.nasa.gov/groups/TLX/

[298] B. Yin and F. Chen, “Towards Automatic Cognitive Load Measurement from Speech Analysis,” in *Human-Computer Interaction. Interaction Design and Usability*, ser. Lecture Notes in Computer Science, J. A. Jacko, Ed. Berlin, Heidelberg: Springer, 2007, pp. 1011–1020. [Online]. Available: https://doi.org/10.1007/978-3-540-73105-4_111

[299] Q. Wang, S. Yang, M. Liu, Z. Cao, and Q. Ma, “An eye-tracking study of website complexity from cognitive load perspective,” *Decision Support Systems*, vol. 62, pp. 1–10, Jun. 2014. [Online]. Available: https://doi.org/10.1016/j.dss.2014.02.007

[300] S. G. Hart and L. E. Staveland, “Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research,” in *Advances in Psychology*, ser. Human Mental Workload, P. A. Hancock and N. Meshkati, Eds. North-Holland, Jan. 1988, vol. 52, pp. 139–183. [Online]. Available: https://doi.org/10.1016/S0166-4115(08)62386-9

[301] S. Carter and J. Mankoff, “When Participants Do the Capturing: The Role of Media in Diary Studies,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI ’05. New York, NY, USA: ACM, 2005, pp. 899–908, event-place: Portland, Oregon, USA. [Online]. Available: http://doi.acm.org/10.1145/1054972.1055098

[302] K. Dalkir, *Knowledge Management in Theory and Practice.* Routledge, Sep. 2013, google-Books-ID: CU20AAAAQBAJ.

[303] T. H. Davenport and L. Prusak, *Working Knowledge: How Organizations Manage what They Know.* Harvard Business Press, 1998.

[304] A. Tiwana, *The Knowledge Management Toolkit: Practical Techniques for Building a Knowledge Management System.* Upper Saddle River, NJ, USA: Prentice Hall PTR, 2000.

[305] S. Devi Ramachandran, S. Chong, and K. Wong, “Knowledge management practices and enablers in public universities: a gap analysis,” *Campus-Wide Information Systems*, vol. 30, no. 2, pp. 76–94, Mar. 2013. [Online]. Available: https://doi.org/10.1108/10650741311306273

[306] M. Mitri, “A Knowledge Management Framework for Curriculum Assessment,” *Journal of Computer Information Systems*, vol. 43, no. 4, pp. 15–24, Sep. 2003. [Online]. Available: https://doi.org/10.1080/08874417.2003.11647529

[307] V. Ratcliffe-Martin, E. Coakes, and G. Sugden, “Knowledge Management issues in universities,” *VINE*, vol. 30, no. 4, pp. 14–18, Apr. 2000. [Online]. Available: https://doi.org/10.1108/eb040770

[308] P. Barker, “Translating course descriptions from XCRI-CAP to schema.org,” Aug. 2017. [Online]. Available: https://blogs.pjjk.net/phil/course-description-xcri-cap-schema-org/

[309] H. Beetham and R. Sharpe, Eds., *Rethinking Pedagogy for a Digital Age: Designing for 21st Century Learning.* London: Routledge, Apr. 2013.

[310] S. Brown, “Large-scale innovation and change in UK higher education,” *Research in Learning Technology*, vol. 21, Sep. 2013. [Online]. Available: https://journal.alt.ac.uk/index.php/rlt/article/view/1473

[311] T. C. Ling, Y. Y. Jusoh, R. Abdullah, and N. H. Alwi, “User evaluation on curriculum design information system,” in *International Conference on Information Technology.* IEEE, Nov. 2014, pp. 195–199. [Online]. Available: https://doi.org/10.1109/icimu.2014.7066629

[312] R. Martinez-Maldonado, P. Goodyear, L. Carvalho, K. Thompson, D. Hernandez-Leo, Y. A. Dimitriadis, L. P. Prieto, and D. Wardak, “Supporting collaborative design activity in a multi-user digital design ecology,” *Computers in Human Behavior*, vol. 71, pp. 327–342, Jun. 2017. [Online]. Available: https://doi.org/10.1016/j.chb.2017.01.055

[313] W. Wang and Z. Wei, “Design and implementation of a knowledge management prototype system of NII,” in *International Conference on Consumer Electronics.* IEEE, Apr. 2012, pp. 3167–3170. [Online]. Available: https://doi.org/10.1109/cecnet.2012.6201450

[314] M. M. Kashyap, “Curriculum development and design process: A systems approach,” *International Library Review*, vol. 11, no. 3, pp. 353–365, Jul. 1979. [Online]. Available: https://doi.org/10.1016/0020-7837(79)90005-0

[315] D. Pratt, “System Theory, Systems Technology, and Curriculum Design,” *The Journal of Educational Thought (JET) / Revue de la Pensée Éducative*, vol. 12, no. 2, pp. 131–152, 1978. [Online]. Available: https://www.jstor.org/stable/23767934

[316] Y. Refecadu, “Systems approach to curriculum development and instruction for occupational education in ethiopia,” Ph.D. dissertation, Oklahoma State University, Stillwater, Oklahoma, May 1975. [Online]. Available: https://hdl.handle.net/11244/20401

[317] T. Dowding, “Managing Chaos (Or How to Survive the Instructional Development Process),” *Educational Technology*, vol. 31, no. 1, pp. 26–31, 1991. [Online]. Available: https://www.jstor.org/stable/44427496

[318] S. McKenney, N. Nieveen, and A. Strijker, “Information Technology Tools for Curriculum Development,” in *International Handbook of Information Technology in Primary and Secondary Education*, ser. Springer International Handbook of Information Technology in Primary and Secondary Education, J. Voogt and G. Knezek, Eds. Boston, MA: Springer US, 2008, pp. 195–210. [Online]. Available: https://doi.org/10.1007/978-0-387-73315-9_12

[319] H. J. Vos, "Applications of general systems theory to the development of an adjustable tutorial software machine," *Computers & Education*, vol. 22, no. 3, pp. 265–276, Apr. 1994. [Online]. Available: https://doi.org/10.1016/0360-1315(94)90008-6

[320] G. Macgregor, "A Quick Guide for AQ Staff : Using the C-CAP Administration Dashboard - C-CAP Embedding Phase," University of Strathclyde, Glasgow, Report, Oct. 2012. [Online]. Available: https://strathprints.strath.ac.uk/59852/

[321] ——, "Improving the discoverability and web impact of open repositories: techniques and evaluation," *The Code4Lib Journal*, no. 43, Feb. 2019. [Online]. Available: https://journal.code4lib.org/articles/14180

[322] ——, "Enhancing content discovery of open repositories: an analytics-based evaluation of repository optimizations," *Publications*, vol. 8, no. 1, p. 8, 2020. [Online]. Available: https://doi.org/10.3390/publications8010008

[323] C. Armbruster and L. Romary, "Comparing Repository Types: Challenges and Barriers for Subject-Based Repositories, Research Repositories, National Repository Systems and Institutional Repositories in Serving Scholarly Communication," *International Journal of Digital Library Systems (IJDLS)*, vol. 1, no. 4, pp. 61–73, Oct. 2010. [Online]. Available: https://doi.org/10.4018/jdls.2010100104

[324] C. De Laurentis, "Digital Knowledge Exploitation: ICT, Memory Institutions and Innovation from Cultural Assets," *The Journal of Technology Transfer*, vol. 31, no. 1, pp. 77–89, Jan. 2006. [Online]. Available: https://doi.org/10.1007/s10961-005-5014-6

[325] A. Fresa, B. Justrell, and C. Prandoni, "Digital curation and quality standards for memory institutions: PREFORMA research project," *Archival Science*, vol. 15, no. 2, pp. 191–216, Jun. 2015. [Online]. Available: https://doi.org/10.1007/s10502-015-9242-8

[326] H. Robinson, "Remembering things differently: museums, libraries and archives as memory institutions and the implications for convergence," *Museum Management and Curatorship*, vol. 27, no. 4, pp. 413–429, Oct. 2012. [Online]. Available: https://doi.org/10.1080/09647775.2012.720188

[327] C. S. Burns, A. Lana, and J. M. Budd, "Institutional Repositories: Exploration of Costs and Value," *D-Lib Magazine*, vol. 19, no. 1/2, Jan. 2013. [Online]. Available: https://doi.org/10.1045/january2013-burns

[328] S. Gibbons, "Defining an Institutional Repository," *Library Technology Reports*, vol. 40, no. 4, pp. 6–10, Jun. 2009. [Online]. Available: https://journals.ala.org/index.php/ltr/article/view/4378

[329] R. E. Jones, T. Andrew, and J. MacColl, *The Institutional Repository.* Oxford: Chandos Publishing, Jan. 2006, google-Books-ID: swOkAgAAQBAJ.

[330] R. Kennison, S. Shreeves, and S. Harnad, "Point & Counterpoint The Purpose of Institutional Repositories: Green OA or Beyond?" *Journal of Librarianship and Scholarly Communication*, vol. 1, no. 4, p. eP1105, Sep. 2013. [Online]. Available: https://doi.org/10.7710/2162-3309.1105

[331] N. Semple, "Digital Repositories [DCC Briefing Papers : Introduction to Curation]," Digital Curation Centre (DCC), Edinburgh, Tech. Rep. 1842/3372, 2006. [Online]. Available: http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation

[332] G. Macgregor, "ePrints and PURE : Discussion Paper," Dec. 2013. [Online]. Available: https://strathprints.strath.ac.uk/60311/

[333] T. Reimer, "The once and future library: the role of the (national) library in supporting research," *Insights*, vol. 31, no. 0, p. 19, May 2018. [Online]. Available: https://doi.org/10.1629/uksg.409

[334] Y. Li and M. Banach, "Institutional Repositories and Digital Preservation: Assessing Current Practices at Research Libraries," *D-Lib Magazine*, vol. 17, no. 5/6, May 2011. [Online]. Available: https://doi.org/10.1045/may2011-yuanli

[335] M. Klein, H. Van de Sompel, R. Sanderson, H. Shankar, L. Balakireva, K. Zhou, and R. Tobin, "Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot," *PLOS ONE*, vol. 9, no. 12, p. e115253, Dec. 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0115253

[336] H. Van de Sompel, M. Klein, and H. Shankar, "Towards Robust Hyperlinks for Web-Based Scholarly Communication," in *Intelligent Computer Mathematics*, ser. Lecture Notes in Computer Science, S. M. Watt,

J. H. Davenport, A. P. Sexton, P. Sojka, and J. Urban, Eds. Cham: Springer International Publishing, 2014, pp. 12–25. [Online]. Available: https://doi.org/10.1007/978-3-319-08434-3_2

[337] H. Van de Sompel, R. Sanderson, H. Shankar, and M. Klein, "Persistent Identifiers for Scholarly Assets and the Web: The Need for an Unambiguous Mapping," *International Journal of Digital Curation*, vol. 9, pp. 331–342, Jul. 2014. [Online]. Available: https://doi.org/10.2218/ijdc.v9i1.320

[338] J. Adamick and R. Reznik-Zellen, "Trends in Large-Scale Subject Repositories," *D-Lib Magazine*, vol. 16, no. 11/12, Nov. 2010. [Online]. Available: https://doi.org/10.1045/november2010-adamick

[339] M. Armstrong, "Institutional repository management models that support faculty research dissemination," *OCLC Systems & Services*, vol. 30, no. 1, pp. 43–51, Jan. 2014. [Online]. Available: https://doi.org/10.1108/OCLC-07-2013-0028

[340] T. Ferreras-Fernández, J. A. Merlo-Vega, and F. J. García-Peñalvo, "Impact of Scientific Content in Open Access Institutional Repositories: A Case Study of the Repository Gredos," in *Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality*, ser. TEEM '13. New York, NY, USA: ACM, 2013, pp. 357–363, event-place: Salamanca, Spain. [Online]. Available: http://doi.acm.org/10.1145/2536536.2536590

[341] T. Lee-Hwa, A. Abrizah, and A. Noorhidawati, "Availability and visibility of open access digital repositories in ASEAN countries," *Information Development*, vol. 29, no. 3, pp. 274–285, Aug. 2013. [Online]. Available: https://doi.org/10.1177/0266666912466754

[342] I. F. Aguillo, J. L. Ortega, M. Fernández, and A. M. Utrilla, "Indicators for a webometric ranking of open access repositories," *Scientometrics*, vol. 82, no. 3, pp. 477–486, Mar. 2010. [Online]. Available: https://doi.org/10.1007/s11192-010-0183-y

[343] B.-C. Björk, M. Laakso, P. Welling, and P. Paetau, "Anatomy of green open access," *Journal of the Association for Information Science and Technology*, vol. 65, no. 2, pp. 237–250, 2014. [Online]. Available: https://doi.org/10.1002/asi.22963

[344] H. Piwowar, J. Priem, V. Larivière, J. P. Alperin, L. Matthias, B. Norlander, A. Farley, J. West, and S. Haustein, "The state of OA: a large-scale analysis of the prevalence and impact of Open Access articles," *PeerJ*, vol. 6, p. e4375, Feb. 2018. [Online]. Available: https://doi.org/10.7717/peerj.4375

[345] Y. Ioannidis, "From digital libraries to knowledge commons," *European Research Consortium for Informatics and Mathematics - ERCIM News*, no. 66, pp. 14–15, 2006. [Online]. Available: https://www.ercim.eu/publication/Ercim_News/enw66/EN66.pdf

[346] P. Suber, "Creating an intellectual commons through open access," in *Understanding Knowledge as a Commons: From Theory to Practice*, C. Hess and E. Ostrom, Eds. Cambridge, MA: MIT Press, 2007, pp. 171–208. [Online]. Available: http://nrs.harvard.edu/urn-3:HUL.InstRepos:4552055

[347] A. Wittel, "Counter-commodification: The economy of contribution in the digital commons," *Culture and Organization*, vol. 19, no. 4, pp. 314–331, Sep. 2013. [Online]. Available: https://doi.org/10.1080/14759551.2013.827422

[348] R. M. Stallman, "The GNU GPL and the American Way," Free Software Foundation, Boston, MA., Tech. Rep., 2001. [Online]. Available: https://www.gnu.org/philosophy/gpl-american-way.en.html

[349] D. Bollier, "The Growth of the Commons Paradigm," in *Understanding Knowledge as a Commons: From Theory to Practice*, C. Hess and E. Ostrom, Eds. Cambridge, MA: MIT Press, 2007, pp. 27–40. [Online]. Available: http://hdl.handle.net/10535/4975

[350] B. Fecher and S. Friesike, "Open Science: One Term, Five Schools of Thought," in *Opening Science: The Evolving Guide on How the Internet is Changing Research, Collaboration and Scholarly Publishing*, S. Bartling and S. Friesike, Eds. Cham: Springer International Publishing, 2014, pp. 17–47. [Online]. Available: https://doi.org/10.1007/978-3-319-00026-8_2

[351] J. C. Molloy, "The Open Knowledge Foundation: Open Data Means Better Science," *PLOS Biology*, vol. 9, no. 12, p. e1001195, Dec. 2011. [Online]. Available: https://doi.org/10.1371/journal.pbio.1001195

[352] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. L. Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, and T. Yarkoni, "Promoting an open research culture," *Science*, vol. 348, no. 6242, pp. 1422–1425, Jun. 2015. [Online]. Available: https://doi.org/10.1126/science.aab2374

[353] J. Willinsky, "The unacknowledged convergence of open source, open access, and open science," *First Monday*, vol. 10, no. 8, Aug. 2005. [Online]. Available: https://doi.org/10.5210/fm.v10i8.1265

[354] M. Woelfle, P. Olliaro, and M. H. Todd, "Open science is a research accelerator," *Nature Chemistry*, vol. 3, no. 10, pp. 745–748, Oct. 2011. [Online]. Available: https://doi.org/10.1038/nchem.1149

[355] P. Knoth and Z. Zdrahal, "Mining cross-document relationships from text," Barcelona, Spain, 2011. [Online]. Available: http://oro.open.ac.uk/29302/

[356] ——, "CORE: connecting repositories in the open access domain." Geneva, Switzerland: CERN, 2011. [Online]. Available: http://oro.open.ac.uk/32560/

[357] ——, "CORE: Aggregation Use Cases for Open Access," in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, ser. JCDL '13. New York, NY, USA: ACM, 2013, pp. 441–442, event-place: Indianapolis, Indiana, USA. [Online]. Available: http://doi.acm.org/10.1145/2467696.2467787

[358] P. Labropoulou, D. Galanis, A. Lempesis, M. Greenwood, P. Knoth, R. Eckart de Castilho, S. Sachtouris, B. Georgantopoulos, L. Anastasiou, S. Martziou, G. Katerina, N. Manola, and S. Piperidis, "OpenMinTeD: A Platform Facilitating Text Mining of Scholarly Content," in *WOSP 2018 Workshop Proceedings*. Luxemburg: European Language Resources Association (ELRA), Jun. 2018. [Online]. Available: http://oro.open.ac.uk/55790/

[359] D. R. Swanson, "Medical literature as a potential source of new knowledge." *Bulletin of the Medical Library Association*, vol. 78, no. 1, pp. 29–37, Jan. 1990. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC225324/

[360] D. R. Swanson and N. R. Smalheisert, “Link Analysis of MedLine Titles as an Aid to Scientific Discovery,” in *In Proceedings of the AAAI Fall Symposium on Artificial Intelligence and Link Analysis.* Menlo Park, CA.: AAAI Press, 1998, pp. 94–97. [Online]. Available: https://www.aaai.org/Papers/Symposia/Fall/1998/FS-98-01/FS98-01-017.pdf

[361] C. Lagoze, H. Van de Sompel, M. Nelson, and S. Warner, “Open Archives Initiative - Protocol for Metadata Harvesting - Implementation Guidelines,” May 2005. [Online]. Available: https://www.openarchives.org/OAI/2.0/guidelines.htm

[362] H. Van de Sompel, M. L. Nelson, C. Lagoze, and S. Warner, “Resource Harvesting within the OAI-PMH Framework,” *D-Lib Magazine*, vol. 10, no. 12, Dec. 2004. [Online]. Available: https://doi.org/10.1045/december2004-vandesompel

[363] P. Knoth, V. Robotka, and Z. Zdrahal, “Connecting Repositories in the Open Access Domain Using Text Mining and Semantic Data,” in *Research and Advanced Technology for Digital Libraries*, ser. Lecture Notes in Computer Science, S. Gradmann, F. Borri, C. Meghini, and H. Schuldt, Eds. Berlin, Heidelberg: Springer, 2011, pp. 483–487.

[364] P. Knoth and D. Herrmannova, “Towards Semantometrics: A New Semantic Similarity Based Measure for Assessing a Research Publication’s Contribution,” *D-Lib Magazine*, vol. 20, no. 11/12, 2014. [Online]. Available: https://doi.org/10.1045/november14-knoth

[365] N. Pontika, L. Anastasiou, A. Charalampous, M. Cancellieri, S. Pearce, and P. Knoth, “CORE Recommender: a plug in suggesting open access content.” Edinburgh: University of Edinburgh, Aug. 2017. [Online]. Available: http://hdl.handle.net/1842/23359

[366] P. Knoth and M. Cancellieri, “Analysing the performance of open access papers discovery tools,” Hamburg, 2019. [Online]. Available: https://www.slideshare.net/petrknoth/analysing-the-performance-of-open-access-papers-discovery-tools

[367] D. S. Chawla, “Unpaywall finds free versions of paywalled papers,” *Nature*, Apr. 2017. [Online]. Available: https://doi.org/10.1038/nature.2017.21765

[368] B. Haslhofer, S. Warner, C. Lagoze, M. Klein, R. Sanderson, M. L. Nelson, and H. Van de Sompel, "ResourceSync: Leveraging Sitemaps for Resource Synchronization," in *Proceedings of the 22Nd International Conference on World Wide Web*, ser. WWW '13 Companion. New York, NY, USA: ACM, 2013, pp. 11–14, event-place: Rio de Janeiro, Brazil. [Online]. Available: http://doi.acm.org/10.1145/2487788.2487793

[369] H. Van de Sompel, M. L. Nelson, M. Klein, and R. Sanderson, "ResourceSync: The NISO/OAI Resource Synchronization Framework," in *Research and Advanced Technology for Digital Libraries*, ser. Lecture Notes in Computer Science, T. Aalberg, C. Papatheodorou, M. Dobreva, G. Tsakonas, and C. J. Farrugia, Eds. Berlin, Heidelberg: Springer, 2013, pp. 488–489.

[370] K. Shearer, E. Rodrigues, A. Bollini, A. Cabezas, D. Castelli, L. Carr, L. Chan, C. Humphrey, R. Johnson, P. Knoth, P. Manghi, L. Matizirofa, P. Perakakis, J. Schirrwagen, T. Smith, H. Van de Sompel, P. Walk, D. Wilcox, and K. Yamaji, "Next generation repositories: scaling up repositories to a global knowledge commons," in *Open Repositories 2018 (OR2018).* Minho: University of Minho, 2018. [Online]. Available: http://hdl.handle.net/1822/55027

[371] RIOXX, "RIOXX: Governance Group," 2019. [Online]. Available: http://www.rioxx.net/governance/

[372] J. G. Bankier and K. Gleason, "Institutional repository software comparison - UNESCO Digital Library," UNESCO, Paris, Tech. Rep., 2014. [Online]. Available: https://unesdoc.unesco.org/ark:/48223/pf0000227115

[373] A. Acharya, *Indexing repositories: pitfalls and best practices*, Jun. 2015. [Online]. Available: https://media.dlib.indiana.edu/media_objects/9z903008w

[374] A. D. R. Sergiadis, "Evaluating Zotero, SHERPA/RoMEO, and Unpaywall in an institutional repository workflow," *Journal of Electronic Resources Librarianship*, vol. 31, no. 3, pp. 152–176, Jul. 2019. [Online]. Available: https://doi.org/10.1080/1941126X.2019.1635396

[375] V. P. Yadav, "Reference Management Tools: EndNote, Mendeley, RefWorks and Zotero," *Journal of Advancements in Library Sciences*, vol. 6, no. 1, pp. 315–319, Mar. 2019. [Online]. Available: http://sciencejournals.stmjournals.in/index.php/JoALS/article/view/1781

[376] Google, "Learn about sitemaps [Google Search Console]," 2019. [Online]. Available: https://support.google.com/webmasters/answer/156184?hl=en

[377] S. Lewis, P. de Castro, and R. Jones, "SWORD: Facilitating Deposit Scenarios," *D-Lib Magazine*, vol. 18, no. 1/2, Jan. 2012. [Online]. Available: https://doi.org/10.1045/january2012-lewis

[378] G. Mosweunyane and L. A. Carr, "Direct desktop-repository deposits with SWORD," in *2014 IST-Africa Conference Proceedings*. Piscataway, NJ.: IEEE, May 2014, pp. 1–8, iSSN: null. [Online]. Available: https://doi.org/10.1109/ISTAFRICA.2014.6880604

[379] B. d. hOra and J. Gregorio, "The Atom Publishing Protocol [RFC 5023]," 2007. [Online]. Available: https://tools.ietf.org/html/rfc5023

[380] M. Artini, C. Atzori, A. Bardi, S. La Bruzzo, P. Manghi, and A. Mannocci, "The OpenAIRE Literature Broker Service for Institutional Repositories," *D-Lib Magazine*, vol. 21, no. 11/12, Nov. 2015. [Online]. Available: https://doi.org/10.1045/november2015-artini

[381] A. Martinez Garcia, "Enhancing Open Access at Cambridge: Apollo repository and CRIS integrations," Hamburg, Jun. 2019. [Online]. Available: https://doi.org/10.17863/CAM.40634

[382] K. Arlitsch and P. OBrien, "Invisible institutional repositories: addressing the low indexing ratios of IRs in Google Scholar," *Library Hi Tech*, vol. 30, no. 1, pp. 60–81, Mar. 2012. [Online]. Available: https://doi.org/10.1108/07378831211213210

[383] D. Askey and K. Arlitsch, "Heeding the Signals: Applying Web best Practices when Google recommends," *Journal of Library Administration*, vol. 55, no. 1, pp. 49–59, Jan. 2015. [Online]. Available: https://doi.org/10.1080/01930826.2014.978685

[384] P. Knoth, "From open access metadata to open access content: two principles for increased visibility of open access content," Charlottetown, Prince Edward Island, Canada, 2013. [Online]. Available: https://oro.open.ac.uk/45935/

[385] N. Pontika, P. Knoth, M. Cancellieri, and S. Pearce, "Developing Infrastructure to Support Closer Collaboration of Aggregators with Open Repositories,"

LIBER Quarterly, vol. 25, no. 4, pp. 172–188, Mar. 2016. [Online]. Available: https://doi.org/10.18352/lq.10138

[386] M. Klein, M. Cancellieri, and P. Knoth, "Comparing the Performance of OAI-PMH with ResourceSync." Hamburg: Zenodo, Nov. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3554256

[387] A. Lopatenko, "Information retrieval in Current Research Information Systems," *arXiv:cs/0110026*, Oct. 2001, arXiv: cs/0110026. [Online]. Available: http://arxiv.org/abs/cs/0110026

[388] E. H. Zimmerman, "CRIS-Cross: Research Information Systems at a Crossroads." euroCRIS, Aug. 2002. [Online]. Available: http://hdl.handle.net/11366/129

[389] P. de Castro, K. Shearer, and F. Summann, "The Gradual Merging of Repository and CRIS Solutions to Meet Institutional Research Information Management Requirements," *Procedia Computer Science*, vol. 33, pp. 39–46, Jan. 2014. [Online]. Available: https://doi.org/10.1016/j.procs.2014.06.007

[390] K. Jeffery and A. Asserson, "Institutional Repositories and Current Research Information Systems," *New Review of Information Networking*, vol. 14, no. 2, pp. 71–83, Nov. 2009. [Online]. Available: https://doi.org/10.1080/13614570903359357

[391] S. Buranyi, "Is the staggeringly profitable business of scientific publishing bad for science?" *The Guardian*, Jun. 2017. [Online]. Available: https://www.theguardian.com/science/2017/jun/27/profitable-business-scientific-publishing-bad-for-science

[392] G. Macgregor, "Repository and CRIS interoperability issues within a 'connector lite' environment," in *14th International Conference on Open Repositories (OR2019).* Universität Hamburg: University of Strathclyde, Jun. 2019. [Online]. Available: https://strathprints.strath.ac.uk/68240/

[393] S. Moore, J. Gray, D. Lämmerhirt, and A. Swan, "PASTEUR4OA Briefing Paper: Infrastructures for Open Scholarly Communication," National Documentation Centre, Athens, Tech. Rep., 2016. [Online]. Available: http://pasteur4oa.eu/resources/229

[394] B. Brembs, “Prestigious Science Journals Struggle to Reach Even Average Reliability,” *Frontiers in Human Neuroscience*, vol. 12, 2018. [Online]. Available: https://doi.org/10.3389/fnhum.2018.00037

[395] A. Grossmann and B. Brembs, “Assessing the size of the affordability problem in scholarly publishing,” PeerJ Inc., Tech. Rep. e27809v1, Jun. 2019. [Online]. Available: https://doi.org/10.7287/peerj.preprints.27809v1

[396] J. P. Tennant and B. Brembs, “RELX referral to EU competition authority,” Oct. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1472045

[397] J. P. Tennant, F. Waldner, D. C. Jacques, P. Masuzzo, L. B. Collister, and C. H. J. Hartgerink, “The academic, economic and societal impacts of Open Access: an evidence-based review,” *F1000Research*, vol. 5, p. 632, Sep. 2016. [Online]. Available: https://doi.org/10.12688/f1000research.8460.3

[398] K. Arlitsch, “Driving Traffic to Institutional Repositories: How Search Engine Optimization can Increase the Number of Downloads from IR,” Sep. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.894564

[399] P. OBrien, K. Arlitsch, L. B. Sterman, J. Mixter, J. Wheeler, and S. Borda, “Undercounting File Downloads from Institutional Repositories,” *Journal of Library Administration*, vol. 56, no. 7, pp. 1–24, Oct. 2016. [Online]. Available: https://scholarworks.montana.edu/xmlui/handle/1/9943

[400] G. Macgregor, “Reviewing repository discoverability : approaches to improving repository visibility and web impact,” in *Repository Fringe 2017*, John McIntyre Conference Centre, University of Edinburgh, Aug. 2017. [Online]. Available: https://strathprints.strath.ac.uk/61333/

[401] D. K. Harman and E. M. Voorhees, “TREC: An overview,” *Annual Review of Information Science and Technology*, vol. 40, no. 1, pp. 113–155, 2006. [Online]. Available: https://doi.org/10.1002/aris.1440400111

[402] E. M. Voorhees, “TREC: Continuing information retrieval’s tradition of experimentation,” *Communications of the ACM*, vol. 50, no. 11, p. 51, Nov. 2007. [Online]. Available: https://doi.org/10.1145/1297797.1297822

[403] R. MacIntyre, P. Needham, J. Lambert, and J. Alcock, “Measuring the Usage of Repositories via a National Standards-based Aggregation Service: IRUS-UK,”

in *New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust: Proceedings of the 19th International Conference on Electronic Publishing.* Malta: IOS Press, 2015, pp. 83–92.

[404] Google, "Google Analytics," 2019. [Online]. Available: https://analytics.google.com/

[405] Google, *Google Search Console*, 2019. [Online]. Available: https://www.google.com/webmasters/tools/home

[406] R. J. Robertson, "STARGATE : Static Repository Gateway and Toolkit. Final Project Report," JISC, London, Report, 2006. [Online]. Available: https://strathprints.strath.ac.uk/14137/

[407] R. J. Robertson and A. Dawson, "An easy option? OAI static repositories as a method of exposing publishers' metadata to the information environment," in *Proceedings of the 10th International Conference on Electronic Publishing*, Jun. 2006, pp. 59–70. [Online]. Available: https://strathprints.strath.ac.uk/2373/

[408] M. Moffat, S. Chumbe, and R. MacLeod, "'Marketing' with Metadata - Increase Exposure and Visibility of Content. OAI-PMH, Z39.50, SRU/SRW, RSS," Heriot-Watt University, Edinburgh, Tech. Rep. 1.0, 2006. [Online]. Available: http://www.icbl.hw.ac.uk/perx/advocacy/exposingmetadata.htm

[409] E. McCulloch, "Taking stock of open access: progress and issues," *Library Review*, vol. 55, no. 6, pp. 337–343, Jan. 2006. [Online]. Available: https://doi.org/10.1108/00242530610674749

[410] J. Alhuay-Quispe, D. Quispe-Riveros, L. Bautista-Ynofuente, and J. Pacheco-Mendoza, "Metadata Quality and Academic Visibility Associated with Document Type Coverage in Institutional Repositories of Peruvian Universities," *Journal of Web Librarianship*, vol. 11, no. 3-4, pp. 241–254, Oct. 2017. [Online]. Available: https://doi.org/10.1080/19322909.2017.1382427

[411] A. Beisler and G. Willis, "Beyond Theory: Preparing Dublin Core Metadata for OAI-PMH Harvesting," *Journal of Library Metadata*, vol. 9, no. 1-2, pp. 65–97, Aug. 2009. [Online]. Available: https://doi.org/10.1080/19386380903095099

[412] E. Bellini and P. Nesi, “Metadata Quality Assessment Tool for Open Access Cultural Heritage Institutional Repositories,” in *Information Technologies for Performing Arts, Media Access, and Entertainment*, ser. Lecture Notes in Computer Science, P. Nesi and R. Santucci, Eds. Berlin, Heidelberg: Springer, 2013, pp. 90–103. [Online]. Available: https://doi.org/10.1007/978-3-642-40050-6_9

[413] F. McCown, M. Nelson, M. Zubair, and X. Liu, “Search engine coverage of the OAI-PMH corpus,” *IEEE Internet Computing*, vol. 10, no. 2, pp. 66–73, Mar. 2006. [Online]. Available: https://doi.org/10.1109/MIC.2006.41

[414] D. Allison, “OAI-PMH Harvested Collections and User Engagement,” *Journal of Web Librarianship*, vol. 10, no. 1, pp. 14–27, Jan. 2016. [Online]. Available: https://doi.org/10.1080/19322909.2015.1128867

[415] S. Lewis, L. Hayes, V. Newton-Wade, A. Corfield, R. Davis, S. Wilson, and T. Donohue, “If SWORD is the answer, what is the question? Use of the Simple Web service Offering Repository Deposit protocol,” *Program: Electronic Library and Information Systems*, vol. 43, no. 3, pp. 407–418, 2009. [Online]. Available: http://hdl.handle.net/2292/5315

[416] J. Downing, P. Murray-Rust, A. P. Tonge, P. Morgan, H. S. Rzepa, F. Cotterill, N. Day, and M. J. Harvey, “SPECTRa: The Deposition and Validation of Primary Chemistry Research Data in Digital Repositories,” *Journal of Chemical Information and Modeling*, vol. 48, no. 8, pp. 1571–1581, Aug. 2008. [Online]. Available: https://doi.org/10.1021/ci7004737

[417] E. F. Duranceau and R. Rodgers, “Automated IR deposit via the SWORD protocol: an MIT/BioMed Central experiment,” *Serials*, vol. 23, no. 3, pp. 212–214, Nov. 2010. [Online]. Available: https://doi.org/10.1629/23212

[418] CORE, “CORE access to raw data,” 2019. [Online]. Available: https://www.youtube.com/watch?time_continue=1&v=nf7IZIzBo5U&feature=emb_title

[419] G. Dunsire, “Collecting metadata from institutional repositories,” *OCLC Systems and Services*, vol. 24, pp. 51–58, 2008. [Online]. Available: https://doi.org/10.1108/10650750810847251

[420] E. Roel, “The MOSC Project: Using the OAI-PMH to Bridge Metadata Cultural Differences across Museums, Archives, and Libraries,” *Information Technology and Libraries*, vol. 24, no. 1, pp. 22–24, Mar. 2005. [Online]. Available: https://doi.org/10.6017/ital.v24i1.3360

[421] S. L. Shreeves, T. G. Habing, K. Hagedorn, and J. Young, “Current developments and future trends for the OAI Protocol for Metadata Harvesting,” *Library Trends*, vol. 53, no. 4, pp. 576–589, 2005. [Online]. Available: http://hdl.handle.net/2142/609

[422] J. M. Giménez-García, H. Thakkar, and A. Zimmermann, “Assessing Trust with PageRank in the Web of Data,” in *The Semantic Web*, ser. Lecture Notes in Computer Science, H. Sack, G. Rizzo, N. Steinmetz, D. Mladenić, S. Auer, and C. Lange, Eds. Cham: Springer International Publishing, 2016, pp. 293–307.

[423] Google, *About Data Highlighter*, 2018. [Online]. Available: https://perma.cc/92DY-MFZP

[424] J. W. Creswell and V. L. P. Clark, *Designing and Conducting Mixed Methods Research.* London: SAGE, 2007.

[425] A. Katsirikou and C. H. Skiadas, *Qualitative and Quantitative Methods in Libraries: Theory and Applications : Proceedings of the International Conference on QQML2009, Chania, Crete, Greece, 26-29 May 2009.* Singapore: World Scientific, 2010, google-Books-ID: agnJxWzMHhMC.

[426] G. Macgregor and J. Turner, “Revisiting e-learning effectiveness : proposing a conceptual model,” *Interactive Technology and Smart Education*, vol. 6, pp. 156–172, 2009. [Online]. Available: https://doi.org/10.1108/17415650911005375

[427] G. Macgregor, A. Spiers, and C. Taylor, “Exploratory evaluation of audio email technology in formative assessment feedback,” *Research in Learning Technology*, vol. 19, pp. 39–59, 2011. [Online]. Available: https://doi.org/10.1080/09687769.2010.547930

[428] P. Cairns and A. L. Cox, *Research Methods for Human-Computer Interaction.* Cambridge: Cambridge University Press, Aug. 2008, google-Books-ID: VtQFji-FOqhwC.

[429] W. James, *Pragmatism: a new name for some old ways of thinking: popular lectures on philosophy.* Longmans, green.

[430] J. Dewey, *Logic : the theory of inquiry.* Henry Holt.

[431] P. J. Scott and J. S. Briggs, "A pragmatist argument for mixed methodology in medical informatics," vol. 3, no. 3, pp. 223–241, publisher: SAGE Publications. [Online]. Available: https://doi.org/10.1177/1558689809334209

[432] C. Robson, *Real world research: a resource for users of social research methods in applied settings*, 3rd ed. Wiley, OCLC: 729956086.

[433] D. L. Morgan, "Pragmatism as a paradigm for social research," vol. 20, no. 8, pp. 1045–1053. [Online]. Available: https://doi.org/10.1177/1077800413513733

[434] V. Kaushik and C. A. Walsh, "Pragmatism as a research paradigm and its implications for social work research," vol. 8, no. 9, p. 255, number: 9 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://doi.org/10.3390/socsci8090255

[435] J. W. Creswell and J. D. Creswell, *Research design: qualitative, quantitative, and mixed methods approaches*, 5th ed. SAGE, OCLC: 1021400902.

[436] B. Hjørland, "Empiricism, rationalism and positivism in library and information science," vol. 61, no. 1, pp. 130–155. [Online]. Available: https://doi.org/10.1108/00220410510578050

[437] L. S. Giddings and B. M. Grant, "A trojan horse for positivism? a critique of mixed methods research," vol. 30, no. 1, pp. 52–60. [Online]. Available: https://journals.lww.com/advancesinnursingscience/Abstract/2007/01000/A_Trojan_Horse_for_Positivism___A_Critique_of.6.aspx

[438] J. N. Hall, "Pragmatism, evidence, and mixed methods evaluation," vol. 2013, no. 138, pp. 15–26, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ev.20054. [Online]. Available: https://doi.org/10.1002/ev.20054

[439] O. Sundin and J. Johannisson, "Pragmatism, neo-pragmatism and sociocultural theory: Communicative participation as a perspective in LIS," vol. 61, no. 1, pp. 23–43. [Online]. Available: https://doi.org/10.1108/00220410510577998

[440] P. Dalsgaard and C. Dindler, "Between theory and practice: bridging concepts in HCI research," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '14. Association for Computing Machinery, pp. 1635–1644. [Online]. Available: https://doi.org/10.1145/2556288.2557342

[441] J. Lazar, J. H. Feng, and H. Hochheiser, *Research Methods in Human-Computer Interaction.* Wiley Global Education, google-Books-ID: qi1EDwAAQBAJ.

[442] N. M. Su, V. Kaptelinin, J. Bardzell, S. Bardzell, J. R. Brubaker, A. Light, and D. Svanaes, "Standing on the shoulders of giants: Exploring the intersection of philosophy and HCI," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '19. Association for Computing Machinery, pp. 1–8. [Online]. Available: https://doi.org/10.1145/3290607.3299020

[443] J. Euzenat and P. Shvaiko, *Ontology Matching.* Berlin, Heidelberg: Springer-Verlag, 2007. [Online]. Available: https://doi.org/10.1007/978-3-642-38721-0

[444] N. Freire, R. Voorburg, R. Cornelissen, S. de Valk, E. Meijers, and A. Isaac, "Aggregation of Linked Data in the Cultural Heritage Domain: A Case Study in the Europeana Network," *Information*, vol. 10, no. 8, p. 252, Aug. 2019. [Online]. Available: https://doi.org/10.3390/info10080252

[445] R. S. Gonçalves, M. R. Kamdar, and M. A. Musen, "Aligning Biomedical Metadata with Ontologies Using Clustering and Embeddings," in *The Semantic Web*, ser. Lecture Notes in Computer Science, P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A. J. Gray, V. Lopez, A. Haller, and K. Hammar, Eds. Cham: Springer International Publishing, 2019, pp. 146–161. [Online]. Available: https://doi.org/10.1007/978-3-030-21348-0_10

[446] A. Isaac, S. Wang, C. Zinn, H. Matthezing, L. van der Meij, and S. Schlobach, "Evaluating Thesaurus Alignments for Semantic Interoperability in the Library Domain," *IEEE Intelligent Systems*, vol. 24, pp. 76–86, Mar. 2009. [Online]. Available: https://doi.org/10.1109/MIS.2009.26

[447] G. Schreiber, "Principles for Knowledge Engineering on the Web," in *Knowledge Engineering: Practice and Patterns*, ser. Lecture Notes in Computer Science, A. Gangemi and J. Euzenat, Eds. Berlin, Heidelberg: Springer, 2008, pp. 6–6. [Online]. Available: https://doi.org/10.1007/978-3-540-87696-0_3

[448] A. Tordai, B. Omelayenko, and G. Schreiber, “Thesaurus and metadata alignment for a semantic e-culture application,” in *Proceedings of the 4th international conference on Knowledge capture*, ser. K-CAP ’07. Whistler, BC, Canada: Association for Computing Machinery, Oct. 2007, pp. 199–200. [Online]. Available: https://doi.org/10.1145/1298406.1298453

[449] L. van der Meij, A. Isaac, and C. Zinn, “A Web-Based Repository Service for Vocabularies and Alignments in the Cultural Heritage Domain,” in *The Semantic Web: Research and Applications*, ser. Lecture Notes in Computer Science, L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, and T. Tudorache, Eds. Berlin, Heidelberg: Springer, 2010, pp. 394–409. [Online]. Available: https://doi.org/10.1007/978-3-642-13486-9_27

[450] D. M. Weigl, D. Lewis, T. Crawford, I. Knopke, and K. R. Page, “On providing semantic alignment and unified access to music library metadata,” *International Journal on Digital Libraries*, vol. 20, no. 1, pp. 25–47, Mar. 2019. [Online]. Available: https://doi.org/10.1007/s00799-017-0223-9

[451] A. Stolk, L. Verhagen, and I. Toni, “Conceptual Alignment: How Brains Achieve Mutual Understanding,” *Trends in Cognitive Sciences*, vol. 20, no. 3, pp. 180–191, Mar. 2016. [Online]. Available: https://doi.org/10.1016/j.tics.2015.11.007

[452] I. Toni and A. Stolk, “Conceptual alignment as a neurocognitive mechanism for human communicative interactions,” in *Human Language: From Genes and Brains to Behavior.* Cambridge, MA: MIT Press, 2019, pp. 249–256.

[453] U. Priss, “Conceptual Alignment with Formal Concept Analysis,” in *ICFCA 2019 – Proceedings ICFCA 2019 Conference and Workshops*, vol. 2378. Frankfurt, Germany: CEUR Workshop Proceedings, 2019, pp. 14–27. [Online]. Available: http://ceur-ws.org/Vol-2378/longICFCA2.pdf

[454] J. Biggs, “Enhancing teaching through constructive alignment,” *Higher Education*, vol. 32, no. 3, pp. 347–364, Oct. 1996. [Online]. Available: https://doi.org/10.1007/BF00138871

[455] H. Larkin and B. Richardson, “Creating high challenge/high support academic environments through constructive alignment: student outcomes,” *Teaching in Higher Education*, vol. 18, no. 2, pp. 192–204, Feb. 2013. [Online]. Available: https://doi.org/10.1080/13562517.2012.696541

[456] K. Trigwell and M. Prosser, "Qualitative variation in constructive alignment in curriculum design," *Higher Education*, vol. 67, no. 2, pp. 141–154, Feb. 2014. [Online]. Available: https://doi.org/10.1007/s10734-013-9701-1

[457] A. Walsh, "An exploration of Biggs' constructive alignment in the context of work-based learning," *Assessment & Evaluation in Higher Education*, vol. 32, no. 1, pp. 79–87, Feb. 2007. [Online]. Available: https://doi.org/10.1080/02602930600848309

[458] X. Wang, Y. Su, S. Cheung, E. Wong, and T. Kwong, "An exploration of Biggs' constructive alignment in course design and its impact on students' learning approaches," *Assessment & Evaluation in Higher Education*, vol. 38, no. 4, pp. 477–491, Jun. 2013. [Online]. Available: https://doi.org/10.1080/02602938.2012.658018

[459] M. L. Zeng and J. Qin, *Metadata*, 2nd ed. New York: Neal-Schuman/American Library Association, 2014, oCLC: 894201488.

[460] R. V. Guha, D. Brickley, and S. Macbeth, "Schema.org: evolution of structured data on the web," *Communications of the ACM*, vol. 59, no. 2, pp. 44–51, Jan. 2016. [Online]. Available: https://doi.org/10.1145/2844544

[461] B. Mecum, R. Nenuji, M. B. Jones, D. Vieglais, and M. Schildhauer, "DataONE on the web: Using Schema.org and JSON-LD to enhance data search and access," *AGU Fall Meeting Abstracts*, vol. 31, Dec. 2018. [Online]. Available: http://adsabs.harvard.edu/abs/2018AGUFMIN31B..29M

[462] U. Şimşek, E. Kärle, O. Holzknecht, and D. Fensel, "Domain Specific Semantic Validation of Schema.org Annotations," in *Perspectives of System Informatics*, ser. Lecture Notes in Computer Science, A. K. Petrenko and A. Voronkov, Eds. Cham: Springer International Publishing, 2018, pp. 417–429. [Online]. Available: https://doi.org/10.1007/978-3-319-74313-4_31

[463] W. Sun, X. Zhang, C. J. Guo, P. Sun, and H. Su, "Software as a Service: Configuration and Customization Perspectives," in *2008 IEEE Congress on Services Part II (services-2 2008)*, Sep. 2008, pp. 18–25, iSSN: null. [Online]. Available: https://doi.org/10.1109/SERVICES-2.2008.29

[464] M. Yuvaraj, “Cloud Computing Software and Solutions for Libraries: A Comparative Study,” *Journal of Electronic Resources in Medical Libraries*, vol. 12, no. 1, pp. 25–41, Jan. 2015. [Online]. Available: https://doi.org/10.1080/15424065.2014.1003479

[465] A. Miles and D. Brickley, “SKOS Mapping Vocabulary Specification,” 2004. [Online]. Available: https://www.w3.org/2004/02/skos/mapping/spec/2004-11-11.html

[466] R. Hoekstra, “BestMap: context-aware SKOS vocabulary mappings in OWL 2,” *OWLED’09 Proceedings of the 6th International Conference on OWL: Experiences and Directions*, vol. 529, 2009. [Online]. Available: https://hdl.handle.net/11245/1.315856

[467] T. Baker, S. Bechhofer, A. Isaac, A. Miles, G. Schreiber, and E. Summers, “Key choices in the design of Simple Knowledge Organization System (SKOS),” *Journal of Web Semantics*, vol. 20, pp. 35–49, May 2013. [Online]. Available: https://doi.org/10.1016/j.websem.2013.05.001

[468] B. Haslhofer, F. Martins, and J. Magalhães, “Using SKOS vocabularies for improving web search,” in *Proceedings of the 22nd International Conference on World Wide Web*, ser. WWW ’13 Companion. Rio de Janeiro, Brazil: Association for Computing Machinery, May 2013, pp. 1253–1258. [Online]. Available: https://doi.org/10.1145/2487788.2488159

[469] D. A. Evans, S. K. Handerson, I. A. Monarch, J. Pereiro, L. Delon, and W. R. Hersh, “Mapping Vocabularies Using Latent Semantics,” in *Cross-Language Information Retrieval*, ser. The Springer International Series on Information Retrieval, G. Grefenstette, Ed. Boston, MA: Springer US, 1998, pp. 63–80. [Online]. Available: https://doi.org/10.1007/978-1-4615-5661-9_6

[470] F. E. Masarie, R. A. Miller, O. Bouhaddou, N. B. Giuse, and H. R. Warner, “An interlingua for electronic interchange of medical information: Using frames to map between clinical vocabularies,” *Computers and Biomedical Research*, vol. 24, no. 4, pp. 379–400, Aug. 1991. [Online]. Available: https://doi.org/10.1016/0010-4809(91)90035-U

[471] F. Giunchiglia, D. Soergel, V. Maltese, and A. Bertacco, “Mapping large-scale Knowledge Organization Systems,” University of Trento, Trento,

Departmental Technical Report, May 2009. [Online]. Available: http://eprints.biblio.unitn.it/1616/

[472] B. Lauser, G. Johannsen, C. Caracciolo, W. R. v. Hage, J. Keizer, and P. Mayr, "Comparing Human and Automatic Thesaurus Mapping Approaches in the Agricultural Domain," *International Conference on Dublin Core and Metadata Applications*, pp. 43–53, Sep. 2008. [Online]. Available: https://dcpapers.dublincore.org/pubs/article/view/918

[473] J. van Ossenbruggen, M. Hildebrand, and V. de Boer, "Interactive Vocabulary Alignment," in *Research and Advanced Technology for Digital Libraries*, ser. Lecture Notes in Computer Science, S. Gradmann, F. Borri, C. Meghini, and H. Schuldt, Eds. Berlin, Heidelberg: Springer, 2011, pp. 296–307.

[474] J. C. Dos Reis, C. Pruski, and C. Reynaud-Delaître, "State-of-the-art on mapping maintenance and challenges towards a fully automatic approach," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1465–1478, Feb. 2015. [Online]. Available: https://doi.org/10.1016/j.eswa.2014.08.047

[475] S. Gupta, P. Szekely, C. A. Knoblock, A. Goel, M. Taheriyan, and M. Muslea, "Karma: A System for Mapping Structured Sources into the Semantic Web," in *The Semantic Web: ESWC 2012 Satellite Events*, ser. Lecture Notes in Computer Science, E. Simperl, B. Norton, D. Mladenic, E. Della Valle, I. Fundulaki, A. Passant, and R. Troncy, Eds. Berlin, Heidelberg: Springer, 2015, pp. 430–434.

[476] A. Joorabchi, M. English, and A. E. Mahdi, "Automatic mapping of user tags to Wikipedia concepts: The case of a Q&A website – StackOverflow," *Journal of Information Science*, vol. 41, no. 5, pp. 570–583, Oct. 2015. [Online]. Available: https://doi.org/10.1177/0165551515586669

[477] A. Ballatore, M. Bertolotto, and D. C. Wilson, "Linking geographic vocabularies through WordNet," *Annals of GIS*, vol. 20, no. 2, pp. 73–84, Apr. 2014. [Online]. Available: https://doi.org/10.1080/19475683.2014.904440

[478] E. L. Tonkin, S. Taylor, and G. J. L. Tourte, "Cover sheets considered harmful," *Information Services & Use*, vol. 33, no. 2, pp. 129–137, Jan. 2013. [Online]. Available: https://doi.org/10.3233/ISU-130705

[479] S. Foo and Y.-L. Theng, "A snapshot of digital library development: The way forward in the asia pacific," in *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific.* IGI Global, pp. 351–370, ISBN: 9781591404415 Pages: 351-370 Publisher: IGI Global. [Online]. Available: https://doi.org/10.4018/978-1-59140-441-5.ch020

[480] J. Barton, "Digital libraries, virtual museums: same difference?" vol. 54, no. 3, pp. 149–154. [Online]. Available: https://doi.org/10.1108/00242530510588908

[481] O. L. Zavalina, "Complementarity in subject metadata in large-scale digital libraries: A comparative analysis," vol. 52, no. 1, pp. 77–89. [Online]. Available: https://doi.org/10.1080/01639374.2013.848316

[482] R. Fattahi and E. Afshar, "Added value of information and information systems: a conceptual approach," vol. 55, no. 2, pp. 132–147. [Online]. Available: https://doi.org/10.1108/00242530610649620

[483] O. L. Zavalina, "Exploring the richness of collection-level subject metadata in three large-scale digital libraries," vol. 7, no. 3, pp. 209–221, publisher: Inderscience Publishers. [Online]. Available: https://doi.org/10.1504/IJMSO.2012.050182

[484] H. Park, "A conceptual framework to study folksonomic interaction," vol. 38, no. 6. [Online]. Available: https://doi.org/10.5771/0943-7444-2011-6-515

[485] P. Mayr and V. Petras, "Cross-concordances: terminology mapping and its effectiveness for information retrieval." Ithaca, NY.: arXiv.org, Jun. 2008, arXiv: 0806.3765. [Online]. Available: http://arxiv.org/abs/0806.3765

[486] F. Boteram and J. Hubrich, "Specifying intersystem relations: Requirements, strategies, and issues," vol. 37, no. 3, pp. 216–222. [Online]. Available: https://www.nomos-elibrary.de/index.php?doi=10.5771/0943-7444-2010-3-216

[487] T. Zschocke, "Subject classification with DITA markup for agricultural learning resources: A case example in agroforestry," in *Metadata and Semantic Research*, E. García-Barriocanal, Z. Cebeci, M. C. Okur, and A. Öztürk, Eds. Springer Berlin Heidelberg, vol. 240, pp. 500–513, series Title: Communications in Computer and Information Science. [Online]. Available: https://doi.org/10.1007/978-3-642-24731-6_49

[488] S. Cormont, P.-Y. Vandenbussche, A. Buemi, J. Delahousse, E. Lepage, and J. Charlet, “Implementation of a platform dedicated to the biomedical analysis terminologies management,” vol. 2011, pp. 1418–1427. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243140/

[489] P.-Y. Vandenbussche, S. Cormont, C. André, C. Daniel, J. Delahousse, J. Charlet, and E. Lepage, “Implementation and management of a biomedical observation dictionary in a large healthcare information system,” vol. 20, no. 5, pp. 940–946. [Online]. Available: https://doi.org/10.1136/amiajnl-2012-001410

[490] G. Macgregor, “Introduction to a special issue on digital libraries and the semantic web : context, applications and research,” *Library Review*, vol. 57, pp. 173–177, Mar. 2008. [Online]. Available: https://doi.org/10.1108/00242530810865457

[491] U. of Strathclyde, “Enhancement-led institutional review 2019: Reflective analysis.”

[492] R. Gartner, M. Cox, and K. Jeffery, “A cerif-based schema for recording research impact,” *The Electronic Library*, vol. 31, no. 4, pp. 465–482, Jan. 2013. [Online]. Available: https://doi.org/10.1108/EL-11-2011-0156

[493] Z. Zahedi, R. Costas, and P. Wouters, “How well developed are altmetrics? A cross-disciplinary analysis of the presence of ‘alternative metrics’ in scientific publications,” *Scientometrics*, vol. 101, no. 2, pp. 1491–1513, Nov. 2014. [Online]. Available: https://doi.org/10.1007/s11192-014-1264-0

[494] A. Tay. 6 ”inside-out” activities librarians are doing. [Online]. Available: http://musingsaboutlibrarianship.blogspot.com/2018/12/6-inside-out-activities-librarians-are.html

[495] D. Walker, “Libraries and the REF: how do librarians contribute to research excellence?” vol. 33, no. 6, number: 6 Publisher: Ubiquity Press. [Online]. Available: https://doi.org/10.1629/uksg.497

[496] K. Arlitsch, J. Wheeler, M. T. N. Pham, and N. N. Parulian, “An analysis of use and performance data aggregated from 35 institutional repositories,” vol. ahead-of-print. [Online]. Available: https://doi.org/10.1108/OIR-08-2020-0328

Appendix A

Candidate's full publication list

- Nixon, W. J., Andrew, T., Macgregor, G., & Proven, J. (Accepted/In press). The wee country that roared: supporting Open Access in Scotland through institutional repositories. Paper presented at 16th International Open Repositories Conference (OR2021), Stellenbosch, South Africa. Available: http://strathprints.strath.ac.uk/71842/

- Macgregor, G., & Neugebauer, T. (2020). Preserving digital content through improved EPrints repository integration with Archivematica. 1-10. UK Archivematica User Group, Warwick, United Kingdom. Available: http://strathprints.strath.ac.uk/73978/

- Macgregor, G. (2020). Open Science: an exploration of Open Access and Open Data concepts. 1-31. CILIPS Online Learning sessions [Virtual]. Available: http://strathprints.strath.ac.uk/72079/

- Macgregor, G. (2020). Enhancing content discovery of open repositories: an analytics-based evaluation of repository optimizations. Publications, 8(1), [8]. Available: http://doi.org/10.3390/publications8010008

- Macgregor, G. (2019). Repository and CRIS interoperability issues within a 'connector lite' environment. 14th International Conference on Open Repositories (OR2019), Hamburg, Germany. Available: http://strathprints.strath.ac.uk/68240/

- Macgregor, G. (2019). Promoting content discovery of open repositories: reviewing the impact of optimization techniques (2016-2019). 1-10. 14th International Conference on Open Repositories (OR2019), Hamburg, Germany. Available: http://doi.org/10.17868/67963

- Macgregor, G. (2019). Improving the discoverability and web impact of open repositories: techniques and evaluation. Code4Lib Journal, (43). Available: `http://journal.code4lib.org/articles/14180`

- Macgregor, G. (2018). Repository optimisation & techniques to improve discoverability and web impact: an evaluation. (pp. 1-13). University of Strathclyde. Available: `http://doi.org/10.17868/65389`

- De Castro, P., Morrison, A., Macgregor, G. , & Repositories, Open Access & Datasets Team. (2018). Integrated Workflow for Open Access publishing and Research Data Management at the University of Strathclyde. Digital or Visual Products, University of Strathclyde. Available: `http://strathprints.strath.ac.uk/65503/`

- Macgregor, G. (2017). Reviewing repository discoverability: approaches to improving repository visibility and web impact. Poster session presented at Repository Fringe 2017, Edinburgh, United Kingdom. Available: `http://strathprints.strath.ac.uk/61333/`

- Hibbert, D., & Macgregor, G. (Ed.) (2016). RCUK Open Access Report - 2015/2016. University of Strathclyde.

- Macgregor, G. (2016). Disambiguating Yourself: Online Identity Management for Researchers. (pp. 1-17). University of Strathclyde. Available: `http://strathprints.strath.ac.uk/60205/`

- Macgregor, G., & Sheppard, I. (2015). Fraser Economic Commentary - The complete catalogue of reviews, outlooks and articles: 1975-2015. Fraser of Allander Economic Commentary, 39(2), 57-67. Available: `http://strathprints.strath.ac.uk/54776/`

- Hibbert, D., & Macgregor, G. (Ed.) (2015). RCUK Open Access Report - 2014/2015. University of Strathclyde.

- Macgregor, G., & Sheppard, I. (2015). Fraser Economic Commentary: Catalogue of all reviews, outlooks and articles, Part 2 1991 - 2000. Fraser of Allander Economic Commentary, 39(1), 52-54. Available: `http://strathprints.strath.ac.uk/53543/`

- Macgregor, G., & Sheppard, I. (2015). Fraser Economic Commentary: catalogue of all reviews, outlooks and articles, part 1 1975 - 1990. Fraser of Allander Economic Commentary, 38(3), 62-66. Available: http://strathprints.strath.ac.uk/52089/

- Macgregor, G. (2014). 2014 Independent Review of the Implementation of the RCUK Policy on Open Access: University of Strathclyde RCUK Open Access Report. University of Strathclyde. Available: http://strathprints.strath.ac.uk/60245/

- Macgregor, G. (2013). ePrints and PURE: Discussion Paper. (pp. 1-37). University of Strathclyde. Available: http://strathprints.strath.ac.uk/60311/

- Macgregor, G. (2013). Introducing C-CAP: [Principles in Patterns Project - PiP]. Digital or Visual Products, University of Strathclyde. Available: http://strathprints.strath.ac.uk/59854/

- Everett, J., Macgregor, G., & Cullen, D. (2013). Institutional Approaches to Curriculum Design Institutional Story: [Principles in Patterns Project - PiP]. University of Strathclyde. Available: http://strathprints.strath.ac.uk/59850/

- Everett, J., Macgregor, G., & Cullen, D. (2013). Principles in Patterns (PiP): Institutional Approaches to Curriculum Design Institutional Story. University of Strathclyde. Available: http://strathprints.strath.ac.uk/44737/

- Macgregor, G. (2013). Revisiting user acceptance... or resistance? Insights from tech-supported curriculum design. University of Strathclyde. Available: http://strathprints.strath.ac.uk/70392/

- Macgregor, G. (2012). C-CAP: Managing Curriculum Designs as Knowledge Assets – Briefing Paper. University of Strathclyde. Available: http://strathprints.strath.ac.uk/59851/

- Everett, J., Macgregor, G., & Mohamed, R. (2012). An incremental approach to technology-supported curriculum design and approval. Paper presented at IADIS International Conference WWW/Internet 2012, Madrid, Spain. Available: http://strathprints.strath.ac.uk/46292/

- Macgregor, G. (2012). Understanding the social system when embedding tech-supported curriculum design and approval. University of Strathclyde. Available: http://strathprints.strath.ac.uk/69721/

- Macgregor, G. (2012). A Quick Guide for AQ Staff: Using the C-CAP Administration Dashboard - C-CAP Embedding Phase . University of Strathclyde. Available: http://strathprints.strath.ac.uk/59852/

- Macgregor, G. (2012). Principles in Patterns (PiP): C-CAP Embedding and Work Plan. University of Strathclyde. Available: http://strathprints.strath.ac.uk/59853/

- Macgregor, G. (2012). Thin end of the wedge: system resistance and its implications for C-CAP. University of Strathclyde. Available: http://strathprints.strath.ac.uk/69720/

- Macgregor, G. (2012). Academic quality: at the centre of the curriculum approval universe. University of Strathclyde. Available: http://strathprints.strath.ac.uk/70393/

- Macgregor, G. (2012). Process as myth: understanding the mythic core of organisational process with ideal types. University of Strathclyde. Available: http://strathprints.strath.ac.uk/69605/

- Macgregor, G. (2012). Principles in Patterns (PiP): Project Evaluation Synthesis. University of Strathclyde. Available: http://strathprints.strath.ac.uk/46513/

- Macgregor, G. (2012). Principles in Patterns (PiP): Piloting of C-CAP - Evaluation of Impact and Implications for System and Process Development. University of Strathclyde. Available: http://strathprints.strath.ac.uk/46511/

- Macgregor, G. (2012). Information Users and Usability in the Digital Age. Library Review, 61(5), 381-383. Available: http://doi.org/10.1108/00242531211280513

- Macgregor, G. (2012, May 9). Two sides of the same coin: qualitative and quantitative approaches to process analysis. University of Strathclyde. Available: http://strathprints.strath.ac.uk/69608/

- Macgregor, G. (2012, Apr 25). Principles in Patterns (PiP): Evaluation Approach. JISC Webinar: Institutional Approaches to Curriculum Design. Available: http://strathprints.strath.ac.uk/59936/

- Macgregor, G. (2012). Principles in Patterns (PiP): Evaluation of Impact on Business Processes. University of Strathclyde. Available: http://strathprints.strath.ac.uk/46512/

- Macgregor, G. (2012). Innovations in Information Retrieval: Perspectives for Theory and Practice. Library Review, 61(3), 233-235. Available: http://doi.org/10.1108/00242531211259364

- Macgregor, G. (2012, Mar 13). Evaluating C-CAP: reflecting on the dichotomy of curriculum design and approval. University of Strathclyde. Available: http://strathprints.strath.ac.uk/69607/

- Macgregor, G. (2012, Feb 12). Evaluating PiP: optimising user acceptance testing via heuristic evaluation. University of Strathclyde. Available: http://strathprints.strath.ac.uk/69606/

- Macgregor, G. (2012). Principles in Patterns (PiP): User Acceptance Testing of Course and Class Approval Online Pilot (C-CAP). University of Strathclyde. Available: http://strathprints.strath.ac.uk/46510/

- Macgregor, G. (2011). Principles in Patterns (PiP): Heuristic Evaluation of Course and Class Approval Online Pilot (C-CAP). University of Strathclyde. Available: http://strathprints.strath.ac.uk/46509/

- Macgregor, G. (2011). Knowledge Representation in the Social Semantic Web. Library Review, 60(8), 723-735. Available: http://doi.org/10.1108/00242531111166764

- Macgregor, G., & Spiers, A. (2011). Media enhanced learning at LJMU: evaluating audio feedback in formative assessment. 1-29. Paper presented at Media Enhanced Learning SIG, Liverpool, United Kingdom. Available: http://strathprints.strath.ac.uk/59958/

- Macgregor, G. (2011). The Accidental Taxonomist. Library Review, 60(1), 86-88. Available: http://doi.org/10.1108/00242531111100630

- Macgregor, G., Spiers, A., & Taylor, C. (2011). Exploratory evaluation of audio email technology in formative assessment feedback. Research in Learning Technology, 19(1), 39-59. Available: http://doi.org/10.1080/09687769.2010.547930

- Macgregor, G. (2010). Folksonomies: Indexing and Retrieval in Web 2.0. Library Review, 59(7), 566-568. Available: http://doi.org/10.1108/00242531011065181

- Macgregor, G., & Spiers, A. (2010). Enhancing the student learning experience: evaluating the impact of voice email technology in formative assessment strategy. 1-29. Paper presented at School of Sport and Exercise Sciences Learning and Teaching Day, Liverpool, United Kingdom. Available: http://strathprints.strath.ac.uk/59963/

- Spiers, A., & Macgregor, G. (2010). "It's as if the student is in front of you" - Using Wimba Voice email for feedback on formative assessment. 1-29. Paper presented at Wimba Study Break Webinar. Available: http://strathprints.strath.ac.uk/59959/

- Macgregor, G. (2010). The Future of Information Architecture: Conceiving a Better Way to Understand Taxonomy, Network and Intelligence. Library Review, 59(3), 231-234. Available: http://doi.org/10.1108/00242531011031223

- Macgregor, G., Spiers, A., & Taylor, C. (2010). Investigating 'voice email' technology efficacy in information management assessment. In 11th Annual Conference of the Subject Centre for Information and Computer Sciences (pp. 162-166). Higher Education Academy. Available: http://strathprints.strath.ac.uk/56790/

- Spiers, A., & Macgregor, G. (2009). Using audio email feedback in formative assessment. 1-19. Paper presented at A Word in Your Ear 2009, Sheffield, United Kingdom. Available: http://strathprints.strath.ac.uk/59960/

- Macgregor, G. (2009). Metadata. Library Review, 58(8), 621-623. Available: http://doi.org/10.1108/00242530910987136

- Macgregor, G. (2009). Ontology and the Semantic Web. Library Review, 58(2), 141-143. Available: http://doi.org/10.1108/00242530910936998

- Macgregor, G. (2009). E-resource management and the Semantic Web: applications of RDF for e-resource discovery. In The E-Resources Management Handbook - UKSG (pp. 1-20). UKSG. Available: http://doi.org/10.1629/9552448-0-3.20.1

- Macgregor, G., Spiers, A., & Taylor, C. (2009). Exploring the Efficacy of Audio Email Feedback in Information Management Assessment (ExAEF Project): Final Report. The Higher Education Academy Subject Centre for Information and Computer Sciences. Available: http://strathprints.strath.ac.uk/56771/

- Macgregor, G., & Turner, J. (2009). Revisiting e-learning effectiveness: proposing a conceptual model. Interactive Technology and Smart Education, 6(3), 156-172. Available: http://doi.org/10.1108/17415650911005375

- Macgregor, G. (2008). Metadata and its Applications in the Digital Library: Approaches and Practices. New Library World, 109, 295-297. Available: http://doi.org/10.1108/03074800810873650

- Macgregor, G. (2008). Annual Review of Information Science and Technology: Volume 41. Library Review, 57(4), 323-327. Available: http://doi.org/10.1108/00242530810868797

- Macgregor, G. (2008). Introduction to a special issue on digital libraries and the semantic web: context, applications and research. Library Review, 57(3), 173-177. Available: http://doi.org/10.1108/00242530810865457

- McCulloch, E., & Macgregor, G. (2008). Analysis of equivalence mapping for terminology services. Journal of Information Science, 34(1), 70-92. Available: http://doi.org/10.1177/0165551507079130

- Macgregor, G., McCulloch, E., & Nicholson, D. (2007). Terminology server for improved resource discovery: analysis of model and functions. Paper presented at Second International Conference on Metadata and Semantics Research, Corfu, Greece. Available: http://strathprints.strath.ac.uk/3435/

- Macgregor, G. (2007). High-Level Thesaurus (HILT) project, phase IV. Paper presented at JISC Shared Infrastructure Services Workshop, London, United Kingdom. Available: http://strathprints.strath.ac.uk/71129/

- Macgregor, G. (2007). How to do Research: A Practical Guide to Designing and Managing Research Projects : 3rd Revised Edition. Library Review, 56(4), 337-339. Available: http://doi.org/10.1108/00242530710743589

- Macgregor, G., Joseph, A., & Nicholson, D. (2007). A SKOS Core approach to implementing an M2M terminology mapping server. In A. R. D. Prasad, & D. P. Madalli (Eds.), International Conference on Semantic Web & Digital Libraries (ICSD 2007): ICSD 2007, 21-23 February 2007 (pp. 109-120). Documentation Research & Training Centre. Available: http://strathprints.strath.ac.uk/2970/

- Nicholson, D., McCulloch, E., & Macgregor, G. (2007). High-level Thesaurus (HILT) Phase III: Completion Report. University of Strathclyde. Available: http://strathprints.strath.ac.uk/71140/

- Nicholson, D., McCulloch, E., & Macgregor, G. (2007). High-level Thesaurus (HILT) Phase III [Project]: Final Report. University of Strathclyde. Available: http://strathprints.strath.ac.uk/71139/

- Macgregor, George and McCulloch, Emma & Nicholson, Dennis; Isaac, A., Phipps, J. & Rubin, D., eds. (2007) High-level Thesaurus (HILT) : edited use case, Semantic Web Deployment. In: SKOS Use Cases and Requirements. W3C, Cambridge, MA. Available: http://strathprints.strath.ac.uk/74724/

- Macgregor, G. (2007). The Content Management Handbook. Library Review, 56(3), 262-263. Available: http://doi.org/10.1108/00242530710736136

- Macgregor, G. (2007). Virtual Methods: Issues in Social Research on the Internet. Library Review, 56(9), 836-838. Available: http://doi.org/10.1108/00242530710831338

- Nicholson, D., McCulloch, E., & Macgregor, G. (2006). High-level Thesaurus (HILT) Phase III [Final Report]: Evaluation Report. University of Strathclyde. Available: http://strathprints.strath.ac.uk/71138/

- Dunsire, G., Macgregor, G., & Thomson, R. (2006). Subject Classification of Collection-level Descriptions Using DDC for Information Landscaping. (pp. 1-16). University of Strathclyde. Available: http://strathprints.strath.ac.uk/59972/

- Macgregor, G. (2006). UNESCO Thesaurus to DDC Mappings: Third Summary – Thousand Sections [HILT Phase III]. University of Strathclyde. Available: `http://strathprints.strath.ac.uk/71136/`

- Macgregor, G. (2006). Basic Research Methods for Librarians – 4th Edition. Library Review, 55(6), 375-376. Available: `http://doi.org/10.1108/00242530610674785`

- Macgregor, G., & McCulloch, E. (2006). Summary of High-level Thesaurus (HILT) Mapping Work - Principal Results. University of Strathclyde. Available: `http://strathprints.strath.ac.uk/71137/`

- Macgregor, G. (2006). Essential Classification. Library Review, 55(1), 75-76. Available: `http://doi.org/10.1108/00242530610641817`

- Macgregor, G., & McCulloch, E. (2006). Collaborative tagging as a knowledge organisation and resource discovery tool. Library Review, 55(5), 291-300. Available: `http://doi.org/10.1108/00242530610667558`

- Nicholson, Dennis & Macgregor, George (2006) Ensuring Interoperable Digital Object Management Metadata in Scotland : Report of the SLIC-funded CMS Metadata Interoperability Project : Findings, Conclusions, and Guidelines for Best Practice. University of Strathclyde. Available: `http://strathprints.strath.ac.uk/74725/`

- Nicholson, D., Dunsire, G., & Macgregor, G. (2006). SPEIR: developing a common information environment in Scotland. Electronic Library, 24(1), 94-107. Available: `http://doi.org/10.1108/02640470610649272`

- Macgregor, G., & McCulloch, E. (2005). Popularity over Relevance in Collaborative Tagging Systems for General Resource Discovery. (pp. 1-5). University of Strathclyde. Available: `http://strathprints.strath.ac.uk/59962/`

- Macgregor, G. (2005). The Ultimate Digital Library: Where the New Information Players Meet. Library Review, 54(6), 390-391. Available: `http://doi.org/10.1108/00242530510605557`

- Macgregor, G. (2005). Libricide: The Regime-sponsored Destruction of Books and Libraries in the Twentieth Century. Library Review, 54(5), 332-334. Available: `http://doi.org/10.1108/00242530510600589`

- Nicholson, D., McCulloch, E., Dawson, A., Joseph, A., Macgregor, G., Dunsire, G., Rees, C., Medyckyj-Scott, D., Boyle, E. G., Soares, B., Stickland, T., Parkinson, B., McKay, D., Heery, R., Powell, A., Hutchison, B., Peacock, D., & Will, L. (2005). HILT : High-Level Thesaurus Project M2M Feasibility Study: [Final Report]. University of Strathclyde. Available: `http://strathprints.strath.ac.uk/71141/`

- Law, D., Macgregor, G., McCulloch, E., & Wallis, J. (2005). Developing a national information strategy in Scotland. Cadernos de Biblioteconomia, Arquivística e Documentação, 1, 49-53. Available: `http://strathprints.strath.ac.uk/1969/`

- Macgregor, G., & McGill, L. (2005). Digital libraries and information literacy issues within virtual learning environments: an e-learning impasse?. Paper presented at Librarians' Annual Information Literacy Conference (LILAC), London, UK. Available: `http://strathprints.strath.ac.uk/2334/`

- Macgregor, G. (2005). Information Literacy: Essential Skills for the Information Age – 2nd Edition. Library Review, 54(9), 532-534. Available: `http://doi.org/10.1108/00242530510629560`

- Macgregor, G., & Dunsire, G. (2005). Library systems: the trends, the developments, the future. E-MmITS: Newsletter of the Multimedia and Information Technology Group Scotland, 2005(Winter). Available: `http://strathprints.strath.ac.uk/6037/`

- Macgregor, G. (2005). The nature of information in the 21st century: conundrums for the informatics community? Library Review, 54(1), 10-23. Available: `http://doi.org/10.1108/00242530510574129`

- Macgregor, G., & Nicolaides, F. (2005). Towards improved performance and interoperability in distributed and physical union catalogues. Program, 39(3), 227-247. Available: `http://doi.org/10.1108/00330330510610573`

- Macgregor, G. (2005). Z39.50 broadcast searching and Z-server response times: perspectives from CC-interop. Online Information Review, 29(1), 90-106. Available: `http://doi.org/10.1108/14684520510583963`

- Nicolaides, F., & Macgregor, G. (2004). Interoperability: the performance of institutional catalogues & strategies for improvement. Paper presented at Hyper Clumps, Mini Clumps and National Catalogues: Resource Discovery for the 21st Century, London, United Kingdom. Available: `http://strathprints.strath.ac.uk/57714/`

- Macgregor, G., & Dunsire, G. (2004, Oct 13). CC-interop: a post mortem. University of Strathclyde. Available: `http://strathprints.strath.ac.uk/59961/`

- Macgregor, G. (2004). Uncanny Networks: Dialogues with the Virtual Intelligentsia. Library Review, 53(6), 335-336. Available: `http://doi.org/10.1108/00242530410544466`

- Macgregor, G. (2004). Our Modern Times: The New Nature of Capitalism in the Information Age. Library Review, 53(4), 243-244. Available: `http://doi.org/10.1108/00242530410531956`

- Dunsire, G., & Macgregor, G. (2004). Improving Interoperability in Distributed and Physical Union Catalogues through Co-ordination of Cataloguing and Indexing Policies: Report for Work Package B of the JISC CC-interop Project . University of Strathclyde. Available: `http://strathprints.strath.ac.uk/57653/`

- Macgregor, G. (2004). Introduction to Digital Libraries. Library Review, 53(1), 66-67. Available: `http://doi.org/10.1108/00242530410514829`

- Nicholson, D., Dunsire, G., Dawson, A., Macgregor, G., Shiri, A., Joseph, A., Williamson, A., Jones, E., & SLIC (Scottish Library and Information Council) (Funder) (2004). SPEIR: Scottish Portals for Education, Information and Research. Final Project Report: Elements and Future Development Requirements of a Common Information Environment for Scotland. University of Strathclyde. Available: `http://strathprints.strath.ac.uk/14124/`

- Dunsire, G., & Macgregor, G. (2003). Make it easy. Information Scotland, 1(6). Available: `http://strathprints.strath.ac.uk/6016/`

- Macgregor, G. (2003). Scottish Collections Access Management Portal: a staff portal to support collaborative collecting. Paper presented at "Electric Connections", the COSMIC and SPEIR Joint Conference, Perth, United Kingdom. Available: `http://strathprints.strath.ac.uk/57715/`

- Nicholson, D., & Macgregor, G. (2003). 'NOF-Digi': putting UK culture online. OCLC Systems and Services, 19(3), 96-99. Available: `http://doi.org/10.1108/10650750310490298`

- Dunsire, G., & Macgregor, G. (2003). Clumps and collection description in the information environment in the UK with particular reference to Scotland. Program, 37(4), 218-225. Available: `http://doi.org/10.1108/00330330310500694`

- Macgregor, G. (2003). Collection-level description: metadata of the future? Library Review, 52(6), 247-250. Available: `http://doi.org/10.1108/00242530310482015`

- Nicholson, D., & Macgregor, G. (2003). Developing the Scottish cooperative infrastructure: the what, who, where, when and why of SPEIR. WIDWISAWN, 1(2).

- Nicholson, D., & Macgregor, G. (2002). Learning lessons holistically in the Glasgow Digital Library. D-Lib Magazine, 8(7/8). Available: `http://doi.org/10.1045/july2002-nicholson`

- Nicholson, D., Dawson, A., & Macgregor, G. (2002). GDL: a model infrastructure for a regional digital library. WIDWISAWN, 1(1).

Appendix B

Published works: PW1-PW11

1. **Published work 1 (PW1)** Macgregor, G. (2003) Collection-level description: metadata of the future? *Library Review*, 52 (6). pp. 247-250. Available: https://doi.org/10.1108/00242530310482015

2. **Published work 2 (PW2)** Macgregor, George and Nicolaides, Fraser (2005) Towards improved performance and interoperability in distributed and physical union catalogues. *Program*, 39 (3). pp. 227-247. Available: https://doi.org/10.1108/00330330510610573

3. **Published work 3 (PW3)** Macgregor, George (2005) Z39.50 broadcast searching and Z-server response times: perspectives from CC-interop. *Online Information Review*, 29 (1). pp. 90-106. Available: https://doi.org/10.1108/14684520510583963

4. **Published work 4 (PW4)** Macgregor, George and McCulloch, Emma (2006) Collaborative tagging as a knowledge organization and resource discovery tool. *Library Review*, 55 (5). pp. 291-300. Available: https://doi.org/10.1108/00242530610667558

5. **Published work 5 (PW5)** Macgregor, George (2009) E-resource management and the Semantic Web : applications of RDF for e-resource discovery. In: *The E-Resources Management Handbook - UKSG*. UKSG, Newbury, pp. 1-20. ISBN 9780955244803. Available: https://doi.org/10.1629/9552448-0-3.20.1

6. **Published work 6 (PW6)** Macgregor, George and Joseph, Anu and Nicholson, Dennis; Prasad, A.R.D and Madalli, Devika P., eds. (2007) A SKOS Core

approach to implementing an M2M terminology mapping server. In: *International Conference on Semantic Web & Digital Libraries (ICSD 2007)*. Documentation Research & Training Centre, Bangalore, India, pp. 109-120. Available: `https://strathprints.strath.ac.uk/2970/`

7. **Published work 7 (PW7)** Macgregor, G. and McCulloch, E. and Nicholson, D. (2007) Terminology server for improved resource discovery: analysis of model and functions. In: *Second International Conference on Metadata and Semantics Research*, 2007-10-11 - 2007-10-12. Available: `https://strathprints.strath.ac.uk/3435/`

8. **Published work 8 (PW8)** McCulloch, E. and Macgregor, G. (2008) Analysis of equivalence mapping for terminology services. *Journal of Information Science*, 34 (1). pp. 70-92. Available: `https://doi.org/10.1177/0165551507079130`

9. **Published work 9 (PW9)** Macgregor, George (2012) *Principles in Patterns (PiP) : User Acceptance Testing of Course and Class Approval Online Pilot (C-CAP)*. [Report]. University of Strathclyde, Glasgow. Available: `https://strathprints.strath.ac.uk/46510/`

10. **Published work 10 (PW10)** Macgregor, George (2019) Improving the discoverability and web impact of open repositories: techniques and evaluation. *Code4Lib Journal* (43). Available: `https://journal.code4lib.org/articles/14180`

11. **Published work 11 (PW11)** Macgregor, George (2020) Enhancing content discovery of open repositories: an analytics-based evaluation of repository optimizations. *Publications*, 8(1), 8. Available: `https://doi.org/10.3390/publications8010008`

Digital directions

Collection-level descriptions: metadata of the future?

George Macgregor

The author

George Macgregor is a Researcher, at the Centre for Digital Library Research, University of Strathclyde, Glasgow, UK.

Keywords

Digital libraries, Trends, Cataloguing,
Collection management

Abstract

The potential for digital library growth has recently drawn into question the ability of users to navigate large distributed and heterogeneous collections. This column attempts to summarise some of the potential benefits to be derived through the implementation of collection-level descriptions for both user resource discovery and institutional collection management. In particular, the concept of "functional granularity" is introduced and some related issues are briefly explored.

Electronic access

The Emerald Research Register for this journal is available at
http://www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
http://www.emeraldinsight.com/0024-2535.htm

Library Review
Volume 52 · Number 6 · 2003 · pp. 247-250
© MCB UP Limited · ISSN 0024-2535
DOI 10.1108/00242530310482015

Introduction

It is now almost farcical to think that the accommodation of the newer formats such as films or sound recordings were considered as a "revolution" in the 1970s and 1980s. Since then, of course, libraries and information services have undergone developments of near seismic proportions as they attempt to tame what Manoff (2000, p. 861) refers to as "the information monsters". The proliferation of electronic information, mainly via the Web, has forced information professionals to swallow that bitter pill: "access vs. ownership". Many may have swallowed it, but few have truly digested it. We still find ourselves trying to exert the same degree of control over electronic information resources as that of print based resources. In the past decade we have discovered that libraries increasingly provide access to highly volatile information, information with an apparent lack of fixity, and information that is often bereft of permanence. The transient nature of this information is such that users are directed to a plethora of information (e-journals, Web sites, related collections, and suchlike) held outwith the traditional confines of the "collection".

Whilst the arguments continue to rage as to whether this constitutes a desirable model of information provision, what is certain is that such arrangements thwart efforts to assimilate them into traditional bibliographic forms. This perception of library collections, particularly in the realm of digital libraries, has been changing. The use of collection-level descriptions has become an increasingly topical and relevant issue in recent years, especially since digital libraries represent a more heterogeneous manifestation than traditional libraries (Hill *et al.*, 1999, p. 1169). The emergence of digital and hybrid libraries, maturing library catalogue systems, the exponential gathering of digital resources into "collections" and the further aggregation of these collections has renewed interest in the use of collection-level description as a means of enhancing resource discovery and collection management. The purpose of this issue's column is therefore to illuminate the emerging potential of collection-level description (CLD) and to perhaps raise some issues worthy of further thought.

CLD: the scenario

CLDs are nothing particularly new. Archives, for instance, have long been using such resource discovery tools. Items within archival fonds can only be understood and appreciated within the context of those other items belonging to the fonds, and the descriptive practice employed by archivists reflects this approach. Yet, in the library and information science domain the collection has not traditionally been at the forefront of resource discovery. Obviously special collections and other significant collections have always existed, but such a view of libraries and information services has not traditionally underpinned the delivery of services in the same way that it has in the world of archives and museums. Emphasis has been on item-level activities such as cataloguing and circulation, while collection-level activities have been implicit in the local service environment. Nevertheless, the rampant march of the digital libraries has perpetuated the rise of information repositories of unforeseen magnitude and of tremendous diversity, often spanning a variety of domains. More importantly, the potential for digital library growth far exceeds the humble parameters established by the print based library. It is thus appropriate to aid user navigation of such "information landscapes" in order that information contained therein is not rendered wholly redundant by its apparent abundance.

The brief nature of this column prohibits any detailed explanation of CLDs; references are provided for that purpose. However, for our interim purposes we can consider a CLD to be a structured, open, standardised and machine-readable form of metadata providing a high-level description of an aggregation of individual items. Such descriptions disclose information about their existence, characteristics and availability, and employ the use of implicit item-level metadata and, more particularly, contextualise that aggregation of item-level descriptions. CLDs are clearly desirable since they can enable the discovery of collections of interest, particularly prior to item-level discovery or data mining. Providing us with an eloquent analogy, Heaney (2000, p. 3) states that the:

> information landscape can be seen as a contour map in which there are mountains, hillocks, valleys, plains and plateaux . . . The scholar surveying this landscape is looking for the high points. A high point represents an area where the potential for gleaning desired information by visiting that spot is greater than in other areas.

We are therefore able to harness the potential of CLDs to provide an overview of groups of items, perhaps even uncatalogued items or those items where item-level details are inappropriate. Such an approach is conducive to the "high-level" navigation of large and often distributed or heterogeneous resource bases. A scholar, for instance, may wish to utilise CLDs to discover the existence of collections spanning numerous domains but with a common characteristic such as subject or collector, and then to subsequently rationalise and direct their item-level queries on the basis of the characteristics intrinsic to that collection. In essence, we can deliver improved distributed networked services for users with the uptake of such metadata, particularly when clear opportunities arise to augment interoperability – for example, the implementation of agreed schemas such as the Research Support Libraries Programme (RSLP) Schema (Powell *et al.*, 2000).

"Functional granularity"

Leaving aside the contentious issue of what actually constitutes a "collection", an institution that agrees upon the particular aggregations that form its collections will invariably discover that these collections are related on a variety of levels. Thus, relationships could be applied to collections of varying sizes and granularity so that, for instance, a "collection" may contain numerous "sub-collections", and vice versa ("super-collections"). The use of granularity is of obvious importance in the context of CLD resource discovery, and Geisler *et al.* (2002, p. 216) have already commented that the relational attributes will be essential, not only for discovering resources within single repositories, but also across libraries of all types, and across different domains:

> By explicitly representing not only a wealth of collections, but also the relationships among them, regardless of their physical location, a collection level metadata schema should greatly improve the navigability of the [digital library].

Such conceptualisations of resource navigation have already been instantiated by CLD projects such as SCONE (2002). In

SCONE rich forms of CLD are capable of being "drilled down" from the highest level of granularity through related sub-collections (many of them distributed) until the desired degree of specificity is reached. The ability to exploit descriptions created by institutions for practical and functional reasons means that the user is capable of surveying the landscape for the elusive "high points". Not only is this approach functional for the user, but it is also functional for the institution. A CLD, based on a recognised schema or standard, can provide a simple, less labour-intensive and standardised means of disclosing an institution's curatorial responsibilities. Disclosing such responsibilities can underpin collection management duties and related initiatives by providing a convenient tool for coordinating collection development, bibliographic access, storage, and preservation, and by enabling informed strategic planning at institutional, cross-institutional, regional, sectoral and national levels. The CLD then assumes a brand of re-usable or recyclable metadata.

Functional for whom?

The concept of functional granularity is unquestionably an intriguing proposition. Heaney's paper, "An analytical model of collections and their catalogues", which has informed the work of UKOLN's collection description focus (UKOLN, 2003) and CDLR projects like SCONE and CC-interop (CC-interop, 2003), suggests that a functional granularity approach should be adopted by an institution in the description of its collections in order to:

> . . . make explicit those elements of the collection descriptions which the institution deems to be useful or necessary for the purposes of resource discovery or collection management (i.e. should adopt a "functional granularity" approach) (Heaney, 2000, p .5).

Clarifying this supposition, Dunsire (2002) states that, "if intellectual or administrative effort has gone into the definition of the collection, then it is probably worth recording". The concept of functional granularity is therefore entirely flexible and relies on the judgements of the administrators of that collection (or those closely associated with the collection) to make informed decisions over what they consider to be a relevant and useful aggregation of items. In doing so they provide flexible tools for collection management, breaking a collection down into manageable sub-collections. More importantly they create levels of granularity that are enlightened by the information professionals' unique knowledge of these collections. These are capable of supporting navigation of the chaotic information environment to which we wish to restore order, as well as providing an efficient filtering mechanism. More succinctly:

> If records are created for both a collection and its significant "sub-collections", then it is possible to choose between presenting only the "super-collection" record (while filtering out the more detailed "sub-collection" records), or presenting the more detailed "hierarchical" view (Johnston, 2002).

The flexibility and functionality of this concept is exemplified yet further when one realises that it is possible to be embraced not just by collection administrators for users (and themselves), but also by users for users. SCONE has collections defined by special groups of users, such as the Scottish Working Group on Official Publications. Such user groups have created their own CLDs expressly for enhancing their own activities. As a collection description service, SCONE has similarly created specific collections to improve the "functions" of general retrieval and data cascade/inheritance.

Yet, there remains the question of whether locally dictated choices are conducive to a globally accessible information infrastructure. This is, after all, the age of "think globally, act locally". Employing the use of functionally granular techniques, especially for resource discovery across distributed networked services, is undoubtedly useful and it does provide a trajectory worth pursuing in digital library research. The success of SCONE bears testament to this. However, a project like SCONE remains within the confines of a distinct geographical and networked area, where the attributes accorded to collections are derived from similar socio-political and cultural perspectives, not to mention similar information science perspectives. Users from outwith these areas may not be so informed by such peculiarities, and nor should they be expected to be. This is particularly true since they are likely to be driven to interrogating the said repository as a result of Belkin's now legendary "Anomalous State of Knowledge" (ASK) conundrum[1] (Belkin *et al.*, 1982). Of course, we can always profess that local expertise informs global expertise, but such an approach smacks of arrogance and few presumptions should ever be made over who your clientele are.

Concluding remarks

It is clear that functional granularity, and the use of granularity generally, is an area of digital library research that should be pursued further and one that should find further applications, especially in conjunction with emerging CLD schemas being championed in the UK and, to a lesser extent, the USA. The real question, however, remains as to whether such an approach to resource discovery lends itself to applications outwith the locality from whence it originated. If not, how can we best tinker with functional granularity so as to maximise its relevance to those disparate communities that do not share common views on what is functional and what is not? What is indisputable is that there is an evident paradigm shift afoot: item level description to collection level description. Digital libraries, in their various permutations, have ushered in an era whereby the importance of item level description is diminishing. It will always be significant (can you imagine a world without it?), but the gargantuan size of digital libraries and their potential for growth have emphasised its limitations and demonstrated the untenable and unwieldy nature of item level description for searching large distributed and heterogeneous collections. So is it fair to say that item level description has had its day? Not quite. It's just not the answer to everything anymore.

Note

[1] Belkin famously remarked that it is "unrealistic to ask the user of an IR system to say exactly what it is that she/he needs to know, since it is just the lack of that knowledge which has brought her/him to the system in the first place" (Belkin *et al.*, 1982, p. 66).

References

Belkin, N.J., Oddy, R.N. and Brooks, H.M. (1982), "Ask for information retrieval: Part 1. Background and theory, *Journal of Documentation*, Vol. 38 No. 2, June, pp. 61-71.

CC-interop (2003), available at: http://ccinterop.cdlr.strath.ac.uk/ (accessed 12 March 2003).

Dunsire, G. (2002), "Guide to the SCONE database", Glasgow, CDLR, available at: http://scone.strath.ac.uk/service/Guide/gContents.cfm (accessed 12 March 2003).

Geisler, G., Giersch, S., McArthur, D. and McCelland, M. (2002), "Creating virtual collections in digital libraries: benefits and implementation issues", *Joint Conference on Digital Libraries 2002*, Portland, OR, pp. 210-218, available at: www.ukoln.ac.uk/metadata/rslp/model/amcc-v31.pdf (accessed 12 March 2003).

Heaney, M. (2000), *An Analytical Model of Collections and their Catalogues*, UKOLN, Bath, available at: www.ukoln.ac.uk/metadata/rslp/model/amcc-v31.pdf (accessed 12 March 2003).

Hill, L.L., Janée, G., Dolin, R., Frew, J. and Larsgaard, M. (1999), "Collection metadata solutions for digital library applications", *Journal of the American Society for Information Science*, Vol. 50 No. 13, pp. 1169-81.

Johnston, P. (2002), *Creating Reusable Collection Level Descriptions – Collection Description Focus, Guidance Paper 1*, UKOLN, Bath, available at: www.ukoln.ac.uk/cd-focus/guides/gp1/ (accessed 12 March 2003).

Manoff, M. (2000), "Hybridity, mutability, multiplicity: theorizing electronic library collections", *Library Trends*, Vol. 49 No. 10, Summer, pp. 857-76.

Powell, A., Heaney, M. and Dempsey, L. (2000), "RSLP collection description", *D-Lib Magazine*, Vol. 6 No. 9, available at: www.dlib.org/dlib/september00/powell/09powell.html (accessed 12 March 2003).

SCONE (2002), available at: http://scone.strath.ac.uk/ (accessed 12 March 2003).

UKOLN (2003), available at: www.ukoln.ac.uk/cd-focus/ (accessed 12 March 2003).

The Emerald Research Register for this journal is available at
www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0033-0337.htm

Towards improved performance and interoperability in distributed and physical union catalogues

George Macgregor
Department of Computer and Information Sciences, Centre for Digital Library Research, University of Strathclyde, Glasgow, UK, and
Fraser Nicolaides
University of Greenwich, London, UK

Received November 2004
Revised January 2005
Accepted January 2005

Abstract

Purpose – Detail research undertaken to determine the key differences in the performance of certain centralised (physical) and distributed (virtual) bibliographic catalogue services, and to suggest strategies for improving interoperability and performance in, and between, physical and virtual models.

Design/methodology/approach – Methodically defined searches of a centralised catalogue service and selected distributed catalogues were conducted using the Z39.50 information retrieval protocol, allowing search types to be semantically defined. The methodology also entailed the use of two workshops comprising systems librarians and cataloguers to inform suggested strategies for improving performance and interoperability within both environments.

Findings – Technical interoperability was permitted easily between centralised and distributed models, however, the various individual configurations permitted only limited semantic interoperability. Significant prescription in cataloguing and indexing guidelines, greater participation in the program for collaborative cataloguing, consideration of future functional requirements for bibliographic records migration, and greater disclosure to end users are some of the suggested strategies to improve performance and semantic interoperability.

Practical implications – This paper not only informs the library and information science research community and union catalogue administrators, but also has numerous practical implications for those establishing distributed systems based on Z39.50 and search/retrieve web services as well as those establishing centralised systems.

Originality/value – The paper moves the discussion of Z39.50-based systems away from anecdotal evidence and provides recommendations based on testing, and is intimately informed by the UK cataloguing and systems librarian community.

Keywords Online catalogues, Cataloguing, Information retrieval, Open systems, Worldwide web, Information searches

Paper type Research paper

1. Introduction

Union catalogues are by no means a new phenomenon. As Cannell and Guy (2001) note, the emergence of library co-operatives in the 1970s stimulated the evolution of

The authors extend their gratitude to Dr Shirley Cousins and Ashley Sanders for providing access to COPAC records in MARC format and assisting greatly in the interpretation of search results. Also they extend their appreciation to those cataloguers and systems librarians that participated in the cataloguing/interoperability workshops.

Program: electronic library and information systems
Vol. 39 No. 3, 2005
pp. 227-247
© Emerald Group Publishing Limited
0033-0337
DOI 10.1108/00330330510610573

co-operative cataloguing systems and resource sharing, which in turn began the evolution of what would currently be termed a centralised "union catalogue". Perhaps the most obvious example came from the USA where the Ohio College Libraries Centre (OCLC) has successfully developed automated systems designed to expedite the goals of the co-operative since the late 1960s. Such initiatives initially tended to produce catalogue records for participating libraries either in computer output microfilm (COM) or on catalogue cards to be consolidated with the other catalogue cards. Enthused by developments across the Atlantic, a series of similar initiatives, some more successful than others, were developed in the UK, including the Birmingham Libraries Co-operative Mechanisation Project (BLCMP), South West Academic Libraries Co-operative Automation Project (SWALCAP), and the Scottish Libraries Co-operative Automation Project (SCOLCAP) (Tedd, 1994). Of course, by 1979 the shared cataloguing system used by OCLC had attracted libraries, not just from across the USA, but from across the world, and by 1981 OCLC decided to change their name to the OCLC Online Computer Library Centre (Jordan, 2003). Today OCLC has become a leading international library organisation and presides over the largest union catalogue in the world: WorldCat (OCLC, 2004).

As user requirements and expectations have grown in tandem with massive web development, making these union catalogues accessible to academics and students has long been a keen area of interest for library and information science (LIS) practitioners and researchers around the globe. The main focus for the UK activity was the Consortium of University Research Libraries (CURL). CURL recognised early on that its centralised shared cataloguing database (established in 1987) constituted a valuable resource for the academic community and that access to such a resource should be widened (Cousins, 1997). In 1995, CURL established the CURL OPAC (COPAC), providing web access to the consolidated records (now about 30 million) of the 26 library members.

Such developments have typically made union catalogues more end-user orientated and since the mid-1990s union catalogues have tended to assume one of two manifestations:

(1) the centralised (or physical) model – a centralised approach whereby bibliographic records contributed from a number of participating institutions are incorporated into a single database; and

(2) the distributed (or virtual) model – where the same service is provided via a distributed model, most commonly utilising the Z39.50 information retrieval protocol (Z39.50, 2004).

Indeed, the increasing pervasiveness of Z39.50 "broadcast searching" has thus allowed participating institutions to remain "distinct" and to avoid the maintenance costs typically associated with administering a large centralised system (Gatenby, 2002).

As with most technical service models, each has numerous advantages and disadvantages. Some of these have been widely documented in the literature for some time (Cousins, 1999; Nicholson, 2000; Stubley *et al.*, 2001; Gatenby, 2002; Friesen, 2002; Taylor, 2003) or examined (Moen, 2001a, b; Moen and Murray, 2003), whilst others have undergone thorough analysis under the auspices of the UK CC-interop project (Nicolaides, 2003a, b; Gilby and Sanders, 2003; Gilby *et al.*, 2004; Dunsire and Macgregor, 2004). Nevertheless, with increasing library Z39.50 compliance, the creation of larger heterogeneous distributed union catalogues becomes ever more likely, and issues pertaining to the relative performance of each model have consequently been drawn into sharp focus. The need for improved performance is now also essential to secure the confidence of end users, some of whom believe union catalogue services to be unreliable or irrelevant (Booth and Hartley, 2004). This focus has sharpened yet further as it becomes clear that those technologies expected to eventually supersede Z39.50 entirely, web services technologies (WST), still harbour various protocol limitations and often suffer from reliability, security and transaction time difficulties (Yu and Chen, 2003).

Z39.50's recent offshoot initiative, "Z39.50 International: Next Generation" (ZING), has been fronting several exciting developments, particularly search/retrieve web service (SRW) and search/retrieve uniform resource indicator (SRU) (ZING, 2004). Both SRW and SRU represent an attempt to amalgamate the powerful capabilities of Z39.50 by implementing them in parallel with updated web-friendly protocols and technologies, such as HTTP (hypertext transfer protocol) with SOAP (simple object access protocol), a protocol for XML (extensible markup language) messaging, and by utilising WSDL (web services definition language) to define the Z39.50 messages. Whilst ZING promises greater functionality, and although developers are beginning to incorporate SRW/SRU facilities within standard Z39.50 software (Index Data, 2004), developments remain tentative and some would argue that it will be some time before it becomes as widely accepted as Z39.50. Many libraries have recently invested significant resources to become Z39.50 compliant and it is only now that Z39.50 compliance has truly reached the "critical mass" to which the UK's Electronic Libraries (eLib) programme originally aspired in 1998 (Macgregor, 2005).

Still, although Z39.50 has a long history, it is far from outmoded. As Taylor (2003) notes and predicts, Z39.50 may have peculiar problems but it remains capable of adapting to new environments and will experience wider deployment within the LIS sector and beyond for many years. Such predictions are certainly manifest in wider LIS deployment. As in many information-rich countries, the UK is experiencing an increasing deployment of Z39.50 applications. While this is most marked in academic and research libraries, it is extending also to further education (FE) colleges, public libraries and lifelong learning institutions. For example, the Scottish Portals for Education, Information and Research project (SPEIR) has spearheaded the wider roll out of Z39.50 across these sectors in order to facilitate the creation of a Scottish Distributed Digital Library (SDDL) (Dunsire and Macgregor, 2003; Nicholson *et al.*, 2004). The emergence of large scale initiatives in museums provides yet further evidence of Z39.50's deployment potential, and reaffirms the possibilities that wait in creating truly heterogeneous distributed union catalogues (Caplan and Haas, 2004). Quite simply Z39.50 is, and will remain for the immediate future, the "eminent enabling technology for distributed, parallel access to information sources" (Hammer and Andresen, 2002).

2. Research purpose and objective

Given this premise, and the growing expectations of user groups, it is essential to improve the performance of both physical and virtual union catalogue models. Moreover, improving performance of each is essential to improving interoperability between both models. Such reasoning has assumed greater relevance via the CC-interop project and, in particular, the work documented by Gilby and Sanders

(2003), whereby it is possible to treat an entire virtual union catalogue as a single Z39.50 target (or Z-target) during traditional Z39.50 broadcast searching. This presupposes the future pre-eminence of ZING technology since SRW/SRU will not provide relief in respect to semantic interoperability and those variations in cataloguing and indexing practices that continue to impair optimal performance of virtual union catalogues. Therefore, the overarching purpose of this paper (and study) is twofold:

- to identify key differences in the performance of certain centralised (physical) and distributed (virtual) bibliographic catalogue services; and
- to suggest strategies for improving interoperability and performance in, and between, physical and virtual models.

The distinct nature of the research objectives will be reflected in the format of the paper, which will essentially follow two parts. Before discussing the research questions, however, it is worthwhile contextualising our study within the remit of the CC-interop project, under the auspices of which much of the said research was undertaken.

3. Background: the CC-interop project

In 1998, the UK's Joint Information Systems Committee (JISC) funded the third phase of the eLib programme, entailing the creation of several virtual union catalogue services (or "clumps" as they became known). Although the creation of widely used and successful services was an objective, the ultimate aim of the clumps was to "kick start critical mass" in the use of Z39.50 and to generate model technical architectures and agreements to precipitate the development of new clumps in their assorted incarnations, perhaps even nationally (Whitelaw and Joy, 2001, p. 2).

By 2000 four clumps had been created:

(1) M25 Link – for libraries within the M25 motorway around London;
(2) CAIRNS – Co-operative Academic Information Retrieval Network for Scotland;
(3) RIDING – libraries in Yorkshire and Humberside; and
(4) Music Libraries Online (MLO).

All these projects successfully established fully functioning clumps, each with common and unique features. Most were regionally defined and were built upon existing library co-operatives. For our purposes, however, the two most significant clumps were M25 Link and CAIRNS:

(1) M25 Link had six partners drawn from the M25 Consortium of Academic Libraries based in the London area (www.m25lib.ac.uk). The eventual distributed catalogue, now comprising 36 institutional Z-servers, forms part of the InforM25 service. It is maintained for the consortium by the M25 systems team.
(2) CAIRNS (http://cairns.lib.strath.ac.uk/) included members of the Scottish Confederation of University and Research Libraries (SCURL) and is now developed and maintained by the Centre for Digital Library Research (CDLR) at the University of Strathclyde. CAIRNS comprises 33 institutional Z-servers, including numerous non-higher education (HE) Z-servers.

To build on the results and findings of eLib phase three, JISC provided a two-year funding grant to the COPAC/Clumps Continuing Technical Cooperation Project (CC-interop: http://ccinterop.cdlr.strath.ac.uk/), which aimed to "bring together, in a virtual modus operandi, distributed catalogues to facilitate richer search and retrieval possibilities for users" (Gilby and Dunsire, 2004, p. 4). Beginning in mid-2002, CC-interop was a collaborative project involving: the M25 systems team, CDLR, Manchester Information and Associated Services (MIMAS), RIDING, and latterly the Centre for Research in Library and Information Management (CERLIM). The inclusion of the COPAC service (http://copac.ac.uk/) at MIMAS epitomised the co-operative nature of the project and emphasised the dialectic nature of the research being undertaken.

The project comprised three work packages, each investigating a plethora of issues, including:

- inter-linking between very large physical union catalogues (i.e. COPAC) and large virtual union catalogues (i.e. InforM25);
- the ability to "clump the clumps" thus producing a "hyper-clump";
- thorough research of collection-level description requirements for such environments;
- improving interoperability in distributed and physical environments; and
- investigating user requirements and behaviour for union catalogues.

For a greater discussion of the project outcomes and findings refer to Gilby and Dunsire (2004).

4. Methodology for first research objective

Our first objective was to identify key differences in the performance of certain centralised (physical) and distributed (virtual) bibliographic catalogue services. COPAC was used as the physical union catalogue for study, and the distributed services selected for testing were those CURL institutions that were also members of InforM25 as seen in Table I.

The focus of the performance evaluation was to determine why any given query might elicit a different result set from each of the two types of system. As such, consideration was given to several aspects of the respective systems: from their interpretation of the structured format in which the queries were submitted to the policies and practices affecting the indexes against which the query was executed. This necessary approach therefore limited the use of quantitative techniques, and instead

Distributed services (distributed Z-servers)	Abbreviation	Library system
Imperial College of Science, Technology and Medicine	Imperial	Unicorn
London School of Economics	LSE	Unicorn
School of Advanced Study	SAS	Innopac
University College London	UCL	Aleph
University of London Library	ULL	Innopac
Wellcome Library for the History and Understanding of Medicine	Wellcome	Innopac

Table I.
Table detailing those distributed systems used for the experiment that are members of both InforM25 and COPAC

a methodical qualitative approach was adopted from which broader conclusions could be drawn.

4.1 Searching and search types

Searches of COPAC and each of the selected distributed catalogues were conducted using the Z39.50 information retrieval protocol and connections to the relevant Z-servers were made using a Yaz Z-client (Index Data, 2004). This allowed search types to be semantically defined in ways additional to those publicly available through the COPAC and InforM25 search interfaces. Searches were constructed using the Bib-1 attribute set, a standard used in Z39.50 to define how search terms are to be treated by the local catalogue (Z39.50, 2003). The search types used for this study included:

- author;
- author/title;
- key title (serial title);
- subject; and
- any (keyword).

Searches of these types were most inclined, we hypothesised, to elicit a different result set from each system since such search types are subject to greater variation in index scope and content, particularly author and subject searches.

No attempt was made to assess the precision of the result sets. This concept is wholly dependent upon the definition of relevance and, as such, was beyond the scope of our research. Instead, in examining the relative performance of COPAC and one or more of the distributed systems, we have sought to account for any differences in result set content (in short, why certain records might be present or absent). It is also worth noting that in examining result sets from the centralised and distributed systems, we have been concerned to identify comparable bibliographic records. Any assessment of the presence or quality of any associated holdings and location data (number of copies, enumeration and chronology, etc.) has therefore been omitted.

Although certain significant differences were observed with respect to the capabilities of the examined services, the superiority of either the physical or distributed model is, and will not be, inferred. Rather, the primary concern was to consider the opportunities for effecting greater interoperability between all components, particularly via any potential operational scenarios, such as within the UK National Union Catalogue (UKNUC). Moreover, it was not the purpose of this research to describe all of the potential or current functionality of COPAC or the several distributed Z-servers.

4.2 Caveats

Not all the distributed systems listed in Table I were included in each test. In order of importance, there were three reasons for this:

(1) not all of the distributed systems were enabled to use precisely the sets of Bib-1 attributes supported by COPAC;

(2) in some circumstances, the various institutional implementations of a particular system type performed in a consistent way and so, testing more than one implementation was consequently not always necessary; and

(3) duplicate searches were not performed when it was felt that the issue or aspect of performance had already been adequately illustrated.

5. Findings

The study revealed a variety of differences in system performance between the physical and the virtual models. Based on the characteristics of these observations, and to provide greater focus for the results discussion, the authors have deemed these differences to fall into the following three broad categories.

(1) "Consolidated and Individuated Indexes" (issues pertaining to those indexes in the tested physical and union catalogues, respectively).

(2) "Data Currency and Comprehensiveness" (issues pertaining to the currency and comprehensiveness of the records retrieved).

(3) "Support and Treatment of Bib-1 Attributes" (the manner in which the tested systems interpreted the search queries and any issues therein).

5.1 Consolidated and individuated indexes

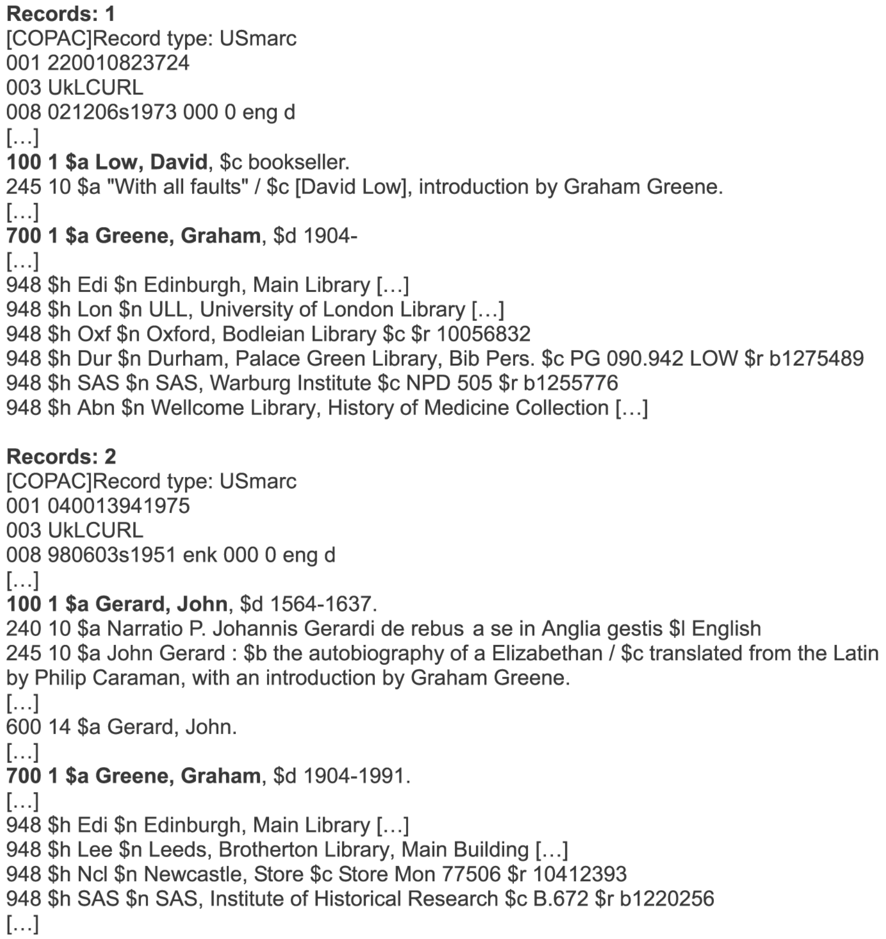
As an example of the physical model, the COPAC system exploits a feature peculiar to union catalogues in that any bibliographic entity is able to derive index entries from records submitted by several institutions. As would be expected from any centralised system, testing demonstrated that entities have been catalogued to various degrees of comprehensiveness. Bearing in mind that any (consolidated) COPAC record may be cumulatively enriched by successive contributing institutions, the potential of the search process to retrieve relevant records proportionately improves. Of course, the corollary dictates that any mis-catalogued entity may generate incorrect index entries, and thus, reduce the precision of affected result sets. Although one such instance was encountered during testing, the authors deem this to be a comparatively minor problem given the wider benefits.

By way of example, a right-truncated author query for "greene, g" was submitted to COPAC and to the distributed SAS Z-server. This generated three hits from COPAC and one on the SAS Z-server, which was duplicated in the COPAC result set. The two additional records retrieved from COPAC are shown in Figure 1.

Both records are present on the SAS catalogue, being found using a similarly structured "Author" search for "low, david" and "gerard, john". Their retrieval from COPAC is a function of the additional indexing of subfield 700$a, Author Added Entry. In both cases, this contains the term "Greene, Graham", which matches the requirements of the query. As can be observed from Figure 2, the equivalent records from SAS do not contain this supplemental field.

The same occurrence was observed in respect of the testing against COPAC and Wellcome, and COPAC and the LSE and Imperial, respectively, using an author query without truncation.

The different performance of the two systems may therefore be attributed to the different composition of the records from which the indexes are derived. For COPAC, the index entries for each of these items have been derived from the relevant records of the multiple contributing institutions. This is shown in Figure 1 by the multiple instances of the 948 institutional-holdings field. In each case, at least one of these contributions contained the Author Added Entry subfield.



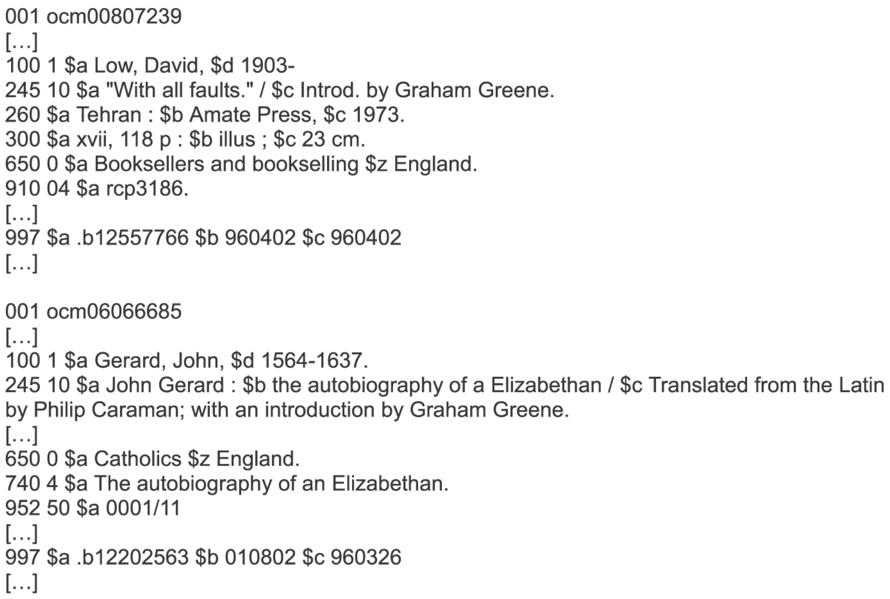
```
Records: 1
[COPAC]Record type: USmarc
001 220010823724
003 UkLCURL
008 021206s1973 000 0 eng d
[…]
100 1 $a Low, David, $c bookseller.
245 10 $a "With all faults" / $c [David Low], introduction by Graham Greene.
[…]
700 1 $a Greene, Graham, $d 1904-
[…]
948 $h Edi $n Edinburgh, Main Library […]
948 $h Lon $n ULL, University of London Library […]
948 $h Oxf $n Oxford, Bodleian Library $c $r 10056832
948 $h Dur $n Durham, Palace Green Library, Bib Pers. $c PG 090.942 LOW $r b1275489
948 $h SAS $n SAS, Warburg Institute $c NPD 505 $r b1255776
948 $h Abn $n Wellcome Library, History of Medicine Collection […]

Records: 2
[COPAC]Record type: USmarc
001 040013941975
003 UkLCURL
008 980603s1951 enk 000 0 eng d
[…]
100 1 $a Gerard, John, $d 1564-1637.
240 10 $a Narratio P. Johannis Gerardi de rebus a se in Anglia gestis $l English
245 10 $a John Gerard : $b the autobiography of a Elizabethan / $c translated from the Latin
by Philip Caraman, with an introduction by Graham Greene.
[…]
600 14 $a Gerard, John.
[…]
700 1 $a Greene, Graham, $d 1904-1991.
[…]
948 $h Edi $n Edinburgh, Main Library […]
948 $h Lee $n Leeds, Brotherton Library, Main Building […]
948 $h Ncl $n Newcastle, Store $c Store Mon 77506 $r 10412393
948 $h SAS $n SAS, Institute of Historical Research $c B.672 $r b1220256
[…]
```

Figure 1.
Two additional records retrieved from COPAC using right-truncated author search

Some variations in the indexing policies and practices of the reviewed institutions and services were also identified. Thus, for any bibliographic record held simultaneously on COPAC and the contributing institution's database, index entries may have been derived from differing sets of (sub)fields. This problem is further exacerbated by possible variations in the mapping from the indexes to the Bib-1 use (access point) attributes. For example, an institution or service may have created several "Author" indexes, each of which is derived from a differing set of relevant fields and each of which is mapped to a different use attribute (author (1003); name (1002); personal name (1); author – personal name (1004); etc.) What constitutes any given "Author" index could therefore vary considerably between databases of essentially the same records.

5.2 Data currency and comprehensiveness

COPAC maintain an "update" page (http://copac.ac.uk/about/updated/) to keep users informed as to the currency of the database. Although COPAC receives updates from



```
001 ocm00807239
[…]
100 1 $a Low, David, $d 1903-
245 10 $a "With all faults." / $c Introd. by Graham Greene.
260 $a Tehran : $b Amate Press, $c 1973.
300 $a xvii, 118 p : $b illus ; $c 23 cm.
650 0 $a Booksellers and bookselling $z England.
910 04 $a rcp3186.
[…]
997 $a .b12557766 $b 960402 $c 960402
[…]

001 ocm06066685
[…]
100 1 $a Gerard, John, $d 1564-1637.
245 10 $a John Gerard : $b the autobiography of a Elizabethan / $c Translated from the Latin
by Philip Caraman; with an introduction by Graham Greene.
[…]
650 0 $a Catholics $z England.
740 4 $a The autobiography of an Elizabethan.
952 50 $a 0001/11
[…]
997 $a .b12202563 $b 010802 $c 960326
[…]
```

Figure 2.
Equivalent records from the SAS Z-server

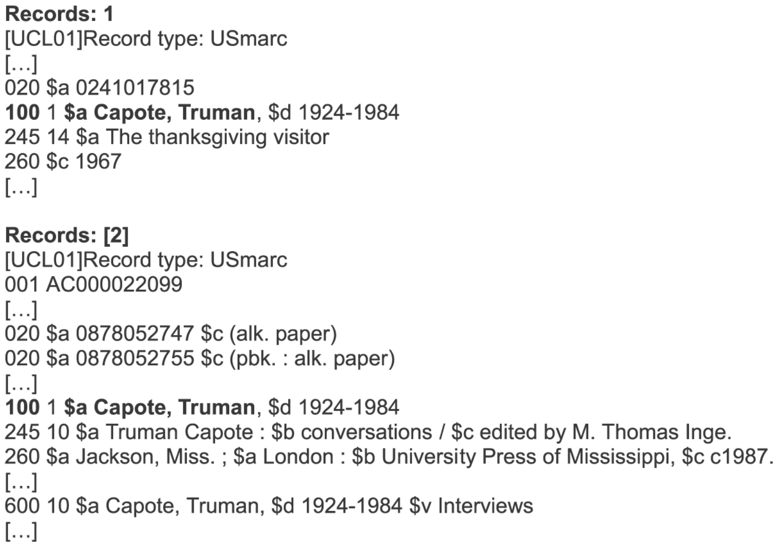
the British Library weekly, updates from other contributing libraries may not be as frequent. As such, the result sets obtained from the distributed systems were, by and large, found to be more up-to-date than the equivalent sets from COPAC. These instances largely concerned a single institution (UCL), which at the time of testing had last submitted data to COPAC in August 1999. The particular problem encountered in the tests was that records on the institutional database were absent from COPAC. (The theoretical corollary is that records pertaining to items withdrawn from stock may still appear on the union database.) This issue concerns the relative frequency with which records are updated on local and third-party databases, such as COPAC.

One such example was observed when testing the search responses from UCL and COPAC. The author query (without truncation) “capote, truman”, returned six records from COPAC and eight from UCL. The two additional records from UCL can be seen in Figure 3.

ISBN searches on COPAC revealed that both items were in fact recorded on the COPAC database, but neither had a current holdings statement for UCL (in the 948 field). This discrepancy evidently occurred due to the temporary obsolescence of COPAC’s UCL data.

Such discrepancies were also manifested in the “policy determined” omission of records relating to certain classes of online resource. Indeed, at the time of testing, both LSE and Imperial had elected not to submit records to COPAC for those bibliographic records describing (and providing links to) electronic resources, such as licensed full-text services or equivalent resources. Conversely, these records were accessible to any third-party Z-client through the institutional Z-servers; though, for consistent policy application and service delivery, it is arguable that they should not be.

Whilst it is unnecessary to over-emphasise the importance of what essentially are administrative processes, such discrepancies as outlined above will tend to undermine



```
Records: 1
[UCL01]Record type: USmarc
[…]
020 $a 0241017815
100 1 $a Capote, Truman, $d 1924-1984
245 14 $a The thanksgiving visitor
260 $c 1967
[…]

Records: [2]
[UCL01]Record type: USmarc
001 AC000022099
[…]
020 $a 0878052747 $c (alk. paper)
020 $a 0878052755 $c (pbk. : alk. paper)
[…]
100 1 $a Capote, Truman, $d 1924-1984
245 10 $a Truman Capote : $b conversations / $c edited by M. Thomas Inge.
260 $a Jackson, Miss. ; $a London : $b University Press of Mississippi, $c c1987.
[…]
600 10 $a Capote, Truman, $d 1924-1984 $v Interviews
[…]
```

Figure 3.
The two additional records from UCL, not found in COPAC

coherent and consistent results across physical and virtual union catalogues, and could potentially deliver inaccurate results within a hyper-clump environment (particularly one incorporating several third-party databases).

5.3 Support and treatment of the Bib-1 attributes

As explained previously, the Bib-1 attribute set is designed to enable the definition of all semantic structures relevant to the identification of bibliographic records. Simply, the Bib-1 attribute set provides semantic definition to the search types by deciding their precise nature. For example, a title search for "ancient american civilizations" (Figure 4) could be interpreted in several ways.

Even if it was interpreted by the system as a "title keyword" search, the issue of truncation still has to be resolved before the search can be undertaken. Such searches are defined by the Bib-1 attribute set in an attempt to resolve these issues pertaining to query interpretation. Table II outlines a basic attribute set for "title keyword" or "title

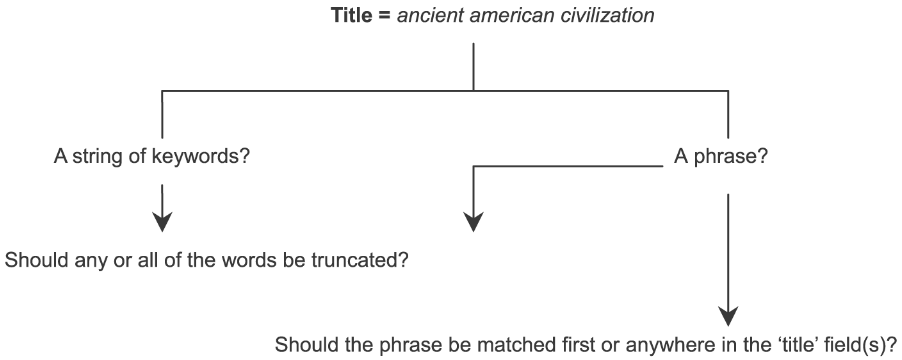

Title = ancient american civilization
A string of keywords?
A phrase?
Should any or all of the words be truncated?
Should the phrase be matched first or anywhere in the 'title' field(s)?

Figure 4.
Process of defining the semantic nature of the search query

exact match" searches, as defined by use, relation, position, structure, truncation and completeness.

To calculate the potential number of attribute combinations would itself be a mathematical challenge. Attribute combinations have consequently been subject to further specification via internationally recognised library profiles such as the Bath Profile (Bath Profile Maintenance Agency, 2004). As an adjunct to the development of commercial and public-domain Z39.50 services, the adoption of Bib-1 has, understandably, been somewhat selective, as has adherence to the related profiles. Thus, in practice, some systems do not support all six attribute types, and, more commonly, most systems support only a selection of the individual attributes and attribute combinations. In many cases, this selectivity has been determined by the limitations of catalogue indexes and local database search routines to which the attributes are mapped. Nevertheless, these tests have revealed several noteworthy aspects of Z-server support and behaviour.

The very scope of the tests was determined by the comparatively limited extent of mutual support for specific attributes and attribute sets. This was patently manifested in the variable support for the "Position" attributes, first- and anywhere-in-field, which perforce can markedly influence retrieval. In some instances, where an attribute (or attribute combination) was not supported, it was replaced with an alternative attribute (or combination) by the Z-server. Such default Z-server behaviour was exemplified with the treatment of "Truncation" by those distributed services using Unicorn systems (Imperial and LSE) and can impact significantly on the consistency of the result sets obtained from differently implemented databases of the same records.

A distinctive variation on this concerns COPAC's ability to interpret the query term (rather than, or in addition to, the attribute set). Testing detailed an "Author" search that, because the term was in a normalised format was, by definition, submitted to a "Quick Author" index. This operation effectively negated one of the specified attributes, "No Truncation", a function that is actually supported by COPAC. Whilst these various default modifications are usually intended to maximise the efficiency of the search and to optimise system performance in a one-to-one (Z-client to single Z-server) relationship, it may not be entirely appropriate in operational environments where a Z-client wishes to affect some measure of semantic consistency between multiple Z-servers.

One final notable behaviour of the tested services related to the processing of the "Structure" attribute, "Phrase". As we discovered, COPAC and certain institutional

	Title search			
	Keyword		Exact match	
Attribute type	Value	Attribute	Value	Attribute
Use (1)	4	Title	4	Title
Relation (2)	3	Equal	3	Equal
Position (3)	3	Any-position-in-field	1	First-in-field
Structure (4)	2	Word	1	Phrase
Truncation (5)	100	Do not truncate	100	Do not truncate
Completeness (6)	1	Incomplete subfield	3	Complete field

Table II.
Basic Bib-1 attribute set for Z39.50 title search

Z-servers responded differently to words within queries that were defined as stopwords or those that could be interpreted as Boolean operators.

In short, the relatively few systems that we examined have been shown to support disparate varieties of search type (semantically defined using the Bib-1 attribute set) and the consequent requirement for some measure of semantic interoperability is clearly evident.

6. Implications

Although the small scale of the study does not permit us to be absolutely authoritative in determining the relative importance of each of the above issues, it is reasonable to assume that the least important concerned the time-lag with which records were updated on the centralised database. Remedial action, if indeed it were deemed to be necessary, would essentially require an organisational rather than a technical solution. The two other factors, to which equal importance should be ascribed, concern variations in, first, cataloguing and indexing practices and, second, support for Bib-1 attributes and attribute sets.

A general observation made by Heiler (1995, p. 271) continues to ring true: "Semantic agreements are often lacking when old data or procedures are used for new purposes not anticipated by their original developers". Variations in cataloguing and indexing policies and practices have long been recognised within the Z39.50 community as an impairment to semantic interoperability (Lynch, 1997; Moen, 2001b; Nicholson *et al.*, 2001; Friesen, 2002). This affects all search types, but is perhaps most pronounced for "Subject" and "Keyword" type searches. Such variations are, of course, the product of historical and local requirements and contingencies, the legitimacy of which should not be challenged. Nevertheless, such issues are all pervasive and, as Simeoni (2004) notes, are beginning to blight the performance of Federated Digital Libraries (FDLs) founded upon the open archives initiative protocol for metadata harvesting (OAI-PMH) also.

Nicholson and Shiri (2003) note that semantic interoperability constitutes the largest obstacle to providing coherent distributed digital libraries, and although McCulloch (2004) and McCulloch *et al.* (2005) note some of the exciting "terminology mapping" developments and initiatives underway to provide a technical solution to these problems, it is not unreasonable to assume that it will be many years before such solutions are capable of being readily deployed within distributed digital library architectures. As McCulloch (2004) observes, even before such solutions can take root, information providers need to champion and implement international standards where multiple terminologies are in use. Operational difficulties might arise with the possibly historical use of multiple schemes, the use of *ad hoc* institutionally specific schemes, the irregular application of schemes, and so forth.

Of course, the scope of the Bib-1 attribute set has allowed for multiple disparate implementations to be made, as demonstrated by the variations in those of COPAC and each of the distributed services under analysis. Within many profiles, semantic interoperability is addressed through the definition of a core suite of search types (constructed using specified sets of attributes). For the current and possible future application scenarios in which COPAC and the distributed systems might operate, the most relevant profile would be the Bath Profile since certain commercial vendors supplying Z-server modules to higher-educational and other institutions are committed to adoption of the profile (Nicolaides, 2003b).

The technical challenges, however, should not be underestimated. In order to support the required attribute sets, it may be necessary for a library to engage in technically demanding and financially onerous tasks, such as re-indexing their catalogue. Essentially, the technical interoperability of COPAC with other distributed systems was never in doubt. Rather, what testing documented was that the various individual configurations permit only limited semantic interoperability. Thus, any supra-national system (or "hyper-clump") that seeks to integrate or otherwise utilise such component services must address the above noted issues.

7. Strategies for improving interoperability and performance

Given some of these short- to medium-term challenges, how best can the LIS community improve interoperability and performance of, and between, physical and virtual union catalogues? The most obvious strategy is to initiate some form of co-ordination of cataloguing and indexing practices. Such concerted initiatives have hitherto been few and far between. Indeed, the only visible attempt in the literature to arrest interoperability problems caused by variations in cataloguing and indexing practices was undertaken by CAIRNS Cataloguing and Indexing Working Group (2000). CAIRNS appended to these guidelines a variety of suggested short- and long-term strategies for alleviating interoperability problems, some of which met with at least nominal success (Nicholson *et al.*, 2001).

Yet, a more general lack of activity is unsurprising since semantic interoperability is inextricably tied to "communities of practice" (Friesen, 2002). Moen (2001b) clarifies this supposition by defining Networked Information Discovery and Retrieval (NIDR) (of which virtual union catalogues constitute one such incarnation) as falling into one of three communities: focal, extended and extra. Moen's definition therefore dictates that the further a virtual union catalogue moves away from a "Focal" community (typified by minimal interoperability issues and a large degree of homogeneity), the greater the challenges and cost are to achieving true interoperability. By acknowledging the work of Gilby and Sanders (2003), we soon recognise that the potential for creating supra-national distributed catalogues and hyper-clumps will inevitably dictate that member libraries will be party to an "Extra" community where there exist numerous factors affecting interoperability. Such an assertion doubtless requires the definition of suitable strategies for improving interoperability and performance in, and between, physical and virtual, particularly with respect to coordinating cataloguing and indexing practices to maximise interoperability.

8. Methodology for second research objective

To ascertain which strategies and mechanisms would prove most effective in providing some degree of homogeneity in the UK cataloguing and indexing practices, a qualitative approach was adopted whereby the opinions and views of the UK cataloguing and systems fraternity were canvassed. The objectives of this approach were threefold.

- To identify or suggest strategies capable of addressing or alleviating variant cataloguing and indexing practices in the UK.
- To ensure such strategies or proposed recommendations were intimately informed by the UK cataloguing and systems fraternity, thus ensuring

the legitimacy and authenticity of championing such strategies in the literature and beyond.

- To ascertain whether the UK-wide initiative, based primarily on the CAIRNS experience, could be adopted and whether the UK cataloguing and systems fraternity would be receptive to such an initiative were it to be rolled out at a strategic level.

Two one-day workshops were organised in London (A) and Glasgow (B), respectively, with invitations issued on relevant UK e-mail lists. A revised and more generic version of the CAIRNS cataloguing guidelines was then distributed in advance of the workshop and participants were encouraged to review these guidelines in preparation for the event. Participants were also encouraged to bring along examples of policy and practice from their local institutions, and issues they had encountered in using union catalogues, to support, contradict, and otherwise inform the workshops.

Both workshops consisted of a number of short presentations in the morning to outline and to refresh participants of the "issues", followed by a facilitated and semi-structured group discussion in the afternoon. With permission of the participants, these discussions were tape-recorded and, together with notes taken during the session, were amalgamated to produce a report summary of the discussions. These reports were then distributed to the participants for amendment, comment and correction, and then consolidated to reflect views expressed at both workshops.

Both events were well attended and attracted representation from many HE and large research libraries, as well as representation from FE colleges. In total, 52 people attended the workshops.

9. Findings from the workshops: strategies and recommendations

A clear consensus emerged at both workshops A and B that the UK cataloguing community requires, and would welcome, the creation of guidelines that were more prescriptive than the current CAIRNS guidelines. Such prescription would assist local cataloguers in actively improving interoperability, whilst simultaneously placing a degree of leverage in the hands of cataloguers and systems administrators to encourage acknowledgement by senior library management of the consequences local policies can have on global interoperability. Nevertheless, participants agreed that the continuing globalisation of cataloguing, and the future potential for hyper-clump creation, dictated that it would be more constructive to produce a set of recommendations for a wider, more active and co-ordinated approach to improving interoperability. More specific findings, strategies or recommendations from workshop A and B participants fell into four categories:

(1) collaboration within distributed or physical union catalogues;

(2) standards;

(3) strategic developments; and

(4) end users.

9.1 Collaboration within distributed or physical union catalogues

Participants at the workshops were unanimous in their recommendation that consortia of libraries contributing to union catalogues should, in the absence of any immediate strategic guidance, develop their own prescriptive guidelines covering catalogue record scope and content, whilst accounting for both local and "global" needs. Such guidelines might include a minimum input standard for the level of cataloguing and the content of entry points or headings. As argued earlier, it is no longer sufficient for such guidelines to be developed for one mode or level of aggregation. Any one library may belong to more than one union catalogue, requiring local needs to be matched against more than one set of global needs.

By way of example, the National Library of Scotland contributes to the CAIRNS distributed union catalogue, the COPAC physical union catalogue, and the British National Bibliography. Any union catalogue may in turn be treated as a single component catalogue of a larger distributed union catalogue; so that what constitutes global in one environment constitutes local in another. Similarly, Strathclyde University Library is a member of CAIRNS, but CAIRNS itself may become a member of a hyper-clump such as a distributed UKNUC. Once again, CAIRNS would be global in the first environment, but local in the second. By continuing this theme yet further, a UKNUC could feasibly become a local component in a distributed union catalogue for the Anglophone world, and so forth. It therefore becomes clear that guidelines for improving interoperability need to be developed at national and international levels and suitable mechanisms for doing so should be identified or even created. One such existing mechanism identified by workshop participants for the UK was the Full Disclosure initiative hosted at the British Library (British Library Board, 2004).

Greater participation by consortia in international activities, such as the PCC (2004 www.loc.gov/catdir/pcc/ (accessed 12 January 2005)), should be encouraged. This would reconcile clashes between local and global name and subject headings, and ensure future interoperability with international distributed union catalogues. In addition, consortia should consider developing a shared cataloguing service for digital resources, involving the creation of only one catalogue record to be used, or copied, by all member libraries. Rules for cataloguing digital resources tend to offer more choice, and therefore, greater opportunity for variations and increased interoperability difficulties. As workshop participants recognised, there is much less need, if any, for local data in the catalogue record for a resource that is not circulated or shelved.

The role of communication was also identified by cataloguers and systems librarians as particularly important to ensure that local reviews pertaining to cataloguing and indexing practice resonate with the wider globalisation of bibliographic records. Though e-mail communication was deemed useful, participants agreed that catalogue consortia should develop mechanisms to ensure regular opportunities to discuss issues and review policy or practices. Indeed, participants were unanimous in their concern that providing proper professional advice to colleagues would be unforthcoming if they were unable to discuss views, concerns, and experiences with fellow professionals, or inform themselves of cataloguing developments occurring within their immediate locale. Such concerns appear to be increasing as pressures to reduce costs and develop new services increase and become ubiquitous within LIS circles.

9.2 Standards

Whilst greater acknowledgement of the Bath Profile was deemed necessary amongst libraries, Z39.50 implementers, and library system vendors, it was recognised that further development work on the Bath Profile should encompass recommendations for the scope and content of specified indexes. The Bath Profile offers greater prescription, yet there still remains a plethora of local choices to be made during Z39.50 installation and implementation. Such choices are not informed by global interoperability requirements and constitute further obstacles to improving interoperability. For example, the title index could be scoped to cover alternate titles, uniform titles, group and part titles, and related titles. Further still, a normalisation rule could be applied to all scoped titles. This, for example, might entail the removal of leading articles, such as "The" and "An". Such improvements in the Bath Profile would give cataloguing consortia, system vendors, and Z39.50 service developers a sound basis for establishing standard index mappings from metadata formats such as MARC21. In any case, participants suggested that consortia using Z39.50 should consider producing guidelines on required conformance with the Bath Profile, specifying conformance areas and specific indexes and searches. This would be more prescriptive than the profile itself, and by reducing choice would arguably improve interoperability.

In addition, standard rules for index content normalisation could be specified (and adopted) at as wide a level as possible. Such rules would obviously cover punctuation in names, titles and subjects, the inversion of personal names, and the treatment of leading articles in titles. Standard rules would allow system vendors and service developers to ensure more uniformity in Z-indexes. The adaptation of existing rules, such as those used by the Name Authority Control (NACO), was deemed by workshop participants to be feasible and wholly desirable.

9.3 Strategic developments

The most significant strategic development likely to impinge on future library system design is the gradual shift towards IFLA's Functional Requirements for Bibliographic Records (FRBR) model (IFLA, 1998). Indeed, as Tillett (2004, p. 7) notes, "vendors and bibliographic utilities like VTLS, OCLC, and RLG have already embraced the FRBR conceptual model in designing their future systems". The FRBR approach is based on an entity-relationship model as a generalised view of the global bibliographic universe. Such a model offers a new perspective on the composition and relationships of bibliographic and authority records, as well as greater precision in the vocabulary used to describe information entities. Whilst the adoption of FRBR has been slower in the US, enthusiastic application in Europe and Australia has compelled the LIS fraternity to begin co-ordinated planning to ensure a smooth transition.

Consequently, workshop participants noted that consortia and individual libraries should monitor the implementation of FRBR to plan for large-scale machine processing of catalogue data to improve interoperability. Upgrading a cataloguing system to the FRBR model requires disaggregation of existing catalogue record components and reaggregation into a significantly different structure (Delsey, 2004). In particular, the true benefits to be derived from the FRBR model are obtained when the catalogues are used in a global environment. This dictates that the effectiveness of the FRBR model depends on precision in name and title indexes, thus facilitating a degree of automated conversion whereby local records are matched and upgraded against fuller, more authoritative global files (Tillett, 2004). Of course the costs of implementing the FRBR model within a local catalogue are likely to be significant and it was recognised that a better return on investment would be secured if a "global context" was applied to all operations involving library catalogues where possible, rather than simply upgrading to FRBR because it is what their library system vendor is offering. More significantly, by applying a global context to all operations, interim and future interoperability would be improved and the future "FRBR-isation" of catalogue data optimised and rendered more manageable.

9.4 End users

An intriguing outcome of the workshops was the suggestion that disclosure of local practices could affect interoperability for end users by influencing their search behaviour. Such information might be embedded within the catalogue interface, or offered via help, orientation, or training screens. Opinion was divided as to how frank such information should be, particularly if it emphasised potentially negative issues such as incomplete catalogues or poor quality records. In point of fact, many questioned whether end users would be interested in, or use, this kind of support. Other participants suggested that service administrators may incur the displeasure of more experienced searchers (academics, research staff, etc.) if they were not informed of those factors that could affect their entire search strategy. Nevertheless, participants were in agreement that consortia should consider agreeing a standard set of information about each catalogue which should be disclosed as part of the union catalogue service, allowing additional information to be disclosed on the local catalogue interface at the discretion of the library.

10. Conclusions and wider relevance

Whilst the various observations and conclusions drawn in this paper are derived from our analysis of established services within the UK, most, if not all, of the issues are likely to have an international resonance. The primary purpose of this paper has been to identify key differences in the performance of certain physical and virtual bibliographic catalogue services and to provide illustrative examples, as well as to suggest strategies for improving interoperability and performance in physical and virtual systems. Such assessments have to be undertaken since the future potential for creating ever larger heterogeneous Z-based union catalogues increases in parallel with growing library Z-compliance, thus drawing performance issues into an ever sharper focus.

Miller (2000) identifies several types of interoperability, including technical, semantic, political/human and international interoperability. As a Z-enabled service, the technical interoperability of COPAC with other distributed systems was never in doubt. Indeed, as Miller notes, technical interoperability "is the most straightforward aspect of maintaining interoperability, as there are often clear 'right' and 'wrong' answers to be found". However, what the crux of this paper has sought to illustrate is that the various individual configurations permit only limited semantic interoperability, as evidenced by those issues relating to consolidated indexes, data currency and the support or treatment of Bib-1 attributes. Perhaps more importantly, these considerations on semantic interoperability will apply equally to SRW, which, although lowering the barriers to future Z39.50-style implementations, will suffer from sub-optimal performance as a consequence of poor semantic interoperability. Such sub-optimal performance has the potential to be more pronounced in coming years with the increasing prevalence of FRBR, where the requirement for semantic specificity will be essential in order to expedite the coherent and meaningful distributed services that users have come to expect.

To meet these expectations it is imperative that significant prescription be introduced into any cataloguing and indexing guidelines adopted by library consortia (or a union catalogue service) in order to thwart those variations in cataloguing and indexing practices that are currently compromising services. Such prescription is not only essential to improve interoperability and performance, but is necessary to secure the confidence of end users, some of whom already harbour little confidence in union catalogue services (Booth and Hartley, 2004). More generally, however, greater strategic guidance is required from international LIS bodies to plan for future supra-national catalogues. It is therefore not unfeasible to suggest that prescriptive guidelines for the Anglophone be developed via a partnership of national libraries (Library of Congress, British Library, National Library of Australia, etc.), in tandem with further Bath Profile development as outlined previously. Whilst interoperability at an international level should be aspired to, such a partnership would function as a catalyst for wider international interoperability initiatives and, if nothing else, would unquestionably constitute a lesson in "political/human interoperability". Nevertheless, greater participation in the PCC, consideration of future FRBR migration, and greater disclosure to end users are all activities in which individual libraries and library consortia can actively influence and improve interoperability. In short, libraries need to think globally before acting locally.

References

Bath Profile Maintenance Agency (2004), *The Bath Profile: An International Z39.50 Specification for Library Applications and Resource Discovery – Release 2.0*, available at: www.collectionscanada.ca/bath/tp-bath2-e.htm (accessed 12 January 2005), Library and Archives of Canada, Ottawa.

Booth, H. and Hartley, R.J. (2004), *User Behaviour in the Searching of Union Catalogues: An Investigation for Work Package C of CC-interop*, available at: http://ccinterop.cdlr.strath.ac.uk/documents/finalreportWPC.pdf (accessed 12 January 2005), Centre for Digital Library Research, Glasgow.

British Library Board (2004), *What is Full Disclosure?*, available at: www.bl.uk/about/cooperation/fdhome.html (accessed 12 January 2005), British Library, London.

CAIRNS Cataloguing and Indexing Working Group (2000), *CAIRNS Project Recommendations for a Cataloguing and Indexing Strategy for Scottish Libraries*, available at: http://cairns.lib.gla.ac.uk/docs/CAIRNSCatStrat.pdf (accessed 12 January 2005), University of Glasgow, Glasgow.

Cannell, S. and Guy, F. (2001), "Cross-sectoral collaboration in the choice and implementation of a library management system: the experience of the University of Edinburgh and the National Library of Scotland", *Program: Electronic Library and Information Systems*, Vol. 35 No. 2, pp. 135-56.

Caplan, P. and Haas, S. (2004), "Metadata rematrixed: merging museum and library boundaries", *Library Hi-Tech*, Vol. 22 No. 2, pp. 263-9.

Cousins, S.A. (1997), "COPAC: the new national OPAC service based on the CURL database", *Program: Electronic Library and Information Systems*, Vol. 31 No. 1, pp. 1-21.

Cousins, S. (1999), "Virtual OPACs versus union database: two models of union catalogue provision", *The Electronic Library*, Vol. 17 No. 2, pp. 97-103.

Delsey, T. (2004), "Functional requirements for bibliographic records – user tasks and cataloguing data: part 2", *Catalogue and Index*, No. 151, pp. 1-4.

Dunsire, G. and Macgregor, G. (2003), "Clumps and collection description in the information environment in the UK, with particular reference to Scotland", *Program: Electronic Library and Information Systems*, Vol. 37 No. 4, pp. 218-25.

Dunsire, G. and Macgregor, G. (2004), *Improving Interoperability in Distributed and Physical Union Catalogues Through Co-ordination of Cataloguing and Indexing Policies*, available at: http://ccinterop.cdlr.strath.ac.uk/documents/CCICatInterop.pdf (accessed 12 January 2005), Centre for Digital Library Research, Glasgow.

Friesen, N. (2002), *Semantic Interoperability and Communities of Practice*, available at: www.cancore.ca/documents/semantic.html (accessed 12 January 2005), Canadian Core Learning Resource Metadata Applications Profile, St Edmonton.

Gatenby, J. (2002), "Aiming at quality and coverage combined: blending physical and virtual union catalogues", *Online Information Review*, Vol. 26 No. 5, pp. 326-34.

Gilby, J. and Dunsire, G. (2004), *COPAC/Clumps Continuing Cooperation Project (CC-interop): Final Report*, available at: http://ccinterop.cdlr.strath.ac.uk/documents/CCiFinalReportVersion1.pdf, Centre for Digital Library Research, Glasgow.

Gilby, J. and Sanders, A. (2003), *Transforming a Clump into a Z-Target: A Feasibility Study*, available at: http://ccinterop.cdlr.strath.ac.uk/documents/Clump_ztarget_issue1_0.pdf (accessed 12 January 2005), Centre for Digital Library Research, Glasgow.

Gilby, J., Sanders, A. and Cousins, S. (2004), *Bibliographic Union Catalogue Results and Display Issues*, available at: http://ccinterop.cdlr.strath.ac.uk/documents/WPALastReportIssue1.pdf (accessed 12 January 2005), Centre for Digital Library Research, Glasgow.

Hammer, S. and Andresen, L. (2002), *Issues in Z39.50 Parallel Searching*, available at: www.deflink.dk/upload/doc_filer/doc_alle/895_Issues%20in%20Z39_50%20Parallel%20 Searching.htm (accessed 12 January 2005), Denmark's Electronic Research Library, Copenhagen.

Heiler, S. (1995), "Semantic interoperability", *ACM Computing Surveys*, Vol. 27 No. 2, pp. 271-3.

IFLA (1998), "Functional requirements of bibliographic records," Final Report, K.G. Saur, Munchen, available at: www.ifla.org/VII/s13/frbr/frbr.pdf (accessed 12 January 2005), .

Index Data (2004), *Yaz*, Index Data, available at: www.indexdata.dk/yaz/ (accessed 12 January 2005).

Jordan, J. (2003), "OCLC and the emerging worldwide library cooperative", *Library Management*, Vol. 24 No. 3, pp. 107-15.

Lynch, C.A. (1997), "The Z39.50 information retrieval standard – Part 1: a strategic view of its past, present and future", *D-Lib Magazine*, Vol. 3 No. 4, available at: www.dlib.org/dlib/april97/04lynch.html (accessed 12 January 2005).

Macgregor, G. (2005), "Z39.50 broadcast searching and Z-server response times: perspectives from CC-interop", *Online Information Review*, Vol. 29 No. 1, pp. 90-106.

McCulloch, E. (2004), "Multiple terminologies: an obstacle to information retrieval", *Library Review*, Vol. 53 No. 6, pp. 297-300.

McCulloch, E., Shiri, A. and Nicholson, D. (2005), "Challenges and issues in terminology mapping: a digital library perspective", *Electronic Library*, Vol. 23 (in press).

Miller, P. (2000), "Interoperability: what is it and why should I want it?", *Ariadne*, No. 24, available at: www.ariadne.ac.uk/issue24/interoperability/ (accessed 12 January 2005).

Moen, W.E. (2001a), "Improving Z39.50 interoperability: Z39.50 profiles and testbeds for library applications", *Proceedings of the 67th International Federation of Library Associations Council and General Conference, Universal Dataflow and Telecommunications Workshop*, Boston, MA, 16-25 August 2001, IFLA, The Hague, available at: www.ifla.org/IV/ifla67/papers/050-203e.pdf (accessed 12 January 2005).

Moen, W.E. (2001b), "Mapping the interoperability landscape for networked information retrieval", *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries*, Roanoke, VA, 24-28 June 2001, University of North Texas, Denton, available at: www.unt.edu/wmoen/publications/MapInteropJCDLFinal.pdf (accessed 12 January 2005).

Moen, W.E. and Murray, K.R. (2003), *Z39.50 Server Implementation Issues and Recommendations*, available at: www.unt.edu/zlot/Deliverables/WA4_del_z-server_v2_krm_23jun2003.doc (accessed 12 January 2005), Texas Center for Digital Knowledge, Denton, TX.

Nicholson, D. (2000), "Clumping towards a UK national catalogue?", *Ariadne*, No. 22, available at: www.ariadne.ac.uk/issue22/distributed/distukcat.html (accessed 12 January 2005).

Nicholson, D. and Shiri, A. (2003), "Interoperability in subject searching and browsing", *OCLC Systems and Services*, Vol. 19 No. 2, pp. 58-61.

Nicholson, D., Denham, M., Dunsire, G. and Gillis, H. (2001) CAIRNS Final Report: An Embryonic Cross-sectoral, Cross-domain National Networked Information Service for Scotland?, available at: http://cairns.lib.gla.ac.uk/cairnsfinal.pdf (accessed 12 January 2005), Glasgow University, Glasgow.

Nicholson, D., Dunsire, G., Dawson, A., Macgregor, G., Shiri, A., Joseph, A., Williamson, A. and Jones, E. (2004) Elements and Future Development Requirements of a Common Information Environment for Scotland: Final Report to the Scottish Library and Information Council (SLIC) on the SPEIR Project, available at: http://speir.cdlr.strath.ac.uk/documents/SPEIRPublishedVersions/0SPEIRFINALReport261004.pdf (accessed 12 January 2005), Centre for Digital Library Research, Glasgow.

Nicolaides, F. (2003a), *A Comparative Study of the Performance of COPAC and Selected Independent Z39.50 Servers*, available at: http://ccinterop.cdlr.strath.ac.uk/documents/WPA_server_tests_issue1.pdf (accessed 12 January 2005), Centre for Digital Library Research, Glasgow.

Nicolaides, F. (2003b), "The bath profile four years on: what's being done in the UK?", *Ariadne*, No. 36, available at: www.ariadne.ac.uk/issue36/bath-profile-rpt/intro.html (accessed 12 January 2005).

OCLC (2004), *WorldCat at a Glance*, available at: www.oclc.org/worldcat/about/default.htm (accessed 12 January 2005), OCLC, Dublin, OH.

Simeoni, F. (2004), "Servicing the federation: the case for metadata harvesting", in Heery, R. and Lyon, L. (Eds), *Proceedings of the 8th European Conference on Digital Libraries: Research and Advanced Technology for Digital Libraries*, 12-17 September 2004, Springer, Berlin.

Stubley, P., Bull, R. and Kidd, T. (2001), *Feasibility Study for a National Union Catalogue – Final Report*, available at: www.uknuc.shef.ac.uk/NUCrep.pdf (accessed 12 January 2005), University of Sheffield, Sheffield.

Taylor, S. (2003), "A quick guide to Z39.50", *Interlending and Document Supply*, Vol. 31 No. 1, pp. 25-30.

Tedd, L.A. (1994), "OPACs through the ages", *Library Review*, Vol. 43 No. 4, pp. 27-37.

Tillett, B. (2004), *What is FRBR? A Conceptual Model for the Bibliographic Universe*, available at: www.loc.gov/cds/downloads/FRBR.PDF (accessed 12 January 2005), Library of Congress Cataloging Distribution Service, Washington, DC.

Whitelaw, A. and Joy, G. (2001), *Summative Evaluation of Phase 3 of the eLib Initiative: Final Report*, available at: www.ukoln.ac.uk/services/elib/papers/other/summative-phase-3/elib-eval-main.pdf (accessed 12 January 2005), UKOLN, Bath.

Yu, S-C. and Chen, R-S. (2003), "Web services: XML-based system integrated techniques", *Electronic Library*, Vol. 21 No. 4, pp. 358-66.

ZING (2004), *ZING – Z39.50 International: Next Generation*, available at: www.loc.gov/z3950/agency/zing/zing-home.html (accessed 12 January 2005), Library of Congress, Washington, DC.

Z39.50 (2003), *Bib-1 Attribute Set, Z39.50: International Standards Maintenance Agency*, available at: www.loc.gov/z3950/agency/defns/bib1.html (accessed 12 January 2005), Library of Congress, Washington, DC.

Z39.50 (2004), *Z39.50: International Standards Maintenance Agency*, available at: www.loc.gov/z3950/agency/ (accessed 12 January 2005), Library of Congress, Washington, DC.

Further reading

Dovey, M. (2005), "So you want to build a union catalogue?", *Ariadne*, No. 23, available at: www.ariadne.ac.uk/issue23/dovey/intro.html (accessed 12 January 2005).

TZIG (2003), *Z Texas Profile: A Z39.50 Specification for Library Systems Applications in Texas – Release 3.0*, available at: www.tsl.state.tx.us/ld/projects/z3950/tzigprofilerelease30.html (accessed 12 January 2005), Texas State Libraries and Archives Commission, Austin.

The Emerald Research Register for this journal is available at
www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1468-4527.htm

Received 1 October 2004
Accepted 7 October 2004

Z39.50 broadcast searching and Z-server response times

Perspectives from CC-interop

George Macgregor
Centre for Digital Library Research (CDLR), Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK

Abstract

Purpose – Aims to focus on research and findings relating to the Z-server response times and the performance of Z39.50 for parallel searching.

Design/methodology/approach – This paper begins by briefly outlining the evolution of Z39.50 and the current trends, including the work of the JISC CC-interop project. The research crux of the paper focuses on an investigation conducted with respect to testing Z39.50 server (Z-server) response times in a broadcast (parallel) searching environment. Customised software was configured to broadcast a search to all test Z-servers once an hour, for 11 weeks. The results were logged for analysis.

Findings – Most Z-servers responded rapidly. "Network congestion" and local online public catalogue usage were not found to influence Z-server performance significantly. Response time issues encountered by implementers may be the result of non-response by the Z-server and how Z-client software deals with this. The influence of "quick and dirty" Z39.50 implementations is also identified as a potential cause of slow broadcast searching.

Research limitations/implications – The paper indicates various areas for further research, including setting shorter time-outs and greater end-user behavioural research to ascertain user requirements in this area. The influence more complex searches, such as Boolean, have on response times and suboptimal Z39.50 implementations are also emphasised for further study.

Practical implications – This paper informs the library and information science (LIS) research community and has practical implications for those establishing Z39.50 based distributed systems, as well as those in the web services community.

Originality/value – The paper challenges popular LIS opinion that Z39.50 is inherently sluggish and thus unsuitable for the demands of the modern user.

Keywords Z39.50, Online catalogues, Information retrieval

Paper type Case study

Introduction

It is often forgotten that Z39.50 protocol has existed, in one form or another, for almost 30 years. Still, it was only in 1995, with approval granted by the National Information Standards Organisation (NISO), that the standard attracted significant attention from the library and information science (LIS) community, as well as some minor acknowledgement from beyond the library world (Needleman, 2002, p. 248). By the late-1990s, this attentiveness had spread internationally and had manifested itself in a

The author wishes gratefully to acknowledge Ashley Sanders and. Shirley Cousins of Copac for their contribution to JAFER configuration, Z39.50 testing and results processing. Acknowledgement is also extended to John Gilby of the M25 systems team, for his insightful comments and assistance throughout.

Online Information Review
Vol. 29 No. 1, 2005
pp. 90-106
© Emerald Group Publishing Limited
1468-4527
DOI 10.1108/14684520510583963

flurry of Z-based research projects and activity, particularly in the UK where the third phase of the Electronic Libraries programme (eLib) stimulated the creation and evolution of several virtual union catalogues (or "clumps" as they became colloquially known) (Dovey, 2000).

Yet perhaps more incredibly, it is only now that deployment of Z39.50 within the library and information services sectors is truly reaching "critical mass". Z-enabled OPACs are, as Needleman (2002, p. 249) notes, now commonplace within the academic and research library fraternities, an observation that could not have been made until recently. Indeed in the UK, as in many information rich countries, Z39.50 is now gaining prevalence within further education and public library sectors, thus facilitating the creation of ever larger, heterogeneous, virtual union catalogues and cracking open the possibilities for distributed searching by end-users (Dunsire and Macgregor, 2003). More intriguingly, it is predicted that with the next revision of Z39.50 scheduled for 2005, those sectors that have hitherto expressed tepid enthusiasm for the standard (museums, archives, and others) will edge closer to Z39.50 compliance (Taylor, 2003). Although this development would undoubtedly uncover a plethora of difficulties and interesting issues pertaining to the interoperability between, and distributed searching of, cross-domain catalogues, it underlines the pervasive nature of Z39.50 and further illustrates the confidence sought by others in a standard that is, by now, ubiquitous in the library community, as well being internationally recognised as the "global standard" for networked information search and retrieval (NISO, 2002, p. 5).

While the advantages of any standard are manifest in its original introduction and adoption, Z39.50 is not without its faults. Some of these have been widely documented (Gatenby, 2002; East, 2003) and examined (Moen, 2001a; Moen and Murray, 2002), while others have undergone thorough analysis under the auspices of the CC-interop project (Nicolaides, 2003; Gilby and Sanders, 2003; Gilby *et al.*, 2004; Dunsire and Macgregor, 2004). Nevertheless it remains true that despite whatever difficulties Z39.50 might present, it continues to rule distributed searching for the library world and will do for the foreseeable future. It constitutes a significant cornerstone in the technical architecture of the UK Joint Information Systems Committee (JISC) Information Environment (IE) (Powell, 2004), continues to be assiduously bandied by library system vendors, and remains a central component of many commercial content management systems (CMS), such as ENCompass (Dietz and Noerr, 2004).

Those technologies expected to eventually supersede Z39.50 entirely, Web Services Technologies (WST), are currently thought to fall short of providing the rich access already offered by Z39.50 (McDonald, 2003) and, as Yu and Chen (2003) note, there are limitations and barriers to be overcome by Web Services, many of which are similar to Z39.50. However, the "Z39.50 International: Next Generation" initiative (ZING, 2004) have been spearheading a flood of immensely exciting experiments and developments, particularly Search/Retrieve Web Service (SRW) and Search/Retrieve URI (SRU). SRW/SRU is an attempt to conflate the powerful capabilities of Z39.50 by implementing them in tandem with updated Web-friendly protocols and technologies, such as Hypertext Transfer Protocol (HTTP) with Simple Object Access Protocol (SOAP), a protocol for Extensible Markup Language (XML) messaging, and by utilising Web Services Definition Language (WSDL) to define the Z39.50 messages. Although promising far greater functionality, developments remain tentative with the first official specification (Version 1.1) only released in early 2004, but coinciding with some tantalising "real life" applications of the protocol via the European Library project (van Venn and Oldroyd, 2004). Indeed, although ZING (2004) are aiming to "lower the barriers to implementation while preserving the existing intellectual contributions of Z39.50" – a move that is hoped will eventually assist wide adoption in the larger web-based community – it will be many years before it is as widely accepted as Z39.50 in the library community. In addition, and perhaps ultimately, SRW will not provide deliverance in respect to semantic interoperability and those variations in cataloguing and indexing practices that continue to blight optimal performance of Z39.50 virtual union catalogues will linger. In any case it would appear that Z39.50 will retain, at least for some time yet, its crown as the "eminent enabling technology for distributed, parallel access to information sources" (Hammer and Andresen, 2002).

To this end JISC in the UK (http://www.jisc.ac.uk/) has been funding research via the CC-interop project (http://cc-interop.cdlr.strath.ac.uk/) into numerous issues, including testing the feasibility of inter-linking between union catalogues, both physical and virtual, as well as investigating the use of collection-level description schemas in relation to physical and virtual union catalogues. The crux of this paper, however, will focus on research and findings relating to Z-Server response times and the performance of Z39.50 for parallel searching.

Before discussing this, it is worth contextualising the said research within the remit of CC-interop. For those unacquainted with the technology, there is also some merit in briefly summarising how Z39.50 functions, however it is not the purpose of the authors to provide an exhaustive explanation of the technical operations of the protocol. For this refer to NISO (2002), Moen (2001b). Lynch (1997) and Taylor (2003).

Z39.50

ANSI/NISO Z39.50 is a communications protocol maintained by the Z39.50 Maintenance Agency at the Library of Congress (Z39.50, 2004), enabling standard messaging between a Z39.50 client (Z-client) and a Z39.50 server (Z-server), and supporting the searching and retrieval of information in all formats in a distributed networked environment. NISO defines Z39.50 yet more simply, as a "standard protocol used by networked computer systems for information retrieval" (NISO, 2002, p. 3).

Essentially Z39.50 functions as a common language allowing interpretation by Z-enabled systems, irrespective of what software, systems, or platforms are in operation at the client or server. Most implementations use the standard TCP/IP internet communications protocol to connect systems and Z39.50-compliant software in order to decipher messages between them for searching and retrieval. By normalizing the messages used by the client and the server, technical interoperability can be achieved. Thus, any search query initiated by the end-user (at the client interface) is immediately translated by the client software for sending to the remote "Z-server" (or "Z-target"). Once the server is in receipt of the search details, it utilises those rules dictated by Z39.50 to decode the search into a format recognised by the local database. These exchanges are defined by attribute sets, the most prevalent of which is the Bib-1 attribute set (Z39.50, 2003). The Bib-1 attribute set underpins the dominant library profiles, such as the Bath Profile (Bath Profile Maintenance Agency,

2004) and the Z-Texas Profile (TZIG, 2003). Once the remote server has decoded the search according to the aforementioned conventions, it initiates the search locally and then returns the results of that search to the client. The results will then be displayed to the user in a pre-determined format. This format will depend on the configuration adopted by the client. Increasingly Z-client software conducts this processing, but more often than not Z-client software either has to be customised or custom software has to be deployed in tandem with the Z-client to undertaken this post-results processing.

Virtual union catalogues and clumps

As Z-client software has developed, and as librarians have recognised the potential for distributed search and retrieval for the end-user, the protocol has made feasible the construction of complex distributed information environments whereupon it is possible for the Z-client to connect to multiple Z-targets. Such an approach allows the user to "broadcast" a single search to multiple Z-enabled catalogues simultaneously and have the results from each catalogue returned and merged into a single result set, perhaps with duplicate records removed depending on Z-client configuration. As mentioned, the late-1990s witnessed a spate of Z39.50 activity as various LIS communities across the globe furiously set about developing virtual union catalogues. The UK was no exception and was the hub of significant activity.

Arising from the Moving to Distributed Environments for Library Services (MODELS) initiative, the JISC-funded electronic libraries programme (eLib), funded the creation of four virtual union catalogue services (or clumps) in 1998 to conduct further research and develop Z39.50 for the purposes of expansive information retrieval in the UK (Stubley, 1998). A "clump" was defined as an aggregation of catalogues, including physical union catalogues; this definition has subsequently been refined to refer to those aggregations that are inherently distributed only, and is now more commonly used to specifically describe aggregations based on Z39.50 (Dunsire and Macgregor, 2003). Although creating a service that would experience wide use by end-users was a tacit objective, the overarching purpose of the clumps was to "kick start critical mass" in the use of Z39.50 and to generate model technical architectures and agreements to precipitate the subsequent growth of new clumps in their various permutations, perhaps even nationally (Whitelaw and Joy, 2001, p. 2).

Of the four clumps created, three were regionally oriented and existing library consortia provided the sure foundation for development:

(1) The Co-operative Academic Information Retrieval Network for Scotland (CAIRNS) (http://cairns.lib.strath.ac.uk/) included members of the Scottish Confederation of University and Research Libraries (SCURL) and is now developed and maintained by the Centre for Digital Library Research (CDLR) at the University of Strathclyde.

(2) M25 Link had six partners drawn from the M25 Consortium of Academic Libraries based in the London area (http://www.m25lib.ac.uk). The resulting distributed catalogue now forms part of the InforM25 service and is maintained for the consortium by the M25 Systems Team.

(3) RIDING included members from the Yorkshire and Humberside Universities Association (YHUA) (http://www.riding.ac.uk/).

(4) Music Libraries Online (MLO) was the only clump not to be regionally focused. Comprising nine UK conservatoire libraries, MLO facilitated distributed access to scholarly music resources (http://www.musiconline.ac.uk/).

All these projects successfully established fully functioning clumps, each with common and peculiar features. CAIRNS, for instance, instantiated a "dynamic clumping" mechanism – or "landscaping mechanism" – based on Conspectus subject strength measurements conducted by the SCURL member libraries (Nicholson *et al.*, 2001), while M25 Link investigated dynamic clumping by geographical zones of London and the availability of periodicals holdings via Z39.50 (Brack *et al.*, 2001).

The CC-interop project

By 2002 JISC had provided a two-year funding grant to the Copac/Clumps Continuing Technical Cooperation Project (CC-interop), a collaborative project involving the M25 Systems Team, CDLR, Manchester Information and Associated Services (MIMAS), RIDING, and latterly the Centre for Research in Library and Information Management (CERLIM). Building on the results and findings of the JISC eLib programme, CC-interop enhanced the "distributed" thread of the JISC Information Environment in that it "aims to bring together, in a virtual *modus operandi*, distributed catalogues to facilitate richer search and retrieval possibilities for users" (Gilby and Dunsire, 2004, p. 4). The inclusion of the Copac service (http://copac.ac.uk/) at MIMAS – a physical union catalogue based on the consolidated bibliographic records of the Consortium of University and Research Libraries (CURL) and searching some 30 million bibliographic records – exemplified the cooperative nature of the project: true collaborative research emanating from both the virtual and physical union catalogues schools of thought.

Ending in the summer of 2004, CC-interop comprised three work packages, each investigating a plethora of issues, including:

- Inter-linking between very large physical union catalogues (i.e. Copac) and large virtual union catalogues (i.e. InforM25).
- The ability to "clump the clumps" thus creating a "hyper-clump".
- Thorough research of collection-level description requirements for such environments;
- Improving interoperability in distributed and physical environments;
- Investigating user requirements and behaviour for union catalogues.

For a greater discussion of the project outcomes and findings refer to Gilby and Dunsire (2004).

It was also within the remit of CC-interop to undertake some investigation of certain Z39.50 performance issues. Naturally, as in any research project, an abundance of noteworthy findings were accumulated in relation to this topic alone. Yet within this, particularly interesting findings pertaining to Z-server responses times were gleaned, and hereupon is a detailed exposition of that research and the results attained.

Research: Z39.50 searching and response times

As noted earlier, Z39.50 is not without its faults. Conducting broadcast searches (or "parallel searches") via Z39.50 is often considered to be sluggish and lacking robustness (Stubley *et al.*, 2001). Such perceptions have been borne out by detailed user studies whereby current user expectations are increasingly influenced by Web searching tools such as Google, to such an extent that failure to achieve rapid retrieval often compels users to abandon searches altogether (Booth and Hartley, 2004). While web search engines have a long way to go before they can address their respective lack of precision, ponderous recall and retrieval of base quality information, the unfortunate fact remains that users increasingly appear to rank speed of delivery over quality. As Nicholas *et al.* (2003) note, user behaviour is increasingly "promiscuous", with users progressively surpassing traditional quality concerns and conforming to the so-called "bouncer" paradigm. As Nicholas *et al.* (2003, p. 28) conclude, "time plainly is a rare commodity". Such developments should not be ignored. Rather, they should inform the subsequent improvement of those services that embrace metadata, as well as informing those pioneering the improvement or augmentation of information literacy orientation at colleges and universities.

Yet since the emergence of Z39.50 the precise cause of this anomaly in performance and the potential for broadcast searching has never undergone detailed scientific or exhaustive study. Instead the LIS community has been exposed to a variety of conclusions based on speculation or conjecture. It has become, as Hammer and Andresen (2002) pertinently note, "a 'folk wisdom' among Z39.50 implementers that the maximum, realistic number of servers to search in parallel was somewhere between 10 and 15", and that "Z39.50 is just inherently clumsy and slow to work with". At this juncture it is worth noting that this area of research is not without some contributions. Exciting, albeit "informal", research conducted by Hammer and Andresen (2002) under the auspices of Denmark's Electronic Research Library (DEF (Danmark's Elektroniske Forskningsbibliotek)) have provided insights to some of the issues CC-interop wished to expose in a UK context. However this research, by their own admission, was not particularly "scientific". Rather, it was an "attempt to move the discussion of parallel Z39.50 applications away from guesswork and in the direction of hard information" (Hammer and Andresen, 2002), on which other studies could construct further investigation. It was therefore this anomaly that CC-interop wished to address. Furthermore, it was also an opportunity to study any specific peculiarities within InforM25 – which would constitute the test-bed for investigation – and inform subsequent CC-interop and JISC IE developments.

Methodology

To enable investigation of the research question, Java Access to Electronic Resources (JAFER) software was configured to execute automated search tests across a number of the InforM25 member libraries, thus allowing the recording of search response times over a considerable period of time. JAFER is an open source software package that has recently been developed as part of the JISC 5/99 programme by staff at Oxford University and is described as a "Java based toolkit for building Z39.50 clients and servers" aimed at "programmers and web developers building resources for teaching and learning" (JAFER, 2003). As well as being freely available, it is built using industry standard tools that are themselves freely available for both Unix and Windows platforms. JAFER is also extremely flexible, supporting a broad selection of record syntaxes, including that of UKMARC and MARC21.

The decision to use JAFER for this experiment was dictated by two factors. First, JAFER had already been deployed in CC-interop successfully to investigate the feasibility of transforming a clump into a Z-target (Gilby and Sanders, 2003) and was *ergo* readily available. Second, it was recognised that for this particular research task JAFER exemplified fitness for purpose and could be easily configured to achieve the desired research aims.

The JAFER client was therefore configured to broadcast a search to those InforM25 library Z-servers that were known to respond. This initially meant that 16 InforM25 libraries were included in the testing. However, by early December 2003, Buckinghamshire Chilterns University College (BCUC) appeared to be responding to the search queries and was therefore added also. The libraries for which results are presented are available in Table I. Testing began on 6 October 2003 and concluded on 23 December 2003.

A simple author test search for “Austen” was broadcasted, using Bib-1 “Use” attribute 1003. Exact attribute settings configured in JAFER for the individual Z-servers were the same as used earlier in the project (see Gilby and Sanders, 2003). JAFER was then configured to broadcast the search to all test Z-servers once an hour and the results were logged for analysis. The time recorded was the duration of initiating the Z39.50 connection between the JAFER client to the Z-servers, as well as the time taken to broadcast the query and receive a response from the Z-server giving the number of records in the result set. This specifically does *not* include the time needed to request and receive individual or groups of records from a Z-server, nor does

Abbreviation	Institution	Library System
Birkbeck	Birkbeck, University of London	Horizon
Brunel	Brunel University	Unicorn
BCUC	Buckinghamshire Chilterns University College	Unicorn
City	City University	Innopac
Hertfordshire	University of Hertfordshire	Voyager
IOE	Institute of Education	Unicorn
Kent	University of Kent	Voyager
LBS	London Business School	Unicorn
Metro/LGU	London Metropolitan University (formerly London Guildhall University)	Innopac
Pharmacy	School of Pharmacy, University of London	Unicorn
Queen Mary	Queen Mary, University of London	Unicorn
St. George’s	St. George;s Hospital Medical School, University of London	Unicorn
St. Mary’s	St. Mary’s College, University of Surrey	Innopac
SAS	School of Advanced Study	Innopac
SOAS	School of Oriental and African Studies, University of London	Innopac
ULL/Heythrop	University of London Library and Heythrop College	Innopac
Wellcome Library	Wellcome Library for the History of Understanding of Medicine	Innopac

Table I.
Libraries for which results are presented

it include any post-processing time or the time taken to display the records received via a user interface. The results give an indication of connection/database search times, wholly independent of the number of records, record type and any client specific processing.

Caveats

While the authors are confident in the methodology deployed, there are several caveats that are worth noting:

- Given the large distributed nature of InforM25, the total data set did have the potential to greatly exceed 17. Regrettably, though, there were a number of Aleph and Talis libraries systems that did not function correctly when connected to JAFER and consequently these libraries had to be excluded from the tests and do not feature in the results. The exact cause for Aleph and Talis systems not connecting with JAFER is not yet known, but early tests indicated that it was attributable to way in which the connection is requested by JAFER. This would necessitate further investigation but does not suggest a fundamental deficiency with the software. Institutional firewalls at some of the Talis sites were also identified. Such sites were removed from the data set to avoid the potentially lengthy negotiations required to have them opened for testing.
- Birkbeck, University of London was offline for significant periods during testing so data on Birkbeck does not appear in all the results. Also, as BCUC data were only recorded during December 2003, any local problems that were present may have affected the results more than would have been the case if they had been recorded for a longer time period.
- As noted, testing was undertaken between 6 October 2003 and 23 December 2003. However, data coverage during this period was not entirely comprehensive as JAFER occasionally runs out of system memory after a few days. When this occurred the software needed to be shut down and restarted. Obviously this marginally reduced the comprehensiveness of the recorded data, but this downtime did not always happen at the same time of day or for particularly long periods of time, so it is considered that this will not have significantly affected the results obtained, nor the observations that it is possible to draw.
- One final issue to note is that the tests were done with JAFER, installed on a PC at MIMAS and connected to the UK's education and research network, JANET. All the tested Z-servers were also connected to JANET, most via the London Metropolitan Area Network. Testing did not reveal any influences on response times due to the various network elements.

Results and discussion

The test results are summarised in the graphs shown in Figures 1-3. Figure 1 shows the frequency of response times for the tested Z-servers (rounded to the nearest 5 milliseconds). The second graph in Figure 2 depicts the way in which the response times varied during the day. Figure 3 has been included to illustrate the percentage of searches per Z-server responding within categorised time periods (in seconds).

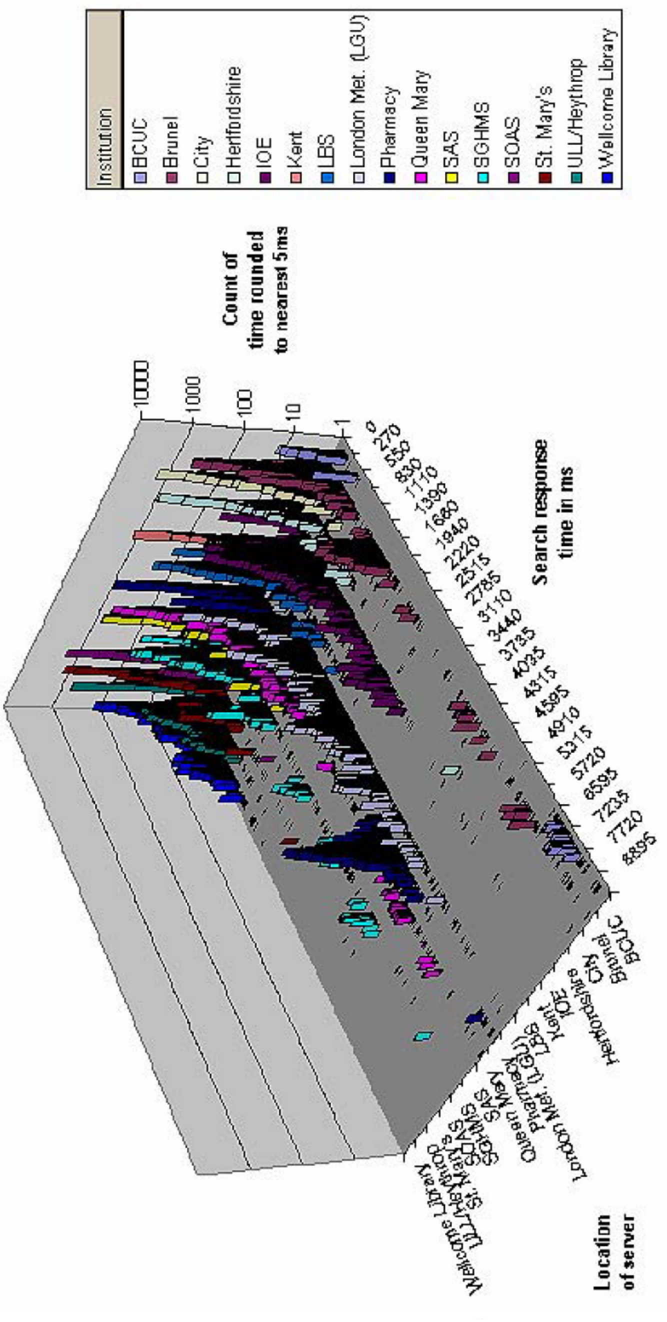


Figure 1.
Frequency of response times for the tested Z-servers

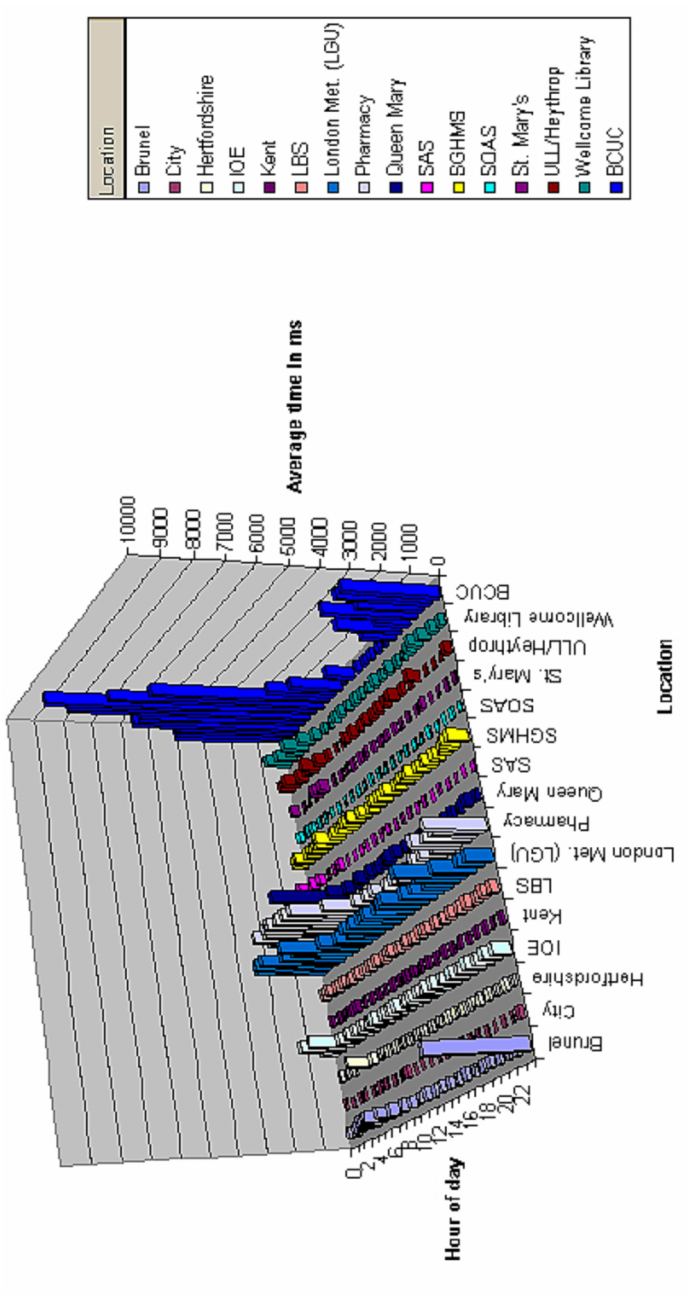


Figure 2.
Average hourly response times for tested Z-servers

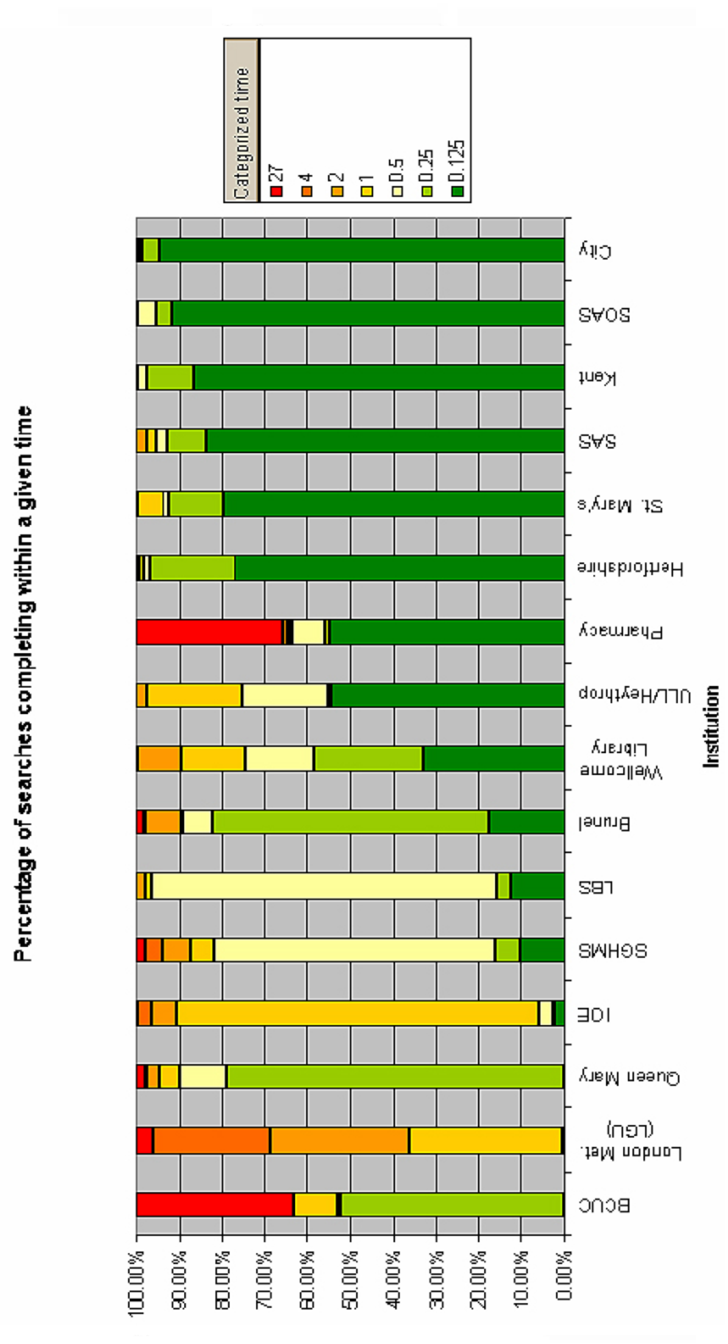


Figure 3.
Percentage of searches per Z-server responding in categorised time periods (seconds)

As can be observed from the clustering of results shown in Figure 1, the majority of responses were received quickly, with approximately 91 percent of searches receiving a response within 1 second. This is what would be anticipated with a very simple query of the type used in the tests. By contrast, approximately 4 percent of all searches took between 4-27 seconds.

As Figure 2 reveals, some Z-servers were consistently fast in their response, indicated in the graph by an almost flat profile. For example, the City Z-server responded to almost 95 percent of searches within 0.125 seconds, with a small number of responses proving lengthier, up to 12.7 seconds in the slowest instance. Other libraries showed a much broader spread of response times, for example London Metropolitan University (LGU) responded to approximately 36 percent of queries within 1 second, approximately 33 percent 1-2 seconds, and approximately 27 percent in 2-4 seconds. BCUC and Pharmacy show a cluster of fast response times, then a cluster of slow ones, with over 34 percent of queries taking 4-27 seconds. In these examples the reasons for the cluster of markedly slower response times are worthy of further investigation, as the systems have revealed that they are perfectly capable of fast responses.

Figure 3 would also suggest that the response time does depend on the type of library system. In most cases the Innopac and Voyager sites (City, SOAS, Kent, SAS, St Mary’s, Hertfordshire) have a very high percentage of response times under 0.25 seconds. Comparing London Met. with the other tested Innopac sites would suggest that there was something different about the Z-server installation at that institution as it constantly responded slowly when compared to other Innopac sites. It is entirely possible, as Moen and Murray (2003) have suggested in other cases, that this delay is attributable to sub-optimal Z-server implementations. Given the constant nature of testing variables, this would be a reasonable assumption to make, but would obviously be no substitution for further testing in our case study. Unicorn sites do generally appear to respond a little more slowly but it is unclear as to the cause of this. As East (2003) and Taylor (2003) have both noted, the implementation of Z39.50 at libraries can be an arduous task for even the most experienced librarians and information professionals (something that ZING developments hope to dissipate). Added to which, those “quick and dirty” implementations favoured by systems vendors often engender yet further obstacles that the librarian has to overcome to ensure an optimal and smooth implementation.

Figure 2 illustrates the range of response times averaged for each hour of the day. As can be observed, the times vary from those Z-servers with a very consistent response time, to others showing large differences in average (mean) response time. For example City and Kent show relatively little variation in response times, while BCUC showed very obvious periods of slow response times, especially during the evening and overnight. It is noteworthy that where response time variations were prominent, the average slowest responses tended to occur early and late in the day, with the fastest responses occurring around mid-day and early afternoon. One probable cause for this is that library system databases often run jobs overnight such as re-indexing and back-ups that tend to take up processing capacity. Library OPACs generally experience higher usage from late morning to early evening so it can arguably be concluded that existing usage of the library system does not directly affect the Z39.50

response times as tested, and vice versa. This resonates with interpretations made by Moen and Murray (2003) and contradicts more popular assumptions that Z39.50 queries are more resource intensive than those queries delivered via the local OPAC (including remote OPAC interrogation over the web). It would also suggest that so-called "network congestion", reputed to occur from late morning to late afternoon, and reputed by the laws of "folk wisdom" to diminish day-to-day Z39.50 performance, is not entirely valid. This latter finding confirms those obtained by Hammer and Andresen (2002). It is worth re-emphasizing, however, that testing carried out by CC-interop did not include the transfer of records, which as well as potentially increasing response times, may perhaps be influenced by the local usage of the library system.

The maximum search time of 27 seconds (Figure 3) is understood to reflect a time-out within JAFER that is initiated so as to avoid the user waiting for slowly responding or non-responding Z-servers. Most distributed systems have a time-out function and if this is too long, searches can appear slow to the user. System designers are presented with a dilemma in that sufficient time needs to be permitted for a slow Z-server to respond, but this is contrasted with the issue of what to do with a Z-server that is not responding at all.

Although there are potentially issues relating to perfunctory Z-server installations, the generally good performance of the Z-servers suggests that many of the response time problems, experienced by searchers conducting broadcast searching for uncomplicated searches, may be the result of non-response by the Z-server and how that is dealt with by the client software. For example, JAFER has a timeout of 27 seconds for non-responding Z-servers, but InforM25 has a cumulative timeout of 65 seconds. More complex searches may, of course, give somewhat different results. It is important to be aware that in InforM25, like many distributed searching environments, the overall searching time experienced by a user is only as fast as the slowest Z-server, so even where most searches are being performed quickly, one slow search is all that is needed to degrade the final response to the user. However it is also important to recognise that not all implementations take this approach and alternatives systems, such as DScovery (Crossnet Systems, 2004) or Metalib (Ex Libris, 2004), can allow users to view results as they are received. This obviously means that the user receives result sets according to Z-target with little post-processing, as opposed to receiving a combined and definitive result set from a service like InforM25.

In an implementation such as InforM25, the test results do appear to suggest that where the slow response of a Z-server adversely affects the overall user query-to-results time, setting a short time-out for the initial Z39.50 connection and search response (e.g. 2 seconds) may help mitigate this. Of course the corollary dictates that those Z-servers that are slow to respond, or which are erratic in their behaviour, may usually be unavailable within a service for searching. Such decisions would have to be taken gingerly by service providers and be taken on a service-by-service basis. Inevitably such a decision would also require detailed analysis of users' requirements. Be that as it may, user behaviour in the searching of union catalogues, as found under CC-interop (Booth and Hartley, 2004), may perhaps suggest that most users would consider such a "trade off" acceptable, especially if it meant that results sets were displayed more quickly.

Conclusions and further research

While the true promise of ZING is afoot and is particularly alluring for the LIS community, it is quite clear that the deployment and uptake of Z39.50 by libraries will not abate for the immediate future. Indeed it is only now, in 2004, that Z-compliance is reaching decisive levels. Such decisive levels of compliance obviously render the creation of large heterogeneous distributed union catalogues ever more likely, and it is therefore imperative that issues pertaining to semantic interoperability and, in our case, performance are addressed to ensure end-users do not consider such retrieval tools as irrelevant in the face of those “low value” tools to which they cling bitterly.

As revealed by the crux of this paper, Z39.50 need not conform to the popular perception that it is “dinosaurian” and too “clunky” or bloated for deployment on the modern web. As Hammer and Andresen (2002) are keen to indicate, Z39.50 is a lightweight protocol, optimised for good performance over large slothful networks. In point of fact, the lightweight genesis and subsequent development of the protocol was necessary as the 1980s imposed severe bandwidth limitations.

The results of this study should hopefully inform further research in the area of Z-server response times. In particular, the community would benefit immeasurably from further research into the effect “quick and dirty” implementations have on Z-server response times, as well as greater technical analysis as to why, in our study, certain library systems appear to greatly influence response rapidity. Moreover, although on site usage of local OPACs did not appear to influence Z39.50 response times as tested, and while Moen and Murray (2003) consider Z39.50 queries to be no more resource intensive on local OPACs than those queries delivered via the local OPAC or over the web, further testing of intensive Z39.50 querying would be prudent so that conclusive data can be gathered on whether such querying could negatively influence Z-server response times and/or local OPAC performance.

There would also be some merit in examining the influence more complex searches, such as Boolean, have on response times. With ever larger virtual union catalogue implementations probable in the future, such research is essential to avoid performance degradation of local OPACs for local user communities, and also to further our collective understanding of Z-server response times generally.

More importantly, greater end-user behavioural research has to be undertaken to ascertain user requirements with respect to the applicability of establishing short time-outs for Z39.50 connections and search responses. Such research would not only inform the Z community (including ZING), but would also inform those champions of Web Services, where the issue of “transaction time” constitutes a significant obstacle for successful Web Services application (Yu and Chen, 2003). In a similar vain, further end-user behavioural studies are required in relation to those Z39.50 implementations like DScovery, as establishing short time-outs may perhaps be a preferable solution if users’ necessity for post-processing is significant.

Ultimately though, semantic interoperability remains the single largest obstacle to improving the overall performance of virtual union catalogues based on Z39.50, an issue that CC-interop grappled with and one that will likely remain atop the LIS agenda even when SRW compliance reaches critical mass.

References

Bath Profile Maintenance Agency (2004), *The Bath Profile: An International Z39.50 Specification for Library Applications and Resource Discover. Release 2.0*, Library and Archives of Canada, Ottawa, available at: www.collectionscanada.ca/bath/tp-bath2-e.htm

Booth, H. and Hartley, R.J. (2004), *User Behaviour in the Searching of Union Catalogues: An Investigation for Work Package C of CC-interop*, Centre for Digital Library Research, Glasgow, //ccinteropcdlrstrathacuk/documents/finalreportWPCpdf

Brack, V., Gilby, J., Gillis, H. and Hogg, M. (2001), "Clumps come up trumps", *Ariadne*, No. 26, available at: www.ariadne.ac.uk/issue26/clumps26/

Crossnet Systems (2004), "Crossnet Systems: DScovery", available at: www.crxnet.com/dscovery.php

Dietz, R. and Noerr, K. (2004), "One-stop searching bridges the digital divide", *Information Today*, Vol. 21 No. 7, p. 24.

Dovey, M. (2000), "So you want to build a union catalogue?", *Ariadne*, No. 23, available at: www.ariadne.ac.uk/issue23/dovey/intro.html

Dunsire, G. and Macgregor, G. (2003), "Clumps and collection description in the information environment in the UK, with particular reference to Scotland", *Program: Electronic Library and Information Systems*, Vol. 37 No. 4, pp. 218-25.

Dunsire, G. and Macgregor, G. (2004), *Improving Interoperability in Distributed and Physical Union Catalogues through Co-ordination of Cataloguing and Indexing Policies*, Centre for Digital Library Research, Glasgow, available at: http://ccinterop.cdlr.strath.ac.uk/documents/finalreportWPC.pdf

East, J.W. (2003), "Z39.50 and personal bibliographic software", *Library Hi-Tech*, Vol. 21 No. 1, pp. 34-43.

Ex Libris (2004), "Metalib: the library portal", available at: www.exlibrisgroup.com/metalib.htm

Gatenby, J. (2002), "Aiming at quality and coverage combined: blending physical and virtual union catalogues", *Online Information Review*, Vol. 26 No. 5, pp. 326-34.

Gilby, J. and Dunsire, G. (2004), *COPAC/Clumps Continuing Cooperation Project (CC-interop): Final Report*, Centre for Digital Library Research, Glasgow, available at: http://ccinterop.cdlr.strath.ac.uk/documents/CCiFinalReportVersion1.pdf

Gilby, J. and Sanders, A. (2003), *Transforming a Clump into a Z-target: A Feasibility Study*, Centre for Digital Library Research, Glasgow, available at: http://ccinterop.cdlr.strath.ac.uk/documents/Clump_ztarget_issue1_0.pdf

Gilby, J., Sanders, A. and Cousins, S. (2004), *Bibliographic Union Catalogue Results and Display Issues*, Centre for Digital Library Research, Glasgow, available at: http://ccinterop.cdlr.strath.ac.uk/documents/WPALastReportIssue1.pdf

Hammer, S. and Andresen, L. (2002), *Issues in Z39.50 Parallel Searching*, Vol. 50, Denmark's Electronic Research Library, Copenhagen, available at: www.deflink.dk/upload/doc_filer/doc_alle/895_Issues%20in%20Z39_50%20 Parallel%20Searching.htm

JAFER (2003), *JAFER Toolkit Website*, available at: www.jafer.org/, Oxford University Library Services, Oxford.

Lynch, C.A. (1997), "The Z39.50 information retrieval standard, part 1: a strategic view of its past, present and future", *D-Lib Magazine*, April, available at: www.dlib.org/dlib/april97/04lynch.html

McDonald, D. (2003), *Web Services Technologies (WST) Report*, JISC, London, available at: www.jisc.ac.uk/uploaded_documents/tsw_03-04.pdf

Moen, W.E. (2001a), “Improving Z39.50 interoperability: Z39.50 profiles and testbeds for library applications”, *Proceedings of the 67th International Federation of Library Associations Council and General Conference, Universal Dataflow and Telecommunications Workshop, IFLA, The Hague, 16-25 August*, available at: www.ifla.org/IV/ifla67/papers/050-203e.pdf

Moen, W.E. (2001b), “Resource discovery using Z39.50: promise and reality”, in Sandberg-Fox, A.M. (Ed.), *Proceedings of the Bicentennial Conference on Bibliographic Control in the New Millennium: Confronting the Challenge of Networked Resources and the Web, 17-20 November*, available at: www.loc.gov/catdir/bibcontrol/moen_paper.html

Moen, W.E. and Murray, K.R. (2002), “A service based approach for virtual libraries: designing and demonstrating a resource discovery service for the Library of Texas”, *Texas Library Journal*, Vol. 78 No. 3, pp. 4-14, available at: www.txla.org/pubs/tlj78/TLJ78_3.PDF

Moen, W.E. and Murray, K.R. (2003), *Z39.50 Server Implementation Issues and Recommendations*, Texas Center for Digital Knowledge, Denton, TX, available at: www.unt.edu/zlot/Deliverables/ WA4_del_z-server_v2_krm_23jun2003.doc

NISO (2002), *Z39.50: A Primer on the Protocol*, NISO Press, Bethesda, MD, available at: www.niso.org/standards/ resources/Z3950_primer.pdf

Needleman, M. (2002), “ZING: Z39.50 international: next generation”, *Serials Review*, Vol. 28 No. 3, pp. 248-50.

Nicholas, D., Dobrowolski, T., Withey, R., Russell, C., Huntington, P. and Williams, P. (2003), “Digital information consumers: players and purchasers: information seeking behaviour in the new digital interactive environment”, *Aslib Proceedings: New Information Perspectives*, Vol. 55 No. 1/2, pp. 23-31.

Nicholson, D., Denham, M., Dunsire, G. and Gillis, H. (2001), *CAIRNS Final Report: An Embryonic Cross-sectoral, Cross-domain National Networked Information Service for Scotland?*, Glasgow University, Glasgow, available at: http://cairns.lib.gla.ac.uk/cairnsfinal.pdf

Nicolaides, F. (2003), *A Comparative Study of the Performance of COPAC and Selected Independent Z39.50 Servers*, Centre for Digital Library Research, Glasgow, available at: http://ccinterop.cdlr.strath.ac.uk/documents/WPA_server_tests_issue1.pdf

Powell, A. (2004), *JISC Information Environment Architecture: Standards Framework: Version 1.1*, UKOLN, Bath, available at: www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/standards/

Stubley, P. (1998), “Clumping in the UK: towards virtual union catalogues”, *New Library World*, Vol. 99 No. 1145, pp. 287-90.

Stubley, P., Bull, R. and Kidd, T. (2001), *Feasibility Study for a National Union Catalogue: Final Report*, University of Sheffield, Sheffield, available at: www.uknuc.shef.ac.uk/NUCrep.pdf

Taylor, S. (2003), “A quick guide to Z39.50”, *Interlending and Document Supply*, Vol. 31 No. 1, pp. 25-30.

TZIG (2003), “Z Texas Profile: a Z39.50 specification for library systems applications in Texas: release 3.0”, available at: www.tsl.state.tx.us/ld/projects/z3950/tzigprofilerelease30.html.

van Venn, T. and Oldroyd, B. (2004), “Search and retrieval in the European Library: a new approach”, *D-Lib Magazine*, Vol. 10 No. 2, available at: www.dlib.org/dlib/ february04/vanveen/02vanveen.html

Whitelaw, A. and Joy, G. (2001), *Summative Evaluation of Phase 3 of the eLib Initiative: Final Report*, UKOLN, Bath, available at: www.ukoln.ac.uk/services/elib/papers/other/summative-phase-3/ elib-eval-main.pdf

Yu, S.-C. and Chen, R.-S. (2003), "Web services: XML-based system integrated techniques", *Electronic Library*, Vol. 21 No. 4, pp. 358-66.

Z39.50 (2003), *Bib-1 Attribute Set, Z39.50: International Standards Maintenance Agency*, Vol. 50, Library of Congress, Washington, DC, available at: www.loc.gov/z3950/agency/defns/bib1.html

Z39.50 (2004), *Z39.50: International Standards Maintenance Agency*, Vol. 50, Library of Congress, Washington DC, available at: www.loc.gov/z3950/agency/

ZING (2004), *ZING – Z39.50 International: Next Generation*, Library of Congress, Washington, DC, available at: www.loc.gov/z3950/agency/zing/zing-home.html

Further reading

Dempsey, L. and Russell, R. (1997), "Clumps or . . . organised access to printed scholarly material: outcomes from the third MODELS workshop", *Program: Electronic Library and Information Systems*, Vol. 31 No. 3, pp. 239-49.

PublicTechnology.net (2004), "Research Libraries Network promises UK researchers joined-up services", 30 July, available at: www.publictechnology.net/ modules.php?op = modloadandname = News&file = article&sid = 1504& newlang = eng

RSLG (2003), *Research Support Libraries Group: Final Report*, Higher Education Funding Council for England, London, available at: www.rslg.ac.uk/final/final.pdf

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0024-2535.htm

DIGITAL DIRECTIONS

Collaborative tagging as a knowledge organisation and resource discovery tool

George Macgregor and Emma McCulloch
Centre for Digital Library Research, Department of Computer & Information Sciences, University of Strathclyde, Glasgow, UK

Received 6 February 2006
Reviewed 19 February 2006
Revised 20 February 2006
Accepted 24 February 2006

Abstract

Purpose – The purpose of the paper is to provide an overview of the collaborative tagging phenomenon and explore some of the reasons for its emergence.

Design/methodology/approach – The paper reviews the related literature and discusses some of the problems associated with, and the potential of, collaborative tagging approaches for knowledge organisation and general resource discovery. A definition of controlled vocabularies is proposed and used to assess the efficacy of collaborative tagging. An exposition of the collaborative tagging model is provided and a review of the major contributions to the tagging literature is presented.

Findings – There are numerous difficulties with collaborative tagging systems (e.g. low precision, lack of collocation, etc.) that originate from the absence of properties that characterise controlled vocabularies. However, such systems can not be dismissed. Librarians and information professionals have lessons to learn from the interactive and social aspects exemplified by collaborative tagging systems, as well as their success in engaging users with information management. The future co-existence of controlled vocabularies and collaborative tagging is predicted, with each appropriate for use within distinct information contexts: formal and informal.

Research limitations/implications – Librarians and information professional researchers should be playing a leading role in research aimed at assessing the efficacy of collaborative tagging in relation to information storage, organisation, and retrieval, and to influence the future development of collaborative tagging systems.

Practical implications – The paper indicates clear areas where digital libraries and repositories could innovate in order to better engage users with information.

Originality/value – At time of writing there were no literature reviews summarising the main contributions to the collaborative tagging research or debate.

Keywords Classification, Controlled languages, Information management, Information retrieval, Knowledge management

Paper type General review

Introduction

Metadata aids the identification, description, management and location of information resources in both digital and non-digital environments. Within the digital environment, the use of metadata to enhance resource discovery continues to be indispensable, particularly within specific communities of practice such as digital libraries or repositories. Metadata can enhance the process of resource discovery by disclosing sufficient information about a resource to enable users or intelligent agents to discriminate between what is relevant and what is irrelevant to a specific information need. Metadata also facilitates approaches to searching or browsing that are simply unfeasible using existing post-coordinate systems (Dawson, 2004). For

Thanks are extended to del.icio.us for granting permission to use screenshots.

Library Review
Vol. 55 No. 5, 2006
pp. 291-300
© Emerald Group Publishing Limited
0024-2535
DOI 10.1108/00242530610667558

example, it enables the location of resources on the subject of "Adam Smith", rather than those written by "Adam Smith".

To facilitate retrieval by subject, information resources are manually assigned subject headings according to their content or, to use cataloguing parlance, "aboutness". Such subject descriptors are commonly known as *index terms* and these are derived from a larger set of index terms known as an *indexing language*. An indexing language constitutes a defined set of terms (or classes) utilising established conventions for ordering and combining terms. The order and arrangement of these terms affect the specificity and exhaustivity of the indexing language. Therefore, terms assigned to resources that are exhaustive will result in high recall at the expense of precision. Conversely, terms that are too specific will result in high precision, but lower recall (Maltby, 1975). To ensure effective indexing and to maintain the overall efficacy of the retrieval system, it is necessary to apply some degree of control to the indexing process. By controlling the indexing process using a so-called *controlled vocabulary*, index terms are standardised and similar or related resources are collocated for ease of discovery by the user (Lancaster, 1972).

Although they yield many benefits, the preeminence of controlled vocabularies has recently been challenged by the appearance of "collaborative tagging" in a variety of prominent Web-based services (del.icio.us: http://del.icio.us/, CiteULike: www.citeulike.org/, Flickr: www.flickr.com/, etc.). Collaborative tagging has emerged as a means of organising information resources on the Web and is contradictory to the ethos of controlled vocabularies. The use of controlled vocabularies – in conjunction with the wider activity of "high quality" metadata creation (i.e. cataloguing) – remains a skilled process normally undertaken by highly trained information professionals. By contrast, collaborative tagging permits any user to assign keywords (or "tags") to Web content (Golder and Huberman, 2005). The purpose of this brief paper is therefore to provide an overview of the collaborative tagging phenomenon, why it has arisen, the emerging literature, and to highlight the problems and the potential of such approaches for knowledge organisation and general resource discovery. Since many of the difficulties associated with collaborative tagging can only be understood via a comparative analysis with controlled vocabularies, we begin by defining the essential properties of controlled vocabularies to which we will refer later in the paper.

Defining controlled vocabularies

Although similar to an authority list, a controlled vocabulary differs in that it generally incorporates some form of semantic and hierarchical structure (Lancaster, 2003). This structure – and the control exerted over vocabulary – performs several functions:

- It controls the use of synonyms (and near-synonyms) by establishing a single form of the term. This ensures that indexers apply the same terms to describe the same or similar concepts, thus reducing the probability that relevant resources will be missed during a user search (Ranganathan, 1967) (e.g. "car", "automobile", "motorcar", or "motor vehicle", etc.).
- It discriminates between homonyms, allowing the indexer to resolve clashes of meaning that arise when several terms assume the same form but assume distinct meanings (e.g. "Java" the programming language, or "Java" the coffee, or "Java" the island belonging to the large south east Asian archipelago of Indonesia). By controlling homonymy, the probability of noise in users' results

sets is reduced (Ranganathan, 1967). By virtue of eliminating homonymy, any other problems associated with homographs – where terms may assume the characteristics of homonyms, but have different pronunciation – are addressed (e.g. "bass" the musical instrument, or "bass" the marine fish of the family Serranidae). Terms that are spelled identically but have different meanings when pronounced differently (i.e. heteronyms) are also resolved (e.g. "reading" the act of comprehending written or printed characters, or "Reading" the town in Berkshire, England, UK).

- It controls lexical anomalies by minimising any superfluous vocabulary or grammatical variations that could potentially create further noise in the users' results set (Chamis, 1991; Garshol, 2004) (e.g. removing vocabulary that is superfluous to describing the intellectual content of the resource, such as leading articles, prepositions, conjunctions, etc. or ensuring consistency in spelling variants, singular and plural forms, verb tenses, and other grammatical variations).
- As noted above, it unites similar terms, or systematically refers the indexer to closely related alternatives, in order to ensure that similar or related resources are collocated. This is normally achieved by displaying the "genus/species" relationship between terms within some form of semantic hierarchical structure, thus indicating when a subordinate class is a species of the super-ordinate class within which it is hierarchically nested (Maltby, 1975) (e.g. "Leninism" is a species of "communism", which in turn, is a species of "political ideology").
- Where appropriate, syntactic relationships (i.e. non-hierarchical relationships) are accommodated (e.g. "language" is syntactically related to "indexing", even though they are not strictly hierarchically related. That is, the relationship between "language" and "indexing" only arises when a compound class of "indexing language" is created).
- The structure also facilitates the use of codes or notation which can then be associated with terms. Such notation is mnemonic, predictable, and language independent (Broughton, 2004). In the physical environment, such notation also assists in the filing, storage and organisation of resources in libraries or information centres (Vickery, 1971).

Lancaster (2003) identifies and defines three major manifestations of controlled vocabulary: bibliographic classification schemes, subject heading lists and thesauri.

The controlled vocabulary "problem"

Traditional classification methods have long been employed in online services. The BUBL Information Service (http://bubl.ac.uk/) organises its content according to Dewey Decimal Classification (DDC). Renardus (www.renardus.org/) employs DDC to help users navigate selected multilingual subject gateways, demonstrating that standard schemes such as DDC have great potential for interoperability and scalability, as well as knowledge organisation and resource discovery. Many digital library services, such as Scotland's Culture service (www.scotlandsculture.org/), have resources indexed using Library of Congress Subject Headings (LCSH), while Artifact (www.artifact.ac.uk/) employs the Art and Architecture Thesaurus (AAT). Although providing many benefits and opportunities for innovative searching or browsing and interoperability, it has long been recognised that traditional controlled vocabularies (in their various permutations) are not always adequate for online resource discovery. Mai (2004) has summarised difficulties of knowledge representation within established bibliographic classification schemes and Nicholson *et al.* (2001) have identified factors including a lack of, or excessive, specificity in the subject areas of some controlled vocabularies as being an impediment to the adequate description of online collections within specific contexts. The need for some services to implement in-house modifications, their general dependency on significant investments of time, money, training, expertise and professional intervention further discourages their wider adoption within particular communities of practice.

The fundamental obstacle preventing wider deployment of controlled vocabularies is that the proliferation of digital libraries and the Web precedes the ability of any one authority to use traditional methods of metadata creation and indexing. While metadata creation is valuable and indispensable within particular communities of practice, it can be costly to implement and can present significant scaling difficulties (Duval *et al.*, 2002). Advances in research of automatic metadata generation applications is increasing (Greenberg, 2004) and indicates that issues of scaling, efficiency and cost can potentially be ameliorated. It is argued by some researchers that such gains in efficiency, were they to be achieved, would allow information professionals to dedicate their efforts on those intellectually demanding metadata activities necessitating some form of human mediation (i.e. assigning controlled index terms) (Anderson and Perez-Carball, 2001; Greenberg *et al.*, 2006). Until such time automatic applications are fully realised, describing or indexing the corpus of information available on the Web will remain beyond the scope of any one authority. The emergence of "collaborative tagging" is therefore considered by some as a useful way in which to supersede the subject indexing role of the information professional and to facilitate resource discovery and knowledge organisation over the Web (Quintarelli, 2005; Shirky, 2005a).

Collaborative tagging

"Collaborative tagging" describes a practice whereby users assign uncontrolled keywords to information resources. Such tags are used to enable the organisation of information within a personal information space, but are also shared, thus allowing the browsing and searching of tags attached to information resources by other users. It also allows users to tag their information resources with those tags that exemplify popularity. The popularity of tags is determined by their level of use and the most popular are often depicted as a "tag cloud" (see Figure 1). Tags are generally single terms, however the assignation of multiple tags to a single resource can be accommodated by omitting essential syntax or punctuation and by using symbols to combine terms (e.g. information+management).

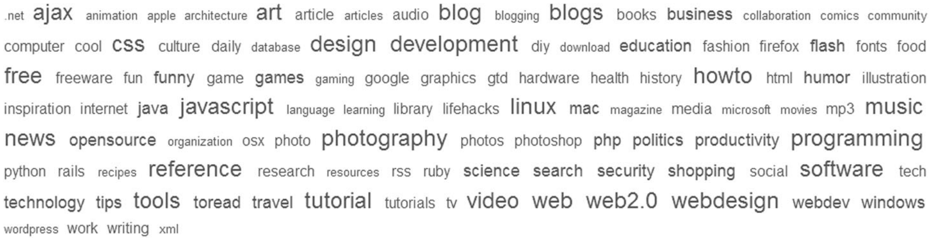

.net ajax animation apple architecture art article articles audio blog blogging blogs books business collaboration comics community
computer cool css culture daily database design development diy download education fashion firefox flash fonts food
free freeware fun funny game games gaming google graphics gtd hardware health history howto html humor illustration
inspiration internet java javascript language learning library lifehacks linux mac magazine media microsoft movies mp3 music
news opensource organization osx photo photography photos photoshop php politics productivity programming
python rails recipes reference research resources rss ruby science search security shopping social software tech
technology tips tools toread travel tutorial tutorials tv video web web2.0 webdesign webdev windows
wordpress work writing xml

Figure 1.
Portion of "tag cloud" as displayed by the collaborative tagging system, "del.icio.us"

The collaborative and ad hoc nature of tagging systems dictates that they lack the essential properties characterising controlled vocabularies (as defined earlier). No control is exerted in collaborative tagging systems over synonyms or near-synonyms, homonyms and homographs, and the numerous lexical anomalies that can emerge in an uncontrolled environment. The probability of noise in a user's result set is therefore very high. The corollary dictates that this impacts negatively upon retrieval precision, as well as limiting the ability to collocate similar or related resources. The inconsistent and ambiguous assignation of tags, and the user proclivity towards exhaustive tags (e.g. "marketing", "technology"), popular tags and personal tags (e.g. "me", "to read") further compromises precision and contributes to high levels of recall and noise also.

Some of the most prominent services incorporating tagging include del.icio.us (http://del.icio.us/), a collaborative bookmarks manager that operates by inviting users to organise their "favourites" in a collaborative environment; Flickr (www.flickr.com/), a Web-based photograph management application; and CiteULike (www.citeulike.org/), a tool for managing and sharing academic papers. Each of these services boasts features geared towards simplifying the process of organising a variety of media, in addition to mechanisms facilitating the future retrieval of such items. Of the aforementioned services, del.icio.us is arguably the most developed and possibly the most collaborative. For example, it combines information gathered from unique identifiers (i.e. the URL) with information gathered about the most popular tags used for that URL. This allows del.icio.us to suggest possible tags when users are bookmarking new resources or to provide users with a list of "common tags" (i.e. popular tags that are assigned to the same resource by multiple users). These common tags can then be used in a subsequent user search strategy. Although Flickr is often discussed as part of the tagging phenomenon, the discrete nature of uploaded objects prohibits such a "close knit society" and thus "collaborative tagging" – as distinct from "tagging" – is not made possible.

Collaborative tagging for knowledge organisation and resource discovery: debate and research

Several authors have documented their thoughts on collaborative tagging but few have done so via the scholarly literature. Discussion of collaborative tagging has instead been most active within the Web blogging community. Vander Wal (2005) and Mathes (2004) have discussed the potential benefits of tagging (as opposed to collaborative tagging) for personal information management (PIM). Vander Wal (2005) has observed that in tagging systems there exists a powerful PIM tool, allowing users to index their information resources with their own vocabulary. Tagging for PIM, however, has inspired far less debate since the benefits for users – although yet to be empirically tested – is quite palpable, understandable, and is not dissimilar to that of file naming or email filtering. Debate has thus concentrated on the use of collaborative tagging for general resource discovery and knowledge organisation on the Web, much of which has been abstract in nature.

Recent debate

Shirky (2005a, b, c) has suggested that the emergence of collaborative tagging on the Web is a "forced move" and hypothesises that tagging will soon supersede controlled vocabularies for the purposes of resource discovery and knowledge organisation. In support of this hypothesis, Shirky (2005a, c) posits the "exclusive" nature of existing controlled vocabularies as impeding their overall usability and suggests that current schemes are incapable of reflecting the transient nature of knowledge and therefore the demands of the modern information user. Shirky suggests that collaborative tagging is inclusive; there is no vocabulary authority imposing a controlled top-down view of knowledge. All users can participate and contribute their own personal vocabularies to generate a collaboratively built "bottom-up" vocabulary which more accurately reflects users' conceptual model of the world around them. The perceived economic advantages of collaborative tagging have also been noted by Shirky (2005a, c). He suggests that the economic advantages will further entrench the practice and make it the preferred strategy for service providers and users in the future. The potential cost reductions available by encouraging communities to undertake indexing themselves, as opposed to relying on professional intervention, undoubtedly contributes to the appeal of collaborative tagging and this particular argument has also been forwarded by other commentators (Quintarelli, 2005; Sterling, 2005).

Davis (2005) has explicitly questioned the economies that can be achieved using tagging. He has argued that any economies achieved in indexing or classifying resources are simply moved onto the price of resource discovery for users, since the lack of collocation increases the number of locations that users have to explore before satisfying their information need. Davis states that the historical purpose of controlled vocabularies has not altered and notes that high costs have always been incurred by a very small number of information professionals in order to reduce the discovery costs for a large number of users. Merholz (2005) has elucidated by providing anecdotal examples from the online reference management service, Connotea (www.connotea.org/). Merholz reveals that a query on the subject of "Avian Flu", for example, exposes twenty-six terms that have been used to describe essentially the same concept.

However, the issue of collocation is considered unimportant by Shirky (2005a). He maintains that the lexical ambiguities inherent in tagging should be permitted to distend since it is through this property that a true representation of knowledge is derived. While cataloguers or indexers will attempt to keep similar or related concepts together, Shirky argues that it is impossible to "collapse" such terms without loosing the essence of what each term conceptually denotes. He therefore states that it is impossible to disentangle terms such as "queer", "gay" or "homosexual" since their meanings are very distinctive and collapsing them together is to misunderstand their conceptual properties. However, Shirky does not discuss how such an approach would scale or impact upon general resource discovery by subject.

Mathes (2004) and Quintarelli (2005) have argued that collaborative tagging can prove beneficial for users' search strategies, providing an increased number of entry points and a measure of serendipity unattainable using controlled vocabularies. Mathes postulates that the serendipitous nature of collaborative tagging, although not necessarily conducive to known-item retrieval or goal-directed browsing, complements *non*-goal-directed searching and browsing by introducing the user to potentially invaluable resources that would otherwise have been undiscoverable. Mathes concludes however that proving or disproving such a hypothesis would require exhaustive large scale qualitative and ethnographic end-user research.

The cognitive processes experienced by users of a collaborative tagging system have been explored by Sinha (2005). She argues that collaborative tagging utilises existing cognitive processes without adding to the cognitive load experienced by the user. She proposes a rudimentary cognitive model of the tagging process and highlights the ability of immediate tagging feedback to circumvent the condition of so-called “post activation analysis paralysis”. According to Sinha, such a condition places the user in a state of cognitive paralysis and is triggered when he/she attempts to tag an information resource to ensure future refindability. Sinha suggests that collaborative tagging reduces the cognitive load experienced by the user because the intellectually onerous task of deciding how a particular resource should be tagged is removed by using system feedback and by observing how others have tagged similar items. Sinha’s hypothesis and conclusions remain untested.

Collaborative tagging research

In one of the few research studies to date, Golder and Huberman (2005) analysed data gathered from del.icio.us to better understand the structure of tagging systems, such as user activity, tag frequencies, the nature of tags used, and so forth. They found that the users of collaborative tagging systems exhibited much variety in the sets of tags they employ. The frequency of tag use and what the tags themselves described was also found to vary greatly between users. However, the data also suggested that there existed some measure of regularity in the tags being assigned by users. On this basis, Golder and Huberman proposed a “dynamical model” of collaborative tagging in which it is possible to predict stable tagging patterns. Their proposed hypothesis remains untested.

Finally, Guy and Tonkin (2006) conducted a small-scale study to assess the “tag literacy” of users and suggest how such literacy might impact on the utility of the tagging approach. Their study involved the analysis of randomly sampled tags from Flickr and del.icio.us. Guy and Tonkin found that 40 per cent and 28 per cent of tags were erroneous in Flickr and del.icio.us respectively. That is, tags were either misspelt, from a language not included in their multilingual dictionary software, in a form that the dictionary could not decode, or were composed of multiple words or a combination of languages. They also found 8 per cent of Flickr tags and 11 per cent of del.icio.us tags to be plural forms and that there existed clear evidence of users deploying the use of various symbols (such as #) at the beginning of tags to influence system filing. Guy and Tonkin consequently propose various system specific strategies for improving the quality of tags (e.g. spelling error checking, suggestion of synonyms, etc.) and encouraging users to observe certain collaborative tagging conventions.

Conclusion: future research and the future of collaborative tagging

Clearly there are numerous difficulties with collaborative tagging, which many proponents have recently been forced to acknowledge. As noted here, most of these difficulties (e.g. low precision, lack of collocation, etc.) originate from the absence of those properties that have come to characterise controlled vocabularies. Commentators on collaborative tagging, such as Quintarelli (2005), consider precision to be unimportant; however it remains difficult to accept that such systems – as an instrument of general resource discovery – will scale and sustain user confidence over the long-term unless they can demonstrate otherwise.

Given some of these basic inadequacies, it is easy to appreciate why tagging has been derided or largely ignored by the LIS community; there appear to be too many irreconcilable problems inherent in the “mass indexing” ethos to envisage it ever superseding more established methods of indexing for knowledge organisation and general resource discovery. Be that as it may, collaborative tagging systems allow users to participate in exciting, highly interactive services and they demonstrate a possible role for users in knowledge organisation and the construction of controlled vocabularies for general resource discovery. There is now momentum behind the development and application of collaborative tagging systems – as the recent acquisition of Flickr and del.icio.us by Yahoo! perhaps demonstrates – and it is quite possible that automated techniques will be deployed to "clean up" and mitigate some of the aforementioned difficulties. However, collaborative tagging systems capable of truly interpreting "the linear unwinding of language" (Foucault, 1977) – as controlled vocabularies and taxonomic classifications do – should not be expected within the foreseeable future.

It is curious to note that during the period in which collaborative tagging has emerged, a reaffirmation of controlled vocabularies has arisen in parallel. The requirement for improved information organisation and management within the corporate sector has facilitated the increased deployment and development of corporate taxonomies (Cruz, 2004; Delphi Group, 2004; Kremer *et al.*, 2005). Similarly, the need for improved subject interoperability within and outside the burgeoning number of distributed digital libraries and digital repositories has also been drawn into sharp focus (McCulloch, 2004; Zeng and Chan, 2004). The need for lexical control, hierarchical structure and associated coding is essential for attaining meaningful subject interoperability across distributed systems (perhaps using different dialects or languages), as well as maintaining the efficacy of subject searching on local systems. To this end librarians and information professionals should be more proactive in extolling the benefits of controlled vocabularies and dispelling the view that controlled vocabularies are inherently non-user-friendly. Controlled vocabularies (or taxonomies) are information tools; a means to an end. For a tool to be useful one has to understand how it operates in order to take advantage of what it offers; such an investment could be considered a "one-off" cost, after which the cost (e.g. time, effort, etc.) of discovery declines. Conversely, collaborative tagging harbours few rules and therefore its use as a "tool" can be quite limited in particular contexts. In stark contrast to a controlled vocabulary, an information literacy session with tagging will not enable a user to improve his/her chances of discovering relevant resources and satisfy an information need since the rules of discovery are not sufficiently predictable nor are they learnable. It therefore becomes a choice between a "perpetual discovery cost" and a "one-off cost".

Equally, collaborative tagging can not be entirely dismissed by librarians or information professionals in the manner that tagging proponents dismiss controlled vocabularies. There are positive lessons to be learned from the interactivity and social aspects exemplified by collaborative tagging systems. Even if their utility for high precision information retrieval is minimal, they succeed in engaging users with information and online communities, and prove useful within PIM contexts. The need to engage users in the development of controlled vocabularies has been recognised by vocabulary experts (Abbott, 2004; Mai, 2004) and collaborative tagging systems could potentially provide a base model for such approaches. Ultimately the dichotomous co-existence of controlled vocabularies and collaborative tagging systems will emerge; with each appropriate for use within distinct information contexts: formal (e.g. academic tasks, industrial research, corporate knowledge management, etc.) and informal (e.g. recreational research, PIM, exploring exhaustive subject areas before formal exploration, etc.).

It is nevertheless clear that the specific factors likely to influence the efficacy of social tagging for resource discovery or knowledge organisation have been ignored and further theoretical analyses are required to facilitate – and to provide focus to – valid and testable hypotheses, and future applied research or enquiry. In particular, the literature to date has focused on the ideological merits (or otherwise) of social tagging, with little attempt being made to understand the theoretical practicalities. This lack of conceptual progress has consequently manifested itself in a lack of testable conceptual models and empirical studies. Librarians and information science researchers – with knowledge of the issues and practicalities surrounding information retrieval with controlled and uncontrolled vocabularies – should therefore be playing a leading role in conducting meaningful research to assess the true value of collaborative tagging in relation to information storage, organisation, and retrieval, and to influence the future development of collaborative tagging systems.

References

Abbott, R. (2004), "Subjectivity as a concern for information science: a Popperian perspective", *Journal of Information Science*, Vol. 30 No. 1, pp. 95–106.

Anderson, J.D. and Perez-Carball, J. (2001), "The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing", *Information Processing and Management*, Vol. 37 No. 2, pp. 231–54.

Broughton, V. (2004), *Essential Classification*, Facet Publishing, London.

Chamis, A.Y. (1991), *Vocabulary Control and Search Strategies in Online Searching*, Greenwood Press, Westport, CT.

Cruz, B. (2004), "Corporate taxonomies can open up the big picture", *Handbook of Business Strategy*, Vol. 5 No. 1, pp. 247–51.

Davis, I. (2005), "Why tagging is expensive", Silkworm Blog, available at: http://silkworm.talis.com/blog/archives/2005/09/why_tagging_is.html (accessed 20 February 2006).

Dawson, A. (2004), "Creating metadata that works for digital libraries and Google", *Library Review*, Vol. 53 No. 7, pp. 347–50.

Delphi Group (2004), *Information Intelligence: Content Classification and the Enterprise Taxonomy Practice*, Delphi Group, Boston, MA, available at: www.delphigroup.com/research/whitepapers/20040601-taxonomy-WP.pdf (accessed 20 February 2006).

Duval, E., Hodgins, W., Sutton, S. and Weibel, S.L. (2002), "Metadata principles and practicalities", *D-Lib Magazine*, Vol. 8 No. 4, available at: www.dlib.org/dlib/april02/weibel/04weibel.html (accessed 20 February 2006).

Foucault, M. (1977), *The Order of Things: An Archaeology of the Human Sciences*, Tavistock Publications, London.

Garshol, L.M. (2004), "Metadata? Thesauri? Taxonomies? Topic Maps! Making sense of it all", *Journal of Information Science*, Vol. 30 No. 4, pp. 378–91.

Golder, S.A. and Huberman, B.A. (2005), *The Structure of Collaborative Tagging Systems*, Information Dynamics Lab: HP Labs, Palo Alto, CA, available at: www.hpl.hp.com/research/idl/papers/tags/tags.pdf (accessed 20 February 2006).

Greenberg, J. (2004), "Metadata extraction and harvesting: a comparison of two automatic metadata generation applications", *Journal of Internet Cataloging*, Vol. 6 No. 4, pp. 59–82.

Greenberg, J., Spurgin, K. and Crystal, A. (2006), "Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions", *International Journal of Metadata, Semantics and Ontologies*, Vol. 1 No. 1, pp. 3–20.

Guy, M. and Tonkin, E. (2006), "Folksonomies: tidying up tags?", *D-Lib Magazine*, Vol. 12 No. 1, available at: www.dlib.org/dlib/january06/guy/01guy/html (accessed 20 February 2006).

Kremer, S., Kolbe, L.M. and Brenner, W. (2005), "Towards a procedure model in terminology management", *Journal of Documentation*, Vol. 61 No. 2, pp. 281–95.

Lancaster, F.W. (1979), *Information Retrieval Systems: Characteristics, Testing and Evaluation*, 2nd ed., John Wiley & Sons, Chichester.

Lancaster, F.W. (2003), *Indexing and Abstracting in Theory and Practice*, 3rd ed., Thomson-Shore Inc., Dexter, MI.

McCulloch, E. (2004), "Multiple terminologies: an obstacle to information retrieval", *Library Review*, Vol. 53 No. 6, pp. 297–300.

Mai, J.E. (2004), "Classification in context: relativity, reality, and representation", *Knowledge Organization*, Vol. 31 No. 1, pp. 39–48.

Maltby, A. (1975), *Sayers' Manual of Classification for Librarians*, 5th ed., Andre Deutsch, London.

Mathes, A. (2004), "Folksonomies – cooperative classification and communication through shared metadata", Adam Mathes.com, USA, available at: http://adammathes.com/academic/computer-mediated-communication/folksonomies.pdf (accessed 20 February 2006).

Merholz, P. (2005), "Clay Shirky's viewpoints are overrated", Peterme.com: links, thoughts, and essays from Peter Merholz, available at: www.peterme.com/archives/000558.html (accessed 20 February 2006).

Nicholson, D., Neill, S., Currier, S., Will, L., Gilchrist, A., Russell, R. and Day, M. (2001), *HILT: High Level Thesaurus Project – Final Report to RSLP & JISC*, Centre for Digital Library Research, Glasgow, available at: http://hilt.cdlr.strath.ac.uk/Reports/Documents/HILTfinalreport.doc (accessed 20 February 2006).

Quintarelli, E. (2005), "Folksonomies: power to the people", *Proceedings of the 1st International Society for Knowledge Organization, UniMIB Meeting, June 24, ISKOI, Milan*, available at: www.iskoi.org/doc/folksonomies.htm (accessed 20 February 2006).

Ranganathan, S.R. (1967), *Prolegomena to Library Classification*, 3rd ed., Asia Publishing House, London.

Shirky, C. (2005a), "Ontology is overrated: categories, links and tags", Shirky.com, New York, USA, available at: http://shirky.com/writings/ontology_overrated.html (accessed 20 February 2006).

Shirky, C. (2005b), "Folksonomies are a forced move: a response to Liz", Many2Many: A group Weblog of social software, available at: www.corante.com/many/archives/2005/01/22/folksonomies_are_a_forced_move_a_response_to_liz.php (accessed 20 February 2006).

Shirky, C. (2005c), "Semi-structured meta-data has a posse: a response to Gene Smith", available at: http://tagsonomy.com/index.php/semi-structured-meta-data-has-a-posse-a-response-to-gene-smith/ (accessed 20 February 2006).

Sinha, R. (2005), "A cognitive analysis of tagging", Rashmi Sinha's weblog, available at: www.rashmisinha.com/archives/05_09/tagging-cognitive.html (accessed 20 February 2006).

Sterling, B. (2005), "Order out of chaos", *Wired Magazine*, Vol. 13 No. 4, available at: www.wired.com/wired/archive/13.04/view.html?pg=4 (accessed 20 February 2006).

Vander Wal, T. (2005), "Explaining and showing broad and narrow folksonomies", vanderwal.net, available at: www.vanderwal.net/random/entrysel.php?blog=1635 (accessed 20 February 2006).

Vickery, B.C. (1971), *Techniques of Information Retrieval*, Butterworths, London.

Zeng, M.L. and Chan, L.M. (2004), "Trends and issues in establishing interoperability among knowledge organization systems", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 5, pp. 377–95.

Corresponding author

George Macgregor can be contacted at: george.macgregor@strath.ac.uk

The E-Resources Management Handbook – UKSG

E-resource management and the Semantic Web: applications of RDF for e-resource discovery

GEORGE MACGREGOR
Information Strategy Group, Information Management & Systems
Liverpool Business School
Liverpool John Moores University

Semantic Web technologies and specifications are increasingly finding applications within digital libraries and other e-resource contexts. The purpose of this chapter is to provide an introduction to some essential Semantic Web concepts and the resource description framework (RDF), a key enabling language of the Semantic Web. Applications of RDF including Dublin Core, FOAF, SKOS and RDFa will be explored with practical examples, and recent implementations of these specifications within a variety of e-resource discovery contexts will be discussed.

Introduction

Recent developments in the Semantic Web offer digital libraries and repositories the opportunity to better expose valuable e-resources using a suite of interoperable standards and technologies. Such tools hold the potential for innovative approaches to the navigation and retrieval of resources within heterogeneous and distributed e-resource environments. The outputs of Semantic Web activity also present opportunities for resolving or ameliorating common problems relevant to digital libraries, such as semantic interoperability and advanced metadata integration. Although the deployment of Semantic Web approaches within digital libraries and repositories is growing, the use of such techniques generally remains confined to particular communities of practice (e.g. research centres, academia, research libraries, etc.). To some extent this is consistent with the wider computing and information profession; however, it is something that has been changing in recent years.

Developments in the Semantic Web are of increasing significance to information professionals. As well as having useful applications within digital libraries, information professionals have an emerging role to play in the development and maintenance of the structured data comprising the Semantic Web (e.g. metadata, ontologies, etc.). The relevance of the Semantic Web to Library and Information Science (LIS) has been reflected in recent research and dissemination activity by information professionals[1,2,3] and many are actively participating in the development of important W3C Semantic Web specifications[4].

Given the relevance of the Semantic Web to LIS, the purpose of this chapter is to provide an introduction to some essential Semantic Web concepts and resource description framework (RDF) specifications. Recent applications of these concepts within a variety of contexts will also be explored, particularly within digital libraries and e-resource discovery. Since RDF and applications of RDF provide a key enabling technology within the Semantic Web, the chapter will introduce RDF using practical examples.

The Semantic Web

The Semantic Web is a research agenda originally initiated by Tim Berners-Lee in 2001[5]. It is now considered to be an evolving extension of the existing web, and the agenda is one that has been reiterated more recently by Berners-Lee and his colleagues as a 'web of data'[6].

The purpose of the Semantic Web is to make the semantics of information and services available on the web interpretable and understandable to machines so that user requests can be more accurately satisfied. The difficulty with the current web is that it has evolved to consist primarily of documents designed for humans to read, rather than for machines. For example, machines can interpret the *syntax* of the web documents (e.g. XHTML) and display these documents to users, but they have little ability to interpret their meaning (i.e. semantics). The intention of the Semantic Web is therefore to deliver a web of data which will better facilitate the extraction of *semantics* from documents by intelligent software agents. Equipped with this semantic knowledge, computers can then actively support users in their information tasks as opposed to passively displaying or delivering information to users.

One obvious area in which this semantic data can be put to good use is information retrieval[7]. For example, if information retrieval systems can better understand the meaning of items within an e-resource collection then it will be easier to design systems that provide greater retrieval precision during users' information-seeking tasks. Increased precision could be achieved by better understanding user context, disambiguating conceptually similar items, performing some of the functions controlled vocabularies might; but improvements in recall could also be achieved by augmenting the results with conceptually related resources, perhaps spanning a variety of media. Although the deployment of the Semantic Web within LIS is our focus, such semantic technologies assume greater potential and complexity when applied to everyday tasks, such as booking a medical appointment[8] or ordering wine for a social event[9]. In such instances numerous applications may be involved, requiring a high level of systems interoperability and a shared level of meaning (i.e. shared semantics) through the use of ontologies.

For the Semantic Web vision to work and for intelligent software agents to have data to harness, resources on the web have to be expressed in a machine-interpretable format. This entails annotating resources with machine-interpretable metadata and other structured data which attempts to capture the semantics of resources. Since the ethos of the web is distributed and since the intention is that Semantic Web data be available for manipulation or reuse by any number of heterogeneous applications, the interoperability of this structured data is absolutely essential. Structured and interoperable data is so fundamental to the success of the Semantic Web that Tim Berners-Lee recently conceded that the 'data web' would have been a better name for his vision[10]. Although there are a number of emerging technologies underpinning the Semantic Web[11], it is the resource description framework (RDF) and its various applications which provide the majority of the structured data required to make the Semantic Web work.

Resource description framework

The resource description framework (RDF)[12] is a framework for modelling and representing data on the web. In fact, RDF is simply a data model in which statements are made about web resources. Each statement made about a resource comprises a collection of 'triples' consisting of a *subject*, *predicate* and *object*. The *subject* denotes the object the triple is describing, the *predicate* identifies the attribute of the subject within the statement, and the *object* defines the value of the predicate. A set of triples is known as an RDF graph and is diagrammed using a series of nodes connected by labelled arcs (Figure 1).

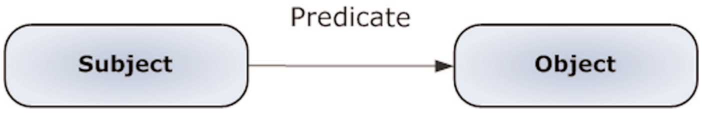


Figure 1. An example of an RDF directed graph

Taken together the subject-predicate-object triple represents a statement of fact about the resource in question and characterizes the nature of the relationship between each node of the directed graph. Consider the following statement as an example:

- 'The title of this chapter is Applications of RDF for e-resource discovery'

Within this statement we can identify the following triple set:

- Subject: `Chapter`
- Predicate: `hasTitle`
- Object: `Applications of RDF for e-resource discovery`

This triple, in turn, could be graphed as in Figure 2.

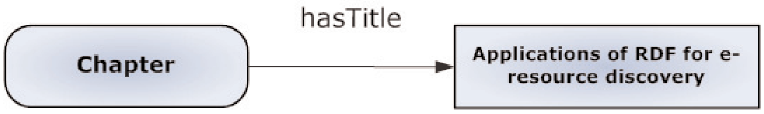


Figure 2. Identifying triples within an RDF graph

Recall that the purpose of the Semantic Web is to provide machine-interpretable statements about resources on the web in order to derive meaning. For the Semantic Web this entails two things: the use of uniform resource identifiers (URIs) and the way of expressing RDF on the web.

Figure 2 illustrates the concept of RDF and triples admirably; however, the English-language text strings used for our triples are more conducive to human interpretation than machine processing. RDF therefore takes advantage of URIs[13] as the principal means of identifying subjects, predicates and objects within RDF triples. Although similar to URLs which *locate* resources, URIs can be far more abstract and can *identify* anything. They can refer to resources available over a network much like a URL but can also refer to non-networked resources (e.g. people, physical documents, places, etc.) and abstract concepts or names which have no physical manifestation (e.g. title, creator, subject). By using URIs within RDF it is therefore possible to describe anything and any type of relationship between these things. The importance of URIs will assume more relevance shortly.

Since RDF is a data model, it remains syntax independent. It is therefore possible to express (or 'serialize') RDF on the web in a variety of ways, including RDF/XML[14], Notation 3 (N3)[15] and Turtle[16]. While the latter two are increasingly popular, RDF/XML continues to be used extensively. The popularity of RDF/XML is attributable to its use of XML[17] to serialize an RDF graph as an XML document. It is used in much of the W3C Semantic Web documentation and continues to be the only serialization recommended by the W3C Semantic Web Activity team[18]. RDF/XML will therefore be the serialization used in examples throughout this chapter.

The importance of RDF/XML and URIs in expressing RDF graphs has been noted and it is now possible to provide an example.

Basic example

In Figure 2 the subject of the RDF graph was `Chapter`. At time of writing, this present chapter lacks an electronic location; however, when it is officially published it will have a URL incorporating the UKSG / MetaPress domain. The URL therefore could be said to be `http://uksg.metapress.com/someURL`.

Dublin Core (DC)[19] metadata allows us to formalize the `hasTitle` predicate from Figure 2 since DC includes a title element fulfilling that purpose. Dublin Core can be expressed as RDF[20] and is defined by an RDF Schema at `http://purl.org/dc/terms`. This allows us to assign a proper predicate for `hasTitle` based not only on a recognized metadata schema, but defined using a URI instead of a text string. In this case `hasTitle` becomes `http://purl.org/dc/terms/title`.

Finally, the object of the RDF graph in Figure 2 is `Applications of RDF for e-resource discovery`. Since this is the value of our object this will remain as a literal (i.e. a text string).

These amendments to the RDF graph allow us to update it accordingly (Figure 3). By doing so we note that the graph now consists of the following triple set:

- Subject: `http://uksg.metapress.com/someURL`
- Predicate: `http://purl.org/dc/terms/title`
- Object: `Applications of RDF for e-resource discovery`

Note also that because the object node in Figure 3 is a literal it is diagrammed as a box.

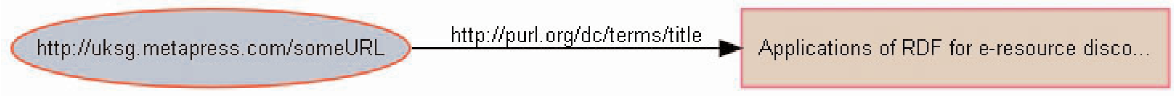

http://uksg.metapress.com/someURL
http://purl.org/dc/terms/title
Applications of RDF for e-resource disco...

Figure 3. A simple RDF statement using Dublin Core

Since providing RDF graphs in a machine-interpretable data format is essential for the Semantic Web to operate, it is possible to express the graph in Figure 3 as RDF/XML. Such a graph would be expressed as follows:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dcterms="http://purl.org/dc/terms/">
<rdf:Description rdf:about="http://uksg.metapress.com/someURL">
 <dcterms:title>Applications of RDF for e-resource discovery</dcterms:title>
</rdf:Description>
</rdf:RDF>
```

The subject and the predicate must always be referenced using a URI. The object is the only component of an RDF triple which is permitted to use literals; but as we have noted in the above example, there are circumstances in which the object must be a literal, often because a URI is inappropriate or unavailable. In the above example the literal was `Applications of RDF for e-resource discovery` and such literals are common when metadata is used. However, the preference in RDF is to use URIs wherever possible to identify triples within an RDF graph so as to aid machine processing and, in many cases, an object URI will be available. Consider the following statement as an example:

- 'The creator of this chapter is George Macgregor'

Within this particular statement we can identify the following triple set:

- Subject: `Chapter`
- Predicate: `Creator`
- Object: `George Macgregor`

With our knowledge of the chapter's URL, of the Dublin Core element set, and of the author's personal homepage (where detailed RDF creator information can be extracted by intelligent software agents), it is possible for us to formalize the triple set using URIs as follows:

- Subject: `http://uksg.metapress.com/someURL`
- Predicate: `http://purl.org/dc/terms/creator`
- Object: `http://www.staff.ljmu.ac.uk/bsngmacg`

Rather than use a literal to describe the creator (i.e. George Macgregor) it is possible for us to reference the creator using a URI. This RDF graph can then be integrated with our previous graph (as in Figure 4) and expressed in RDF/XML as follows:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dcterms="http://purl.org/dc/terms/">
```

```
<rdf:Description rdf:about="http://uksg.metapress.com/someURL">
 <dcterms:title>Applications of RDF for e-resource discovery</dcterms:title>
 <dcterms:creator rdf:resource="http://www.staff.ljmu.ac.uk/bsngmacg/" />
</rdf:Description>

</rdf:RDF>
```

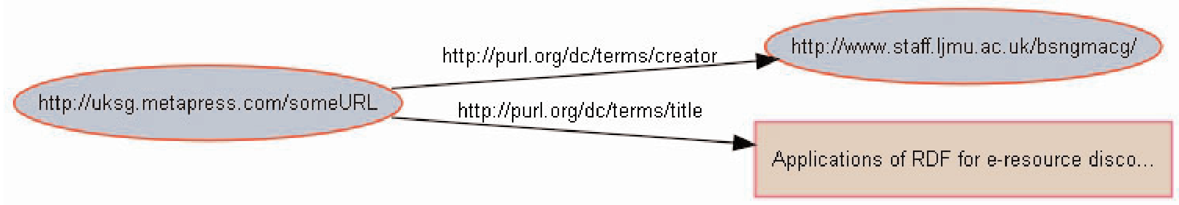

http://uksg.metapress.com/someURL
http://purl.org/dc/terms/creator
http://www.staff.ljmu.ac.uk/bsngmacg/
http://purl.org/dc/terms/title
Applications of RDF for e-resource disco...

Figure 4. Simple RDF graph demonstrating the use of URIs in RDF

If desired, this simple RDF statement could easily be augmented with further Dublin Core metadata elements. For example, publisher information could be included along with Library of Congress Subject Heading (LCSH) descriptor charactering the aboutness of the resource in question, and the rights could be referred to by a Creative Commons licence[21], all of which could be referenced by URI, thus generating the RDF graph in Figure 5 and providing the following RDF/XML:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:dcterms="http://purl.org/dc/terms/">

<rdf:Description rdf:about="http://uksg.metapress.com/someURL"> <dcterms:title>Applications
of RDF for e-resource discovery</dcterms:title>
 <dcterms:creator rdf:resource="http://www.staff.ljmu.ac.uk/bsngmacg/"/> <dcterms:publisher
rdf:resource="http://www.uksg.org/"/> <dcterms:subject
rdf:resource="http://id.loc.gov/authorities/sh2002000569"/> <dcterms:rights
rdf:resource="http://creativecommons.org/licenses/by-nc-sa/2.0/uk/"/></rdf:Description>

</rdf:RDF>
```

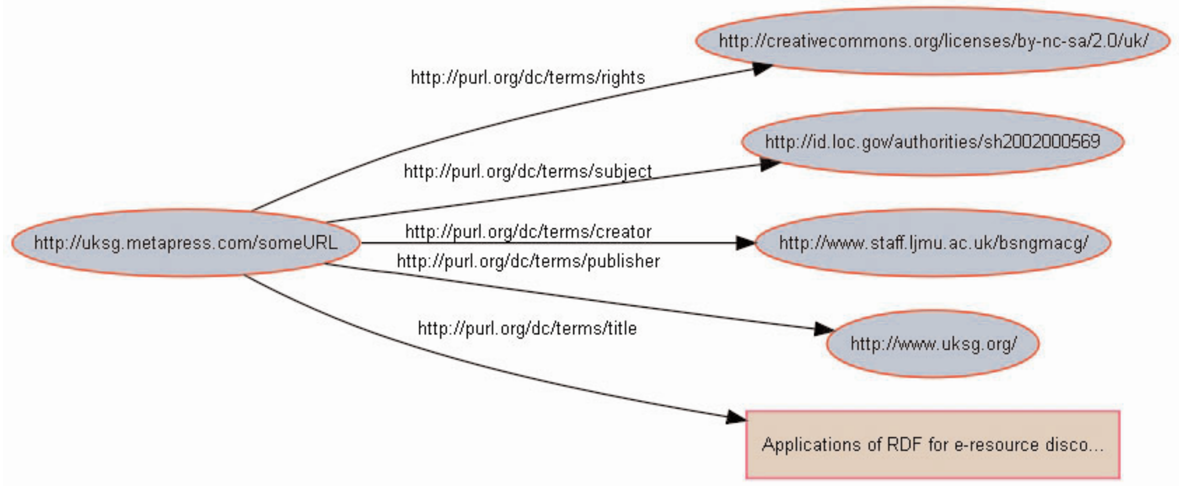

http://creativecommons.org/licenses/by-nc-sa/2.0/uk/
http://purl.org/dc/terms/rights
http://id.loc.gov/authorities/sh2002000569
http://purl.org/dc/terms/subject
http://uksg.metapress.com/someURL
http://purl.org/dc/terms/creator
http://purl.org/dc/terms/publisher
http://www.staff.ljmu.ac.uk/bsngmacg/
http://purl.org/dc/terms/title
http://www.uksg.org/
Applications of RDF for e-resource disco...

Figure 5. Augmenting our simple RDF graph from Figure 4

In Figure 5 we have been able to augment the RDF graph by making greater use of URIs. The decision to use a URI for subject indexing was based on the increasing use of controlled vocabularies on the Semantic Web expressed in RDF. One such example of this is LCSH[22]. The URI of `http://id.loc.gov/authorities/sh2002000569` denotes the LCSH descriptor, 'Semantic Web'. This URI not only defines the concept of the Semantic Web, but at the end of the URI we discover rich terminological data expressed in a variety of Semantic Web-friendly serializations. Referring to controlled vocabularies in this way will be discussed in more detail in the 'Simple Knowledge Organization System (SKOS)' section of this chapter. Of course, in many circumstances literals will suffice and the subject heading used above, for example, could easily be a literal taken from LCSH rather than a URI.

The use of Dublin Core in the Semantic Web is a useful introduction to the basic concepts of RDF and RDF/XML. Additionally, the ability to integrate RDF data on the web means that DC is often used in conjunction with numerous other RDF applications. Note that the RDF/XML examples and the resulting RDF graphs in this section were created using specialist software[23,24]; however, the validity of the RDF/XML examples (and all others in this chapter) can easily be verified by using the W3C RDF Validation Service[25]. This allows the RDF/XML document to be checked and graphed.

The basic concepts and principles of RDF have now been introduced. The remainder of the chapter will now consider some other applications of RDF.

Friend-of-a-friend (FOAF)

Friend-of-a-friend (FOAF)[26] was one of the first applications of RDF and was originally designed as a Semantic Web version of a personal homepage[27]. FOAF is therefore designed to capture metadata about people. The FOAF vocabulary specification[28] provides a rich vocabulary to describe personal information (e.g. name, mailbox addresses, homepage URLs, blogs, etc.), as well as relationships with other people, groups, projects, and other affiliations.

The FOAF vocabulary defines classes (e.g. `foaf:Person`) and numerous properties (i.e. predicates), such as `foaf:name`, `foaf:knows`, `foaf:interests`, `foaf:depiction`, `foaf:weblog`, etc. Once published on the web (e.g. as RDF/XML), FOAF files can be processed by machines to establish relationships between people or organizations and the nature of these relationships. This data can then be used by computers to locate people or groups with similar interests, allow new entrants to a community to understand its structure, manage online personal identities via URIs, and a variety of other uses too numerous to list here[29]. FOAF's ability to characterize social relationships has also led to its use within online social network applications[30].

For example, we might want to state that there exists a person (`foaf:Person`) with the name 'George Macgregor' (`foaf:name`), who has:

- An e-mail address (`foaf:mbox`)
- A homepage (`foaf:homepage`)
- A blog (`foaf:weblog`)
- And, who knows (`foaf:knows`) another person (`foaf:Person`) with the name 'Emma McCulloch', who also has a homepage (`foaf:homepage`).

Such a 'social graph' could be expressed in FOAF RDF/XML as follows:

```
<?xml version=”1.0” encoding=”UTF-8”?>
<rdf:RDF xmlns:foaf=”http://xmlns.com/foaf/0.1/”
xmlns:rdf=”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
xmlns:rdfs=”http://www.w3.org/2000/01/rdf-schema#”>

<foaf:Person rdf:about=”http://www.staff.ljmu.ac.uk/bsngmacg/#me”>
 <foaf:name>Macgregor, George</foaf:name>
 <foaf:mbox rdf:resource=”mailto:g.r.macgregor@ljmu.ac.uk”/>
 <foaf:homepage rdf:resource=”http://www.staff.ljmu.ac.uk/bsngmacg/”/>
 <foaf:weblog rdf:resource=”http://ljmuinfostrategy.blogspot.com/”/>
 <foaf:knows>
  <foaf:Person>
   <foaf:name>McCulloch, Emma</foaf:name>
   <foaf:homepage rdf:resource=”http://cdlr.strath.ac.uk/people/mcculloche.htm”/>
  </foaf:Person>
 </foaf:knows>
</foaf:Person>

</rdf:RDF>
```

Recall that URIs can identify anything, even people. A URI has therefore been used in the above example to identify `foaf:Person` (i.e. George Macgregor). By assigning a URI we eliminate any ambiguity about which 'George Macgregor' is being referred to. Not only that, we enable others in the Semantic Web to refer unambiguously to this 'George Macgregor' rather than others with the same name. This URI could also be used to merge all other RDF data available on the web which happens to reference 'George Macgregor'. Where such a URI is missing, other mechanisms could be used (e.g. e-mail address).

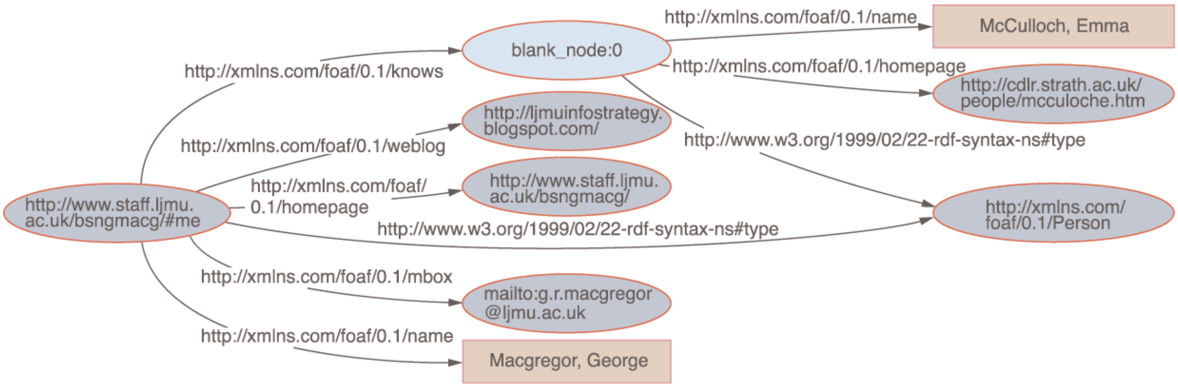

http://xmlns.com/foaf/0.1/name
McCulloch, Emma
blank_node:0
http://xmlns.com/foaf/0.1/knows
http://xmlns.com/foaf/0.1/homepage
http://cdlr.strath.ac.uk/people/mcculoche.htm
http://ljmuinfostrategy.blogspot.com/
http://xmlns.com/foaf/0.1/weblog
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://xmlns.com/foaf/0.1/homepage
http://www.staff.ljmu.ac.uk/bsngmacg/
http://www.staff.ljmu.ac.uk/bsngmacg/#me
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://xmlns.com/foaf/0.1/Person
http://xmlns.com/foaf/0.1/mbox
mailto:g.r.macgregor@ljmu.ac.uk
http://xmlns.com/foaf/0.1/name
Macgregor, George

Figure 6. RDF graph of a FOAF file about 'George Macgregor'

Although the above FOAF RDF/XML example is relatively simple, we can observe from Figure 6 that the resulting RDF graph is already more complex than those featured earlier. A 'blank node' can also be observed in Figure 6. Blank nodes are common in RDF and are often unavoidable. Blank nodes essentially represent nodes which do not have a URI or literal (i.e. they are 'blank'). Such nodes therefore do not contain any data; instead they are used as parent nodes to group data together. For example, in the above example the FOAF RDF/XML essentially states that 'George Macgregor' knows a person whose name is 'Emma McCulloch' and who has a homepage. The `foaf:Person` of 'Emma McCulloch' is not uniquely identified by a URI. Since `foaf:Person` does not have its own URI, properties about 'Emma McCulloch' are grouped together using a blank node. This blank node mimics a URI and provides the necessary linkages between nodes within the RDF graph for it to make sense. In the absence of a URI, the software used to generate the RDF graph in Figure 6 has assigned a blank node identifier (`blank_node:0`). Blank node identifiers have no real meaning within RDF graphs other than allowing us to distinguish between other blank nodes within the same graph, thus most dedicated software applications (including the W3C RDF Validation Service) will assign identifiers automatically. Note that blank node identifiers only identify nodes within the same graph. If there is a need to merge multiple RDF graphs, or if others want to reference a blank node from outside the graph, then URIs have to be used instead.

Of course, it is possible to further augment this FOAF example with properties such as `foaf:gender`, `foaf:depiction`, `foaf:pastProject`, and so forth. More relationships can also be established (`foaf:knows`), as well as personal interests (`foaf:interest`) thus increasing the links within the social graph. The following example augments our FOAF RDF/XML with numerous properties and extra classes. Note also that the blank nodes resulting from the previous example have been resolved by assigning URIs to all instances of `foaf:Person`. The resulting RDF graph is too large to reproduce here but can be verified using the W3C RDF Validation Service:

```
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:foaf="http://xmlns.com/foaf/0.1/"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
```

```
<foaf:Person rdf:about=”http://www.staff.ljmu.ac.uk/bsngmacg/#me”>
 <foaf:name xml:lang=”en”>Macgregor, George</foaf:name>
 <foaf:firstName xml:lang=”en”>George</foaf:firstName>
 <foaf:surname xml:lang=”en”>Macgregor</foaf:surname>
 <foaf:gender>male</foaf:gender>
 <foaf:mbox rdf:resource=”mailto:g.r.macgregor@ljmu.ac.uk”/>
 <foaf:homepage rdf:resource=”http://www.staff.ljmu.ac.uk/bsngmacg/”/>
 <foaf:depiction
rdf:resource=”http://www.staff.ljmu.ac.uk/bsngmacg/img/georgedepiction.jpg”/>
 <foaf:workplaceHomepage rdf:resource=”http://www.ljmu.ac.uk/LBS/92624.htm”/>
 <foaf:publications rdf:resource=”http://www.staff.ljmu.ac.uk/bsngmacg/pubs.html”/>
 <foaf:weblog rdf:resource=”http://ljmuinfostrategy.blogspot.com/” dc:title=”Information
Strategy Group, LJMU - Blog”/>
 <foaf:interest rdf:resource=”http://www.w3.org/2004/02/skos/”/>
 <foaf:interest rdf:resource=”http://hilt.cdlr.strath.ac.uk/”/>
 <rdfs:seeAlso rdf:resource=”http://cdlr.strath.ac.uk/foaf/cdlr.rdf”/>
 <foaf:pastProject>
  <rdf:Description rdf:about=”http://hilt.cdlr.strath.ac.uk/” dc:title=”HILT: High-level
Thesaurus project phase IV”>
  </rdf:Description>
 </foaf:pastProject>
 <foaf:knows>
  <foaf:Person rdf:about=”http://www.staff.ljmu.ac.uk/bsnpkell/#me”>
   <foaf:name>Kelly, Phil</foaf:name>
   <foaf:title>Dr</foaf:title>
   <foaf:homepage rdf:resource=”http://www.ljmu.ac.uk/LBS/92623.htm”/>
  </foaf:Person>
 </foaf:knows>
 <foaf:knows>
 <foaf:Person rdf:about=”http://cdlr.strath.ac.uk/people/mcculloche.htm#me”>
  <foaf:name>McCulloch, Emma</foaf:name>
  <foaf:mbox rdf:resource=”mailto:e.mcculloch@strath.ac.uk”/>
  <foaf:homepage rdf:resource=”http://cdlr.strath.ac.uk/people/mcculloche.htm”/>
  <foaf:depiction rdf:resource=”http://cdlr.strath.ac.uk/people/mcculloche.jpg”/>
 </foaf:Person>
 </foaf:knows>
 </foaf:Person>

</rdf:RDF>
```

Merging of RDF data is where FOAF is potentially of most use to digital libraries. For example, Dublin Core metadata (in RDF) about this chapter could be merged with FOAF metadata (in RDF), thus providing an enhanced metadata record containing rich authorship information. Malmsten[31] describes the use of a series of Semantic Web specifications to build a semantic digital library, in particular the use of FOAF to structure name authority files. A similar approach is demonstrated by Kruk et al.[32]. Their semantic digital library (‘JeromeDL’[33]) uses FOAF to manage an authority file of authors, editors and publishers, but also uses FOAF to connect users and manage user profiles within their system[34]. JeromeDL deploys FOAFRealm[35], a FOAF-based technology developed by members of the same research team, to establish user identities[36]. FOAF is also used to offer novel resource discovery mechanisms described as ‘social semantic collaborative filtering’[37]. For example, two colleagues will often share similar academic interests such that one might be able to find resources relevant to their information need within the profile of the other (e.g. resources held within virtual bookshelves, bookmarks, etc.).

Even less formal tools, such as those optimized for personal information management, increasingly deploy FOAF. BibSonomy[38], the social bookmark and publication management tool, exposes user profiles and interests via publicly available FOAF files, each providing personal information and subject interests which can be discovered by Semantic Web applications wishing to reuse bookmarks or publications stored and tagged by users. BibSonomy also exposes bookmarks in a variety of formats, including RDF/XML, XML, RSS and BibTeX.

Although RDF is optimized for machine processing, an increasing number of freely available tools can be used to explore FOAF files on the web[39, 40, 41]. Browser plug-ins for Mozilla Firefox[42] are also available[43], enabling the automatic extraction of FOAF data (and other RDF data) from web pages and their interrogation using a number of technologies.

Simple Knowledge Organization System (SKOS)

It was noted earlier that an important aim of the Semantic Web is to improve information retrieval and information organization on the web. SKOS[44] is an application of RDF designed to provide a data model for Knowledge Organization Systems (KOS) and is currently under active development by the W3C Semantic Web Deployment Working Group[45]. KOS – also referred to as controlled vocabularies or terminologies, and as 'concept schemes' by the SKOS specification – includes tools such as information retrieval thesauri, taxonomies, classification schemes, subject heading lists, and other forms of authority list or knowledge structure. It is therefore immediately understandable why SKOS will contribute to improvements in resource discovery, and practical examples of this will be discussed later.

SKOS is primarily designed to enable the publication of controlled vocabularies for use in the Semantic Web, thus enabling their machine interpretation to facilitate the retrieval and organization of resources. SKOS also enables KOS interoperability, data sharing, linking and data merging. The ability to merge and link SKOS with other data sources is consistent with RDF generally and enables SKOS data to be linked or merged by Semantic Web applications with other controlled vocabularies or subject indexes. This can be useful for a number of reasons, but particularly in retrieval circumstances where multiple collections have to be queried as it avoids the need for complex database integration[46].

An important Semantic Web specification in the area of knowledge modelling and representation is the W3C Web Ontology Language (OWL)[47]. Discussion of OWL can be complex and is therefore outside the scope of this chapter. Nevertheless, OWL assumes an important role in enabling intelligent software agents to infer and reason over knowledge captured in ontologies[48]; however, it is generally acknowledged that OWL is insufficient to fulfil the Semantic Web vision on its own and the "construction of detailed 'maps' of particular domains of knowledge"[49] are necessary, along with metadata. SKOS is therefore about harnessing LIS expertise in the area of knowledge organization to create these 'maps'. The large number of well-developed vocabularies already in use and under continual revision are well suited to achieving this. Additionally, SKOS enables the easy creation and publication of new vocabularies to fulfil emerging knowledge domains.

SKOS is very flexible and can accommodate most forms of KOS, with special provisions made for modelling arrays, notation and other features peculiar to controlled vocabularies. SKOS essentially consists of a series of classes and properties to express the structural characteristics of KOS. For example, a thesaurus would be a `skos:ConceptScheme` containing a series of `skos:Concepts`, each of which might have properties such as `skos:broader`, `skos:narrower`, `skos:related` and `skos:altLabel` (i.e. BT, NT, RT and UF respectively). Consider the following example taken from the *UNESCO* Thesaurus[50] for the concept, 'Information scientists':

Information scientists

SN A person who works on the theory or application of informatics or information science, i.e. analyses, designs, implements, etc. information systems

UF Information officers
BT Information/library personnel
RT Archive personnel
RT Information science education

Such a thesaurus concept could be expressed in SKOS RDF/XML as follows:

```
<?xml version=”1.0” encoding=”UTF-8”?>

<rdf:RDF xmlns:rdf=”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
xmlns:skos=”http://www.w3.org/2004/02/skos/core#”>

<skos:Concept rdf:about=”http://.../mt5.20/Informationscientists#concept”>
 <skos:prefLabel xml:lang=”en”>Information scientists</skos:prefLabel>
 <skos:scopeNote xml:lang=”en”>A person who works on the theory or application of
informatics or information science, i.e. analyses, designs, implements etc. information
systems.</skos:scopeNote>
 <skos:altLabel xml:lang=”en”>Information officers</skos:altLabel>
 <skos:broader rdf:resource=”http://.../mt5.20/Informationlibrarypersonnel#concept”/>
 <skos:related rdf:resource=”http://.../mt5.20/Archivepersonnel#concept”/>
 <skos:related rdf:resource=”http://.../mt1.50/Informationscienceeducation#concept”/>
</skos:Concept>

</rdf:RDF>
```

The above example produces the RDF graph given in Figure 7. Note that URIs have been used to identify the concepts within the KOS. At time of writing, the *UNESCO Thesaurus* remains unpublished for the Semantic Web so the URIs in the above example are merely illustrative. Increasingly, vocabularies published in SKOS infer their structure or use their notation within URI. The micro-thesaurus notation from the *UNESCO Thesaurus* has therefore been incorporated into the URI. This approach to ‘minting’ URIs is consistent with the ‘Cool URI’ trend within the Semantic Web community[51]; an attempt to maintain the purpose of a URI in uniquely identifying resources (in their various permutations) whilst simultaneously making them more meaningful than simply a random sequence of characters. The significance of minting Cool URIs has recently attracted wider discussion and research by SKOS researchers. For example, Panzer discusses the minting of URIs for publishing DDC for the Semantic Web[52], whilst Summers et al. discuss URIs in their conversion of LCSH from MARCXML to SKOS[53].

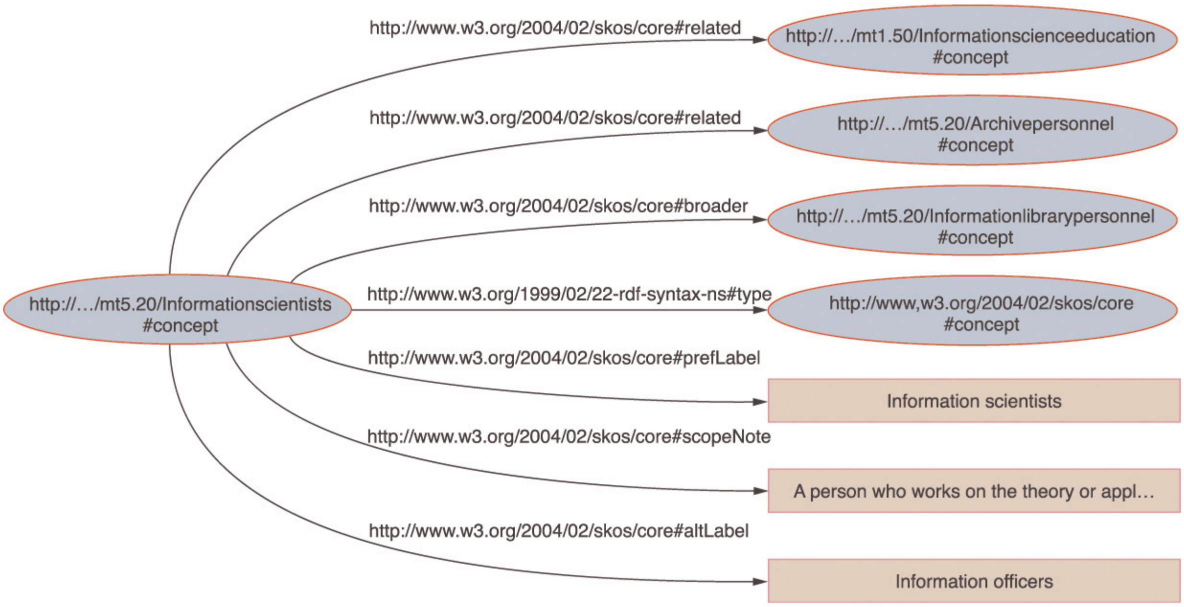

http://www.w3.org/2004/02/skos/core#related
http://.../mt1.50/Informationscienceeducation #concept
http://www.w3.org/2004/02/skos/core#related
http://.../mt5.20/Archivepersonnel #concept
http://www.w3.org/2004/02/skos/core#broader
http://.../mt5.20/Informationlibrarypersonnel #concept
http://.../mt5.20/Informationscientists #concept
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www,w3.org/2004/02/skos/core #concept
http://www.w3.org/2004/02/skos/core#prefLabel
Information scientists
http://www.w3.org/2004/02/skos/core#scopeNote
A person who works on the theory or appl…
http://www.w3.org/2004/02/skos/core#altLabel
Information officers

Figure 7. RDF graph for the UNESCO Thesaurus concept (skos:Concept), ‘Information scientists’

Recall that in our DC RDF/XML example (Figure 5), the subject of our resource (`dcterms:subject)` was indicated by the LCSH descriptor, 'Semantic Web'; however, rather than identify this descriptor by using a literal we elected to identify the concept by URI (`http://id.loc.gov/authorities/sh2002000569)`. This 'concept URI' not only defines the concept of the 'Semantic Web' and the preferred lexical label, but points to rich terminological data (e.g. the PT, BT, RT, SN, etc.) expressed in SKOS by the Library of Congress Authorities & Vocabularies service[54], thus enabling information retrieval which is less dependent on free-text searching and more concerned with the representation of *concepts*. Indeed, it is possible for concept definitions (i.e. URIs) to be reused with alternative lexical labels. This ethos is central to SKOS (and the Semantic Web generally) and forms part of the 'linked data' principle[55, 56]: exposing and reusing RDF data and URIs to maximize data connections and relationships in a manner which is useful to both humans and machines. In essence then, linked data is about creating connections between data which previously may not have existed and exposing this data for sharing on the Semantic Web by using URIs and RDF. Tim Berners-Lee has noted that linked data is essential to connect the components of the Semantic Web[57]. The more connections there are between data, the greater the value and usefulness of that data, thus allowing humans and machines to follow semantic threads across disparate data sources (using URIs). The linked data approach holds great potential for SKOS as it allows "concepts from different concept schemes [to be] connected together […] to form a distributed, heterogeneous global concept scheme. A web of concept schemes can serve as the foundation for new applications that allow meaningful navigation between KOSs"[58].

More generally, the use of SKOS makes it easier to design distributed information retrieval systems because the identification of concepts is based on concept URIs and structured according to KOS rules in RDF. For example, upon retrieving a resource via subject searching, a system could be designed to retrieve other resources on the Semantic Web identified in the same way, thus improving recall whilst maintaining a level of precision. This can be a particularly useful mechanism given the distributed and decentralized nature of resource publication on the web. Since a concept URI links to a detailed description of the concept (e.g. its preferred label, BT, NT, RT, etc.), it is also possible to reuse this data to provide extra retrieval aids for the user. For example, broader and related terms could be used to deliver query expansion search techniques[59], or the terms could be displayed to assist the user in refining their search query, perhaps allowing the user to browse the KOS hierarchically. Visual search interfaces could be created showing the relationships between concepts (e.g. based on the RDF graph), for example see the Library of Congress Authorities & Vocabularies[60]. Software could also be designed to enable users to browse concept schemes and retrieve resources identified using its concept URIs.

Some of the aforementioned techniques have been demonstrated by the Explicator project[61]. Gray et al. demonstrate a web service for searching and exploring concepts within SKOS-encoded astronomical vocabularies[62]. Their 'Vocabulary Explorer' web application enables users to traverse astronomical concepts and formal scientific definitions, their relationships to other concepts, and their relationships with similar concepts in alternative vocabularies. Further work undertaken by the same research team demonstrates how rich semantic relationships within SKOS can be exploited to improve retrieval precision and deliver a variety of searching aids for users[63].

Another interesting feature of SKOS is its ability to capture mappings between concepts in different concept schemes. This can be useful where problems of semantic heterogeneity exist (i.e. a collection is using more than one vocabulary to index resources). To accommodate such scenarios, SKOS provides properties such as `skos:closeMatch`, `skos:exactMatch`, `skos:broadMatch`, etc. These properties can be used to state a conceptual link between SKOS concepts in different concept schemes, thus ameliorating the vocabulary mis-match difficulties which often arise in distributed contexts, or where several heterogeneous collections are merged. For example, Isaac et al. report on the use of SKOS to resolve semantic heterogeneity within digitized cultural heritage collections[64]. They use their methods of 'semantic alignment' to create mappings between different concept schemes, thereby providing users with integrated access to resources which have been indexed using a number of different vocabularies.

An increased need to deliver KOS data (with mappings) in a web services context has emerged in recent years. Such web services are considered necessary to effect improvements in digital library searching functionality and/or to offer users the option of searching multiple third-party repositories indexed using disparate vocabularies. Use of SKOS within a web services context has unsurprisingly attracted attention. For example, the STAR project[65] has created a series of pilot Semantic Web services for KOS data based on SKOS, providing term look-up functionality, browsing and semantic concept expansion[66]. Macgregor et al.[67] demonstrate the use of SKOS in a web services context as part of the High-Level Thesaurus (HILT) project[68]. Their 'terminology mapping server' uses SKOS to structure terminological data (including mappings via a DDC spine) when responding to SRW/U requests from digital libraries. Similar work is also being conducted by the Deutsche Nationalbibliothek[69].

Of course, almost all of the aforementioned is entirely dependent upon KOS being published for the Semantic Web in SKOS. Although SKOS is currently a W3C 'candidate recommendation', several well-known vocabularies have already been made officially available in SKOS for use on the Semantic Web, such as LCSH[70], *STW Thesaurus for Economics*[71], AGROVOC[72], and GEMET[73]. Many others have been temporarily published in SKOS, but these lack provenance and stability owing to their use within research experiments.

RDFa

RDF specifications such as FOAF, SKOS, OWL, and even Dublin Core RDF, necessitate understanding of the underlying RDF data model, as well as knowledge of the various RDF serializations. Such applications of RDF are typically made available independently of the resource(s) they are describing or associated with (i.e. as a separate file).

More recently the W3C has introduced RDFa (Resource Description Framework in attributes)[74]. RDFa provides a series of XHTML[75] extensions which can be used to annotate web pages with semantic data. As the official RDFaWiki[76] and RDFa Primer[77] indicate, RDFa is a simple way of embedding RDF statements within XHTML and an attempt to encourage publishers, bloggers, web developers and the like to participate in the development of the Semantic Web. RDFa enables simple semantic data to be encoded without detailed knowledge of RDF or the need for separate RDF files containing detailed RDF/XML or other RDF serializations. In fact, knowledge of XHTML is the only prerequisite to deploying RDFa in practice, although more detailed applications of RDFa would obviously benefit from a wider knowledge of RDF.

Consider the following snippet of 'vanilla' XHTML. This example represents what typical XHTML might look like in a fictional web page publishing this chapter at the MetaPress domain (`http://uksg.metapress.com/someURL`):

```
<!DOCTYPE html PUBLIC “-//W3C//DTD XHTML 1.1//EN”
“http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd”>
<html xmlns=”http://www.w3.org/1999/xhtml” xml:lang=”en”>

<head>
<title>E-Resource management and the Semantic Web: applications for RDF for e-resource
discovery</title>
<meta http-equiv=”Content-Type” content=”text/html; charset=iso-8859-1”/>
</head>

<body>
<h1>E-Resource management and the Semantic Web: applications for RDF for e-resource
discovery</h1>
<p>George Macgregor</p>
<p>16 April 2009</p>
<p>Keywords: Semantic Web, digital libraries</p>

<h2>Abstract</h2>
<p>Semantic Web technologies and specifications are increasingly finding applications
```

```
within digital libraries and other e-resource contexts. The purpose of this chapter is to
... within a variety of e-resource discovery contexts will be discussed.</p>
<h2>About the author</h2>
<p>George Macgregor is currently a Lecturer in Information Management and a <a
href=”http://www.ljmu.ac.uk/LBS/92624.htm”>member</a> of the Information Strategy Group at
<a href=”http://www.ljmu.ac.uk/LBS/Index.htm”>Liverpool Business School</a>, <a
href=”http://www.ljmu.ac.uk/”>Liverpool John Moores University</a>. George helps maintain
the <a href=”http://ljmuinfostrategy.blogspot.com/”>Information Strategy Group
blog.</a></p>

</body>
</html>
```

The above example is an instance of how the web has evolved to provide a series of documents conducive to human interpretation, but has failed to capture the semantics of these documents for machine interpretation. This web page does little to assist machines in interpreting who the creator of the chapter is, or even what its title is. Humans know who the creator is and what the title is, but only because this is loosely inferred by the page structure when the file is viewed in a web browser.

To embed semantics we could use RDFa to annotate the XHTML (XHTML+RDFa) by embedding the necessary RDF triples. For example, we could annotate the previous example by extending the XHTML to include Dublin Core and FOAF. The relevant RDFa extensions are visible in bold font:

```
<html
  xmlns=”http://www.w3.org/1999/xhtml” version=”XHTML+RDFa 1.0”
  xml:lang=”en”
  xmlns:dcterms=http://purl.org/dc/elements/1.1/
  xmlns:foaf=”http://xmlns.com/foaf/0.1/”>

<head>
<title>E-Resource management and the Semantic Web: applications for RDF for e-resource
discovery</title>
<base href=”http://uksg.metapress.com/someURL” />
</head>

<body>
<h1 property=”dcterms:title”>E-Resource management and the Semantic Web: applications for
RDF for e-resource discovery</h1>

<p><span rel=”dcterms:creator”><span about=”http://www.staff.ljmu.ac.uk/bsngmacg/#me”
typeof=”foaf:Person”>George Macgregor</span></span></p>
<p><span property=”dcterms:date” content=”2009-04-16”>16 April 2009</span></p>
<p>Keywords: <span rel=”dcterms:subject”
resource=”http://id.loc.gov/authorities/sh2002000569”>Semantic Web</span>, <span
rel=”dcterms:subject” resource=”http://id.loc.gov/authorities/sh95008857”>digital
libraries</span></p>

<h2>Abstract</h2>
<p property=”dcterms:abstract”>Semantic Web technologies and specifications are
increasingly finding applications within digital libraries and other e-resource contexts.
The purpose of this chapter is to ... within a variety of e-resource discovery contexts
will be discussed.</p>

<h2>About the author</h2>
<p about=”http://www.staff.ljmu.ac.uk/bsngmacg/#me” typeof=”foaf:Person”><span
property=”foaf:name”>George Macgregor</span> is currently a Lecturer in Information
Management and a <a rel=”foaf:workPlaceHomePage”
```

```
href=”http://www.ljmu.ac.uk/LBS/92624.htm”>member</a> of the Information Strategy Group at
<a rel=”foaf:workInfoHomePage” href=”http://www.ljmu.ac.uk/LBS/Index.htm”><span
property=”dc:title”>Liverpool Business School</span></a>, <a
href=”http://www.ljmu.ac.uk/”><span property=”dc:title”>Liverpool John Moores
University</span></a>. George helps maintain the <a rel=”foaf:weblog”
href=”http://ljmuinfostrategy.blogspot.com/”><span property=”dc:title”>Information Strategy
Group blog</span>.</a> </p>

</body>
</html>
```

One of the advantages of XHTML+RDFa is that it allows semantics to be embedded within running text. This is clearly demonstrated in the paragraph providing biographical information about the author. `foaf:Person` has been used to identify the author and other FOAF and Dublin Core properties have been used.

RDFa Distiller[78] is a W3C tool for ‘scraping’ RDF triples from XHTML+RDFa web pages and for outputting them in standalone RDF serializations (e.g. RDF/XML). By using RDFa Distiller on the above XHTML+RDFa we can observe in the example below that the relevant triples have been extracted and structured in RDF/XML, and in a manner not dissimilar to examples earlier in this chapter. This example generates the RDF graph in Figure 8 and is easier to decipher:

```
<?xml version=”1.0” encoding=”utf-8”?>
<rdf:RDF
 xmlns:dcterms=”http://purl.org/dc/elements/1.1/”
 xmlns:foaf=”http://xmlns.com/foaf/0.1/”
 xmlns:rdf=”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
 xmlns:rdfs=”http://www.w3.org/2000/01/rdf-schema#”
 xmlns:xhv=”http://www.w3.org/1999/xhtml/vocab#”
 xmlns:xml=”http://www.w3.org/XML/1998/namespace”
>
 <rdf:Description rdf:about=”http://uksg.metapress.com/someURL”>
  <dcterms:abstract xml:lang=”en”>Semantic Web technologies and specifications are
increasingly finding applications within digital libraries and other e-resource contexts.
The purpose of this chapter is to ... within a variety of e-resource discovery contexts
will be discussed.</dcterms:abstract>
  <dcterms:creator>
   <foaf:Person rdf:about=”http://www.staff.ljmu.ac.uk/bsngmacg/#me”>
    <foaf:workPlaceHomePage rdf:resource=”http://www.ljmu.ac.uk/LBS/92624.htm”/>
    <foaf:workInfoHomePage rdf:resource=”http://www.ljmu.ac.uk/LBS/Index.htm”/>
    <foaf:name xml:lang=”en”>George Macgregor</foaf:name>
    <foaf:weblog rdf:resource=”http://ljmuinfostrategy.blogspot.com/”/>
   </foaf:Person>
  </dcterms:creator>
  <dcterms:subject rdf:resource=”http://id.loc.gov/authorities/sh2002000569”/>
  <dcterms:subject rdf:resource=”http://id.loc.gov/authorities/sh95008857”/>
  <dcterms:date xml:lang=”en”>2009-04-16</dcterms:date>
  <dcterms:title xml:lang=”en”>E-Resource management and the Semantic Web: applications for
RDF for e-resource discovery</dcterms:title>
 </rdf:Description>
</rdf:RDF>
```

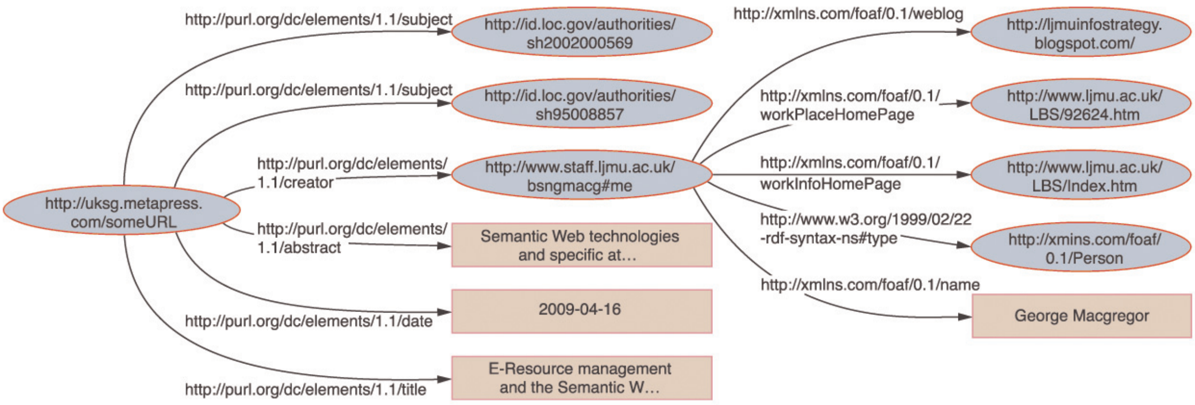

http://uksg.metapress.com/someURL
http://purl.org/dc/elements/1.1/subject
http://id.loc.gov/authorities/sh2002000569
http://purl.org/dc/elements/1.1/subject
http://id.loc.gov/authorities/sh95008857
http://purl.org/dc/elements/1.1/creator
http://www.staff.ljmu.ac.uk/bsngmacg#me
http://purl.org/dc/elements/1.1/abstract
Semantic Web technologies and specific at...
2009-04-16
http://purl.org/dc/elements/1.1/date
E-Resource management and the Semantic W...
http://purl.org/dc/elements/1.1/title
http://xmlns.com/foaf/0.1/weblog
http://ljmuinfostrategy.blogspot.com/
http://xmlns.com/foaf/0.1/workPlaceHomePage
http://www.ljmu.ac.uk/LBS/92624.htm
http://xmlns.com/foaf/0.1/workInfoHomePage
http://www.ljmu.ac.uk/LBS/Index.htm
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://xmins.com/foaf/0.1/Person
http://xmlns.com/foaf/0.1/name
George Macgregor

Figure 8. RDF graph of fictional XHTML+RDFa, based on triples extracted using RDFa Distiller

RDFa remains a relatively new Semantic Web standard and only received W3C recommendation status in late 2008. Implementations have therefore been predominantly confined to those offered by the W3C. Nevertheless, large scale implementations within digital libraries are already visible. Neubert[79] describes the publication of the *STW Thesaurus for Economics*[80] for the Semantic Web. *STW* is a richly interconnected multilingual thesaurus (English and German) accommodating subjects within the economics and business-related disciplines. It provides 'topical entry points' to the German National Library of Economics (ZBW)[81] digital library and aims to provide an economics and business hub within the web of linked data. *STW* is delivered as XHTML+RDFa pages (using Dublin Core, SKOS, OWL and others), with searching and concept tree browsing functionality offered in the interface. A standalone SKOS RDF/XML dump version can also be downloaded.

Conclusion

This chapter has attempted to introduce the key Semantic Web concepts and its principal enabling language using a series of practical examples. As we have noted, applications of RDF, such as Dublin Core, FOAF and SKOS, have clear applications within e-resource discovery contexts and their increased deployment can effect improvements in information retrieval and enable the delivery of other information tools for users. They also enable a level of improved data sharing, linking, merging and interoperability which can enrich the structured data already managed by digital libraries, thus contributing to the web of 'linked data' and better exposing invaluable e-resources. The benefits of interacting, contributing and maintaining the structured data required to support the Semantic Web have been recognized by information professionals and the increased deployment of Semantic Web techniques within digital libraries has proliferated. Fulfilling the vision of the Semantic Web for those outside the information profession is an immense task owing to the lack of structured data available with which to work. It is therefore appropriate that digital libraries and repositories assume an increased responsibility in bringing the Semantic Web vision to fruition.

References

1. Ed. Greenberg, J and Méndez, E, *Knitting the Semantic Web*, 2007, Binghamton, The Haworth Press.
2. Ed. Macgregor, G, Digital libraries and the Semantic Web: context, applications and research, *Library Review (Special issue)*, 2008, 57(3).
3. International Conferences on Digital Libraries and the Semantic Web (ICDC):
http://www.icsd-conference.org/ (Accessed 16 April 2009)
4. Ed. Miles, A and Bechhofer, S, *SKOS Simple Knowledge Organization System Reference. W3C Candidate Recommendation 17 March 2009*, 2009, Cambridge, MA, W3C.
http://www.w3.org/TR/2009/CR-skos-reference-20090317/ (Accessed 16 April 2009)

5. Berners-Lee, T, Hendler, J and Lassila, O, The Semantic Web, *Scientific American*, 2001, May.
http://www.sciam.com/article.cfm?id=the-semantic-web (Accessed 16 April 2009)

6. Shadbolt, N, Berners-Lee, T and Hall, W, The Semantic Web Revisited, *IEEE Intelligent Systems*, 2006, 21(3), 96–101.

7. Guha, R, McCool, R and Miller, E, Semantic search. In: *Proceedings of the 12th International Conference on the World Wide Web, 20–24 May, Budapest, Hungary*, Ed. Hencsey, G and White, B, 2003, New York, ACM, 700–709.

8. Berners-Lee, T et al., ref. 5.

9. Ed. Smith, M K, Welty, C and McGuinness, D L, *OWL Web Ontology Language Guide*, 2004, MA, W3C.
http://www.w3.org/TR/owl-guide/ (Accessed 16 April 2009)

10. Berners-Lee, T, Q&A with Tim Berners-Lee, *Business Week*, 2007, 9 April.
http://www.businessweek.com/technology/content/apr2007/tc20070409_961951.htm (Accessed 16 April 2009)

11. W3C Semantic Web Activity:
http://www.w3.org/2001/sw/ (Accessed 3 August 2009)

12. Ed. Klyne, G and Carroll, J J, *Resource Description Framework (RDF): concepts and abstract syntax*, 2004, Cambridge, MA, W3C.
http://www.w3.org/TR/rdf-concepts/ (Accessed 16 April 2009)

13. Thompson, H S, *What's a URI and why does it matter?*, 2008, Edinburgh, University of Edinburgh / W3C.
http://www.ltg.ed.ac.uk/~ht/WhatAreURIs/ (Accessed 16 April 2009)

14. Ed. Beckett, D, *RDF/XML syntax specification (revised). W3C recommendation 10 February 2004*, 2004, Cambridge, MA, W3C.
http://www.w3.org/TR/rdf-syntax-grammar/ (Accessed 16 April 2009)

15. Berners-Lee, T, *Primer: Getting into RDF & Semantic Web using N3*, 2005, MA, W3C.
http://www.w3.org/2000/10/swap/Primer.html (Accessed 16 April 2009)

16. Ed. Beckett, D and Berners-Lee, T, *Turtle - Terse RDF Triple Language*, 2008, MA, W3C.
http://www.w3.org/TeamSubmission/turtle/ (Accessed 16 April 2009)

17. Ed. Bray, T, Paoli, J, Sperberg-McQueen, C M, Maler, E and Yergeau F, *Extensible Markup Language (XML) 1.0 (Fifth edition)*, 2008, Cambridge, MA, W3C.
http://www.w3.org/TR/2008/REC-xml-20081126/ (Accessed 16 April 2009)

18. W3C Semantic Web Activity, ref. 11.

19. Dublin Core Metadata Initiative:
http://dublincore.org/ (Accessed 16 April 2009)

20. Nilsson, M, Powell, A, Johnston, P and Naeve, A, *Expressing Dublin Core metadata using the Resource Description Framework (RDF)*, 2008, Seoul, Dublin Core Metadata Initiative.
http://www.dublincore.org/documents/dc-rdf/ (Accessed 16 April 2009)

21. Creative Commons:
http://creativecommons.org/ (Accessed 16 April 2009)

22. Authorities & Vocabularies (Library of Congress):
http://id.loc.gov/authorities/ (Accessed 30 April 2009)

23. Morla:
http://www.morlardf.net/ (Accessed 16 April 2009)

24. Altova Semantic Works 2009:
http://www.altova.com/products/semanticworks/semantic_web_rdf_owl_editor.html (Accessed 16 April 2009)

25. W3C RDF Validation Service:
http://www.w3.org/RDF/Validator/ (Accessed 16 April 2009)

26. The FOAF project:
http://www.foaf-project.org/ (Accessed 16 April 2009)

27. Graves, M, Constabaris, A and Brickley, D, FOAF: Connecting people on the Semantic Web, *Cataloging & Classification Quarterly*, 2007, 43 (3/4), 191–202.

28. Brickley, D and Miller, L, *FOAF Vocabulary Specification 0.91: Namespace Document 2 November 2007 – OpenID Edition*, 2007.
http://xmlns.com/foaf/spec/ (Accessed 16 April 2009)

29. Ding, L, Zhou, L, Finin, T and Joshi, A, How the Semantic Web is being used: an analysis of FOAF documents. In: *Proceedings of the 38th Annual Hawaii International Conferences on System Sciences (HICSS'05), 3–6 January, Hawaii, USA*, 2005, IEEE Computer Society, USA.
http://ebiquity.umbc.edu/_file_directory_/papers/120.pdf (Accessed 16 April 2009)

30. Golbeck, J and Rothstein, M, Linking social networks on the web with FOAF: A Semantic Web case study. In: *Proceedings of the Twenty-third AAAI Conference on Artificial Intelligence National Conference on Artificial Intelligence, 13–17 July, Chicago, USA*, 2008, California, AAAI Press, 1138–1143.
http://www.cs.umd.edu/~golbeck/downloads/foaf.pdf (Accessed 16 April 2009)

31. Malmsten, M, Making a library catalogue part of the Semantic Web. In: *Proceedings of the International Conference on Dublin Core and Metadata Applications: Metadata for semantic and social applications, 22–26 September 2008, Berlin, Germany*, Ed. Greenberg, J and Wolfgang, K, 2008, Go_ttingen, Go_ttingen Univ.-Verl.
http://dcpapers.dublincore.org/ojs/pubs/article/view/927/923 (Accessed 16 April 2009)

32. Kruk, S R, Woroniecki, T, Gzella, A and Dπbrowski, M, JeromeDL – a semantic digital library. In: *Proceedings of the Semantic Web Challenge Workshop, 6th International Semantic Web Conference, 11–15 November 2007, Busan, South Korea*, 2007.
http://library.deri.ie/resource/81n2JzC5 (Accessed 16 April 2009)

33. JeromeDL:
http://www.jeromedl.org/ (Accessed 16 April 2009)

34. Kruk, S R, Zimmerman, K and Sapkota, B, Semantically enhanced search services in digital libraries. In: *Proceedings of the International Conference on Internet and Web Applications and Services (ICIW 2006), 23–25 February, Guadeloupe, French Caribbean*, 2006.
http://library.deri.ie/resource/iwqOhveX (Accessed 16 April 2009)

35. FOAFRealm:
http://www.foafrealm.org/ (Accessed 16 April 2009)

36. Kruk, S R, et al, ref. 34.

37. Kruk, S R, Decker, S, Gzella, A and Grzonkowski, S, Social Semantic Collaborative Filtering, *Journal of Web Semantics*, 2008 (pre-print).
http://library.deri.ie/resource/790k6n8a (Accessed 16 April 2009)

38. BibSonomy:
http://www.bibsonomy.org/ (Accessed 16 April 2009)

39. FOAF Explorer:
http://xml.mfd-consult.dk/foaf/explorer/ (Accessed 16 April 2009)

40. Semantically Interlinked Online Communities (SIOC) Browser:
http://sparql.captsolo.net/browser/browser.py (Accessed 16 April 2009)

41. FOAF ('friend-of-a friend') Database:
http://swordfish.rdfweb.org/rweb/who (Accessed 16 April 2009)

42. Mozilla Firefox:
http://www.mozilla-europe.org/en/firefox/ (Accessed 16 April 2009)

43. Semantic Radar for Firefox:
http://sioc-project.org/firefox

44. Ed. Miles, A and Bechhofer, S, ref. 4.

45. W3C Semantic Web Deployment Working Group:
http://www.w3.org/2006/07/SWD/ (Accessed 16 April 2009)

46. Miles, A and Pérez-Agüera, J R, SKOS: Simple Knowledge Organisation for the Web, *Cataloging & Classification Quarterly*, 2007, 43(3/4), 69–83.

47. Ed. Dean, M and Schreiber, G, *OWL Web Ontology Language Reference*, 2004, Cambridge, MA, W3C.
http://www.w3.org/TR/owl-ref/ (Accessed 16 April 2009)

48. Shadbolt, N, et al, ref. 6.

49. Ed. Miles, A and Bechhofer, S, ref. 4.

50. UNESCO Thesaurus:
http://databases.unesco.org/thesaurus/ (Accessed 16 April 2009)

51. Ed. Sauerman, L and Cyganiak, R, *Cool URIs for the Semantic Web - W3C Working Draft, 17 December 2007*, 2007, Cambridge, MA, W3C.
http://www.w3.org/TR/2007/WD-cooluris-20071217/ (Accessed 16 April 2009)

52. Panzer, M, Cool URIs for the DDC: Towards Web-scale accessibility of a large classification system. In: *Proceedings of the International Conference on Dublin Core and Metadata Applications: Metadata for semantic and social applications, 22-26 September 2008, Berlin, Germany*, Ed. Greenberg, J and Wolfgang, K, 2008, Göttingen, Göttingen Univ.-Verl.
http://dcpapers.dublincore.org/ojs/pubs/article/viewFile/932/928 (Accessed 16 April 2009)

53. Summers, E, Isaac, A, Redding, C and Krech, D, LCSH, SKOS and linked data. In: *Proceedings of the International Conference on Dublin Core and Metadata Applications: Metadata for semantic and social applications, 22–26 September 2008, Berlin, Germany*, Ed. Greenberg, J and Wolfgang, K, 2008, Göttingen, Göttingen Univ.-Verl.
http://dcpapers.dublincore.org/ojs/pubs/article/view/916/912 (Accessed 16 April 2009)

54. Authorities & Vocabularies, ref. 22.

55. Berners-Lee, T, linked data, 2007, Cambridge, MA, W3C.
http://www.w3.org/DesignIssues/LinkedData.html (Accessed 16 April 2009)

56. Linked Data - Connect Distributed Data across the Web:
http://linkeddata.org/ (Accessed 16 April 2009)

57. Berners-Lee, T, ref. 55.

58. Ed. Miles, A and Bechhofer, S, ref. 4.

59. Efthimiadis, E N, Query expansion, *Annual Review of Information Systems and Technology (ARIST)*, 31, 121–187.
http://faculty.washington.edu/efthimis/pubs/Pubs/qe-arist/QE-arist.html (Accessed 16 April 2009)

60. Authorities & Vocabularies, ref. 22.

61. Explicator project:
http://explicator.dcs.gla.ac.uk/ (Accessed 16 April 2009)

62. Gray, A J G, Gray, N and Ounis, I, Searching and exploring controlled vocabularies. In: *Proceedings of Exploiting Semantic Annotations in Information Retrieval (ESAIR 2009), 09 February, Barcelona, Spain*, 2009, New York, ACM, 1–5.
http://www.zaragozas.info/esair/final_6.pdf (Accessed 16 April 2009)

63. Gray, A J G, Gray, N and Ounis, I, Finding data resources in a virtual observatory using SKOS vocabularies. In: *Proceedings of 25th British national conference on databases: sharing data, information and knowledge, 07–10 July, Cardiff, UK*, 2008, Berlin, Springer-Verlag, 189–192.

64. Isaac, A, Schlobach, S, Matthezing, H and Zinn, C, Integrated access to cultural heritage resources through representation and alignment of controlled vocabularies, *Library Review*, 2008, 57(3), 187–199.

65. Semantic Technologies for Archaeological Resources project (STAR):
http://hypermedia.research.glam.ac.uk/kos/STAR/ (Accessed 16 April 2009)

66. Tudhope, D and Binding, C, Experiences with Knowledge Organization System services from the STAR Project, *Signum*, 2008, 5.
http://pro.tsv.fi/STKS/signum/200805/3.pdf (Accessed 16 April 2009)

67. Macgregor, G, McCulloch, E and Nicholson, D, Terminology server for improved resource discovery: analysis of functions and model. In: *Proceedings of the 2nd International Conference on Metadata and Semantics Research, 11–12*

October 2007, Ionian Academy, Corfu, Greece, Ed. Miguel-Angel, S and Lytras, M D, 2007, Corfu, Ionian Academy. http://www.mtsr.ionio.gr/proceedings/mcculloch.pdf (Accessed 16 April 2009)

68. High-Level Thesaurus (HILT) project:
http://hilt.cdlr.strath.ac.uk/ (Accessed 16 April 2009)
69. Svennson, L G, National libraries and the Semantic Web: requirements and applications. In: *Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007), 21–23 February 2007, Bangalore, India*, Ed. Prasad, A R D and Madalli, D P, 2007, Bangalore, DRTC, 101–108.
70. Authorities & Vocabularies, ref. 22.
71. *STW Thesaurus for Economics*:
http://zbw.eu/stw/versions/8.04/about.en.html (Accessed 16 April 2009)
72. AGROVOC (SKOS):
http://www.fao.org/aims/ag_download.htm (Accessed 16 April 2009)
73. GEneral Multilingual Environmental Thesaurus (GEMET):
http://www.eionet.europa.eu/gemet/webservices?langcode=en (Accessed 16 April 2009)
74. Ed. Adida, B, Birbeck, M, McCarron, S and Pemberton, S, *RDFa in XHTML: syntax and processing: A collection of attributes and processing rules for extending XHTML to support RDF. W3C recommendation 14 October 2008*, 2008, Cambridge, MA, W3C.
http://www.w3.org/TR/rdfa-syntax/ (Accessed 16 April 2009)
75. Ed. Axelsson, J, Birbek, M, Dubinko, M, Epperson, B, Ishikawa, M, McCarron, S, Navarro, A and Perberton, S, *XHTML 2.0*, 2006, Cambridge, MA, W3C.
http://www.w3.org/TR/xhtml2/ (Accessed 16 April 2009)
76. RDFaWiki:
http://rdfa.info/wiki/RDFa_Wiki (Accessed 16 April 2009)
77. Ed. Adida, B and Birbeck, M, *RDFa primer: Bridging the human and data webs*, 2008, Cambridge, MA, W3C.
http://www.w3.org/TR/xhtml-rdfa-primer/ (Accessed 16 April 2009)
78. RDFa Distiller:
http://www.w3.org/2007/08/pyRdfa/ (Accessed 16 April 2009)
79. Neubert, J, Bringing the 'Thesaurus for Economics' on to the Web of linked data. In: *Proceedings of linked data on the Web 2009 (LDOW2009), 20 April 2009, Madrid, Spain*, Ed. Bizer, C, Heath, T, Berners-Lee, T and Idehen, K, 2009, CEUR-WS.
http://events.linkeddata.org/ldow2009/papers/ldow2009_paper7.pdf (Accessed 16 April 2009)
80. *STW Thesaurus for Economics*, ref. 71.
81. German National Library of Economics (ZBW):
http://www.zbw-kiel.de/index-e.html (Accessed 16 April 2009)

■ **George Macgregor**
Information Strategy Group
Information Management & Systems
Liverpool Business School
Liverpool John Moores University, John Foster Building
98 Mount Pleasant, Liverpool L3 5UZ, UK.
E-mail: g.r.macgregor@ljmu.ac.uk

Biographical note

George Macgregor is a lecturer in Information Management and a member of the Information Strategy Group based in the Information Management & Systems section of Liverpool Business School, Liverpool John Moores University. Prior to this George was a Research Fellow based in the Centre for Digital Library Research at the University of Strathclyde. His research interests lie in the areas of distributed digital libraries, metadata, networked Knowledge Organization Systems (KOS), and the use of Semantic Web technologies within digital libraries and repositories.

A SKOS Core approach to implementing an M2M terminology mapping server

George Macgregor, Anu Joseph and Dennis Nicholson

Centre for Digital Library Research, Department of Computer & Information Sciences, University of Strathclyde, Glasgow, UK
{george.macgregor, anu.joseph, dennis.nicholson}
@cis.strath.ac.uk

Abstract: The proliferation of distributed digital libraries and repositories has increased the need for improved interoperability between terminologies in order to facilitate user access to the discrete heterogeneous digital objects held therein. The emergence of the Simple Knowledge Organization System (SKOS) Core is a useful development in this context. In this paper we describe a SKOS Core approach to implementing a web services (i.e. M2M) terminology server employing terminology mapping and using SKOS Core to wrap terminology responses. Aspects advantageous to this approach are explored, as are issues and areas for future research.

Keywords: terminologies, interoperability, Knowledge Organization Systems, SKOS Core, resource discovery, information retrieval

1 Introduction

Knowledge Organization Systems (KOS) encompasses a variety of disparate *terminologies* designed to present a systemised interpretation of knowledge (Zeng & Chan, 2004). KOS can include term lists (e.g. authority files, glossaries, gazetteers, etc.), classification schemes (e.g. bibliographic classification schemes, taxonomies, etc.) and relational vocabularies (e.g. thesauri, subject heading lists, etc.). The proliferation of digital libraries and repositories has increased the need for improved interoperability between terminologies in order to enhance user access to discrete heterogeneous digital objects (Chan & Zeng, 2002). This is particularly true within distributed resource discovery contexts in which digital objects are indexed and organized according to a variety of terminologies, perhaps deriving from disparate KOS. In such contexts it is impractical for users to query each repository individually or to acquaint themselves with the variety of terminologies in use. Simultaneous searching and browsing of multiple distributed repositories is therefore considered increasingly desirable and research exploring techniques designed to artificially or intellectually

In: *Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007), 21-23 February 2007, Bangalore, India.* Documentation Research & Training Centre, Indian Statistical Institute, Bangalore. pp.109-120.

augment cross-repository subject interoperability continues to be a significant area of study (e.g. Binding & Tudhope, 2004; Doerr, 2001; Koch, Neuroth & Day, 2003; Nicholson, Dawson & Shiri, 2006; Zeng & Chan, 2004).

The Simple Knowledge Organization System (SKOS) Core model (Miles & Brickley, 2005) offers a means of expressing the structure of ‘concept schemes’ on the web, facilitating the implementation of operational terminology services within a machine-to-machine (M2M) web services context. In this paper we describe and propose an approach to implementing a pilot M2M terminology server employing *terminology mapping* and using SKOS Core to mark-up terminology responses.

The remainder of the paper is structured as follows. We contextualise our work by defining terminology mapping and provide brief details of the underlying design of the terminology mapping server in section 2. SKOS Core is briefly introduced in section 3. The crux of the paper (section 4) describes the way in which SKOS Core is deployed in our system and explores the potential for server functions. Discussion, issues, areas for future research and development, and conclusions are addressed in section 5.

2 Terminology mapping

Before introducing SKOS Core, it is first necessary to define what is meant by terminology mapping. Mapping essentially involves relating equivalent terms, concept or hierarchical relationships, from one terminology to another (Doerr, 2001). The process of terminology mapping remains largely an intellectual process and is consequently heavily reliant upon human intervention. Within particular scenarios equivalence between terms can be derived via computational means (Vizine-Goetz, Hickey, Houghton & Thompson, 2004); however, most of these approaches still require significant human resources to verify and/or amend erroneous equivalences (McCulloch, Shiri & Nicholson, 2005). Research has therefore focussed on terminology *switching* to reduce the degree of human intervention required and to simplify the management of numerous terminology-to-terminology mappings.

Switching involves the use of a single terminology as an intermediary to translate requests from one scheme to another. For example, all the terminologies to be used within a retrieval system are mapped to a common terminology (X). This enables user queries entered using terminology A to be translated to X and then switched to the equivalent terms in terminology B.

In: *Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007), 21-23 February 2007, Bangalore, India.* Documentation Research & Training Centre, Indian Statistical Institute, Bangalore. pp.109-120.

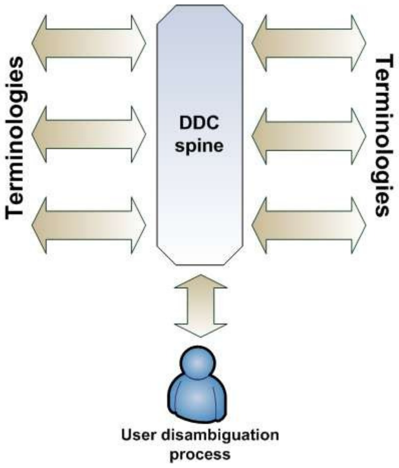


Figure 1: Diagram of the DDC spine-based model employing user 'disambiguation'.

The mapping mechanism employed by the system documented in this paper (HILT: High-level Thesaurus: http://hilt.cdlr.strath.ac.uk/) is similar to switching but differs in that the switching terminology is also central to user *disambiguation* processes (Shiri, Nicholson & McCulloch, 2004). It should be noted that the process of disambiguation not only resolves the existence of homographs (as the term 'disambiguation' may suggest), but encompasses a variety of processes allowing users to qualify their search requirements (Figure 1). This so-called 'spine-based' approach uses the Dewey Decimal Classification (DDC) as a switching spine for searching and permits hierarchical browsing and the discovery of like terms within other terminologies. Although the primary purpose of the terminology mapping server is to enable improved cross-repository searching, it can also provide other terminological functions, such as terminology-based interactive query expansion to assist user query formulation. Such terminology-based techniques have been more formally defined by Efthimiadis (1996) as interactive query expansion based on collection independent knowledge structures.

3 SKOS Core

Simple Knowledge Organization System (SKOS) Core (Miles & Brickley, 2005) is an application of the Resource Description Framework (RDF) and a model proposed by the W3C Semantic Web Best Practices and Deployment Working Group (W3C, 2006). SKOS Core provides a flexible framework for expressing the structure and content of terminologies (or 'concept schemes'), thus enabling efficient machine processing. The framework is flexible enough to accommodate most KOS (as defined in section 1) and essentially consists of a series of RDF properties and RDF Schema (RDFS) classes to encode

In: *Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007), 21-23 February 2007, Bangalore, India.* Documentation Research & Training Centre, Indian Statistical Institute, Bangalore. pp.109-120.

terminologies' content and structural characteristics. For example, a thesaurus could be considered as a series of `skos:Concepts` containing preferred labels (`skos:prefLabel`) and non-preferred labels (`skos:altLabel`). It may also contain various broader terms (`skos:broader`) or related terms (`skos:related`), and so forth. Since the data is encoded in RDF it is inherently pliable and can be utilised or integrated with other RDF data via semantic web applications.

To complement SKOS Core, Miles and Brickley (2004) have proposed the SKOS Core Mapping Vocabulary Specification (MVS). The SKOS Core MVS allows the mapping of concepts between different terminologies using the SKOS Core framework. The properties proposed by SKOS MVS are: exactMatch, broadMatch, narrowMatch, majorMatch and minorMatch. Since like to like mappings are often rare, the MVS also supplements the match types with a series of classes (AND, OR, NOT) for combining or excluding concepts. For example, the 'AND' class is used to denote the intersection of two or more concepts. The term of *Education (United Kingdom)* in terminology A may therefore map to *Education* AND *United Kingdom* in terminology B.

4 Using SKOS Core for terminology services

4.1. SKOS Core: deployment context

The motivation behind the terminology mapping server is to ameliorate the limited terminological interoperability currently afforded between the federation of repositories, digital libraries and information services comprising the UK Joint Information Systems Committee (JISC) Information Environment (JISC, 2003). The expectation is that participant services in the federation will employ Search/Retrieve Web service (SRW) (http://www.loc.gov/standards/sru/srw/) clients to interact transparently with the SRW compliant terminology mapping server during normal service operation (Figure 2). Client requests made to the server will be sent to a database of terminology sets and associated mappings to DDC. Hits identified are then sent back to the server for onward communication to the SRW clients. Testing of this underlying architecture is currently being conducted in collaboration with 'GoGeo!' (http://www.gogeo.ac.uk/) hosted at EDINA (http://www.edina.ac.uk/). GoGeo! provides access to a variety of geospatial datasets, many indexed using disparate terminologies, and constitutes a sound test bed.

In: *Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007), 21-23 February 2007, Bangalore, India.* Documentation Research & Training Centre, Indian Statistical Institute, Bangalore. pp.109-120.

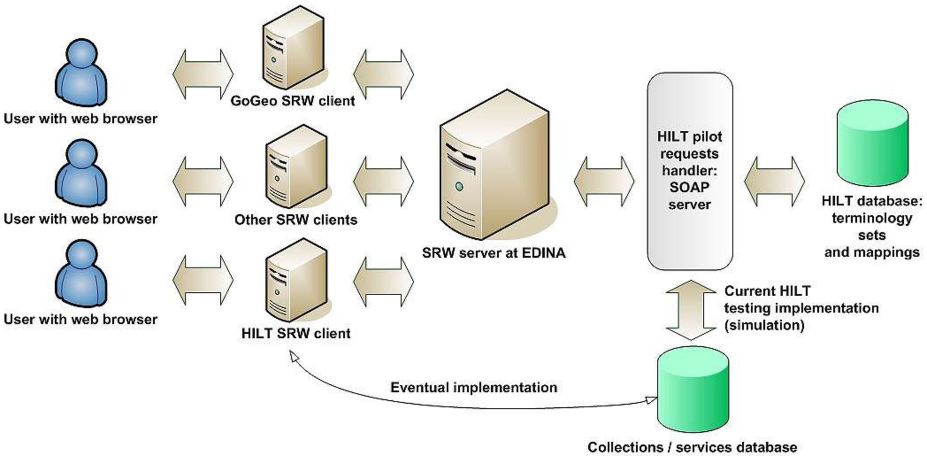

User with web browser
User with web browser
User with web browser
GoGeo SRW client
Other SRW clients
HILT SRW client
SRW server at EDINA
HILT pilot requests handler: SOAP server
HILT database: terminology sets and mappings
Current HILT testing implementation (simulation)
Eventual implementation
Collections / services database

Figure 2: Underlying design of the M2M terminology mapping server

4.2 Wrapping terminologies using SKOS Core

Whilst it is acknowledged that SKOS Core may be deployed in novel and unanticipated ways (Miles, Matthews, Wilson & Brickley, 2005), the main objective of SKOS Core is to facilitate the publication of terminologies for the semantic web, not necessarily for dynamic client/server interactions as described above. However, we propose the use of SKOS Core for the terminology mapping server in order to facilitate meaningful communication with SRW clients regarding the structural nature of the terminological data requested and/or found in the database.

Within the context defined in 4.1, results identified in the database (e.g. scheme information, mapped terms, etc.) are 'wrapped' (i.e. marked-up) in SKOS Core by the SOAP (http://www.w3.org/TR/soap/) server (Figure 2). Since SRW requests (made in Common Query Language (CQL)) are handled using XML over HTTP via the SOAP protocol, terminological data marked-up in SKOS Core can easily be embedded within a SOAP XML envelope for messaging to clients. While such a technical approach potentially permits the future addition of further layers of abstraction, the use of SKOS Core for wrapping is advantageous for several reasons:

- It can accurately model and maintain the structural and semantic properties of the terminological data requested, thus facilitating flexible and reliable re-use by clients in local systems. It is worth noting that such re-use may entail the generation of innovative user interfaces or browsing structures, perhaps displaying results as RDF graphs or providing users with result displays that accurately reflect the hierarchical or semantic structure of the terminological data requested and/or found.

- Although issues exist (to be discussed in section 5), SKOS Core can accommodate the representation of terminology mappings.
- SKOS Core offers opportunities for enhanced interaction with the terminology mapping server, facilitating added functionality such as terminology-based interactive query expansion.

4.3 Server functions in SKOS Core

The M2M pilot terminology mapping server has currently been developed to offer six distinct terminological functions (Nicholson, 2006). These are overly detailed to address in this paper, therefore discussion will concentrate on two (`Get_filtered_set` and `Get_non_DDC_records`).

One purpose of `Get_filtered_set` is to enable the enrichment of users' search vocabulary, provide user feedback and allow limited interactive query expansion. The filtered search can consequently provide (where they exist) related terms (RT), broader terms (BT), narrower terms (NT), preferred terms (PT), and non-preferred terms (NPT). Scope notes may also be provided, depending on the characteristics of the terminology. Although it is possible to envisage such a function being deployed in a variety of user searching scenarios, it is expected that the `Get_filtered_set` function will be of most use to information services that wish to enhance the searching of their local service for users (i.e. enriching users' searching vocabulary to aid query formulation).

For example, a filtered query set to the UK Integrated Public Sector Vocabulary (IPSV) using the term 'Arboriculture' would return a SKOS Core record (Figure 3) providing details necessary to process and invoke a variety of local searching functionality and allowing users to re-interrogate (e.g. using BT, NT, RT, etc. to familiarise users with the topic area within the chosen terminology to aid the subsequent reformulation of search queries, and so forth).

Recall that the primary purpose of the server is to provide terminology mappings using DDC as a spine. There are several functions relating to terminology mapping; one is the `Get_non_DDC_records`. Subsequent to identification of a DDC number by the user via the disambiguation process (as discussed in section 2), `Get_non_DDC_records` provides the client with details of any non-DDC record that includes a mapping to the DDC number sent. Figure 4 provides the SKOS Core output from the terminology mapping server in response to a `Get_non_DDC_records` request for the DDC number 363.34 ('Disasters').

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<SOAP-ENV:Envelope SOAP-
ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/" xmlns:SOAP-
ENV="http://schemas.xmlsoap.org/soap/envelope/"
xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/">
<SOAP-ENV:Body>
<ns1:get_filtered_setResponse xmlns:ns1="http://tempuri.org">
<return xsi:type="SOAP-ENC:Array" SOAP-ENC:arrayType="xsd:string[1]">
<item xsi:type="xsd:string">
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#"
xml:base="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/concepts.php">
<skos:Concept rdf:about="#2715">
          <skos:prefLabel xml:lang="en">Arboriculture</skos:prefLabel>
          <skos:broader rdf:resource="#504"/>
          <skos:narrower rdf:resource="#2633"/>
          <skos:related rdf:resource="#1566"/>
          <skos:related rdf:resource="#15"/>
          <skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/schemes/IPSV.xml"/>
</skos:Concept>
<skos:concept rdf:about="#504">
          <skos:prefLabel xml:lang="en">Horticulture</skos:prefLabel>
          <skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/schemes/IPSV.xml"/>
</skos:concept>
<skos:concept rdf:about="#2633">
          <skos:prefLabel xml:lang="en">Tree planting</skos:prefLabel>
          <skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/schemes/IPSV.xml"/>
</skos:concept>
<skos:concept rdf:about="#1566">
          <skos:prefLabel xml:lang="en">Woodlands</skos:prefLabel>
          <skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/schemes/IPSV.xml"/>
</skos:concept>
<skos:concept rdf:about="#15">
          <skos:prefLabel xml:lang="en">Trees</skos:prefLabel>
          <skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/schemes/IPSV.xml"/>
</skos:concept>
</rdf:RDF>
</item>
</return>
</ns1:get_filtered_setResponse>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

Figure 3: Example of M2M pilot terminology server response to Get_filtered_set in SKOS Core within SOAP envelope. Example illustrates filtered search for the IPSV term 'Arboriculture'. Note: example has been truncated for publication purposes.

The need for encoding the presence of mappings from multiple terminologies in response to client requests invokes the use of the SKOS Core MVS. In Figure 4 most of the mappings are deemed to be exact matches (i.e. `map:exactMatch`). This process enables the identification of a variety of terms from disparate terminologies associated with a particular DDC number

In: *Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007), 21-23 February 2007, Bangalore, India.* Documentation Research & Training Centre, Indian Statistical Institute, Bangalore. pp.109-120.

to be used to search relevant repositories or information services using the correct terminology to match local indexes. Such a function goes some way to ameliorating the current subject interoperability difficulties encountered within the JISC Information Environment. Note that there are several non-DDC terminologies represented in the example: the Global Change Master Directory (GCMD), Library of Congress Subject Headings (LCSH), IPSV, and the UNESCO Thesaurus. The SKOS Core MVS may prove useful for encoding terminology mappings within our model; however, the use of MVS within terminology services highlights particular issues which are further of further study. These will be discussed below in more detail.

5 Conclusion and further research

The M2M pilot implementation proposed in this paper offers terminology mapping as a principal function to enhance user access to disparately indexed heterogeneous digital objects held within multiple repositories, but it also offers terminology-based interactive query expansion functionality. We have demonstrated that SKOS Core can function effectively in a web services environment and that such an approach constitutes a flexible means of implementing various terminological functions for third party terminology services. Our experiments are currently being conducted within in a controlled environment (i.e. the JISC Information Environment); however, we consider the use of SKOS Core for a terminology mapping server (or terminology services generally) to be sufficiently flexible (and scalable, providing the necessary terminology sets exist) so as to permit similar approaches to be used in alternative contexts or global information environments. The wrapping of terminological data within SKOS Core provides a readable means of transporting data over networks (i.e. over HTTP using SOAP) and allows such data to be structured appropriately and modelled correctly, thus facilitating flexible re-use by clients. The authors therefore intend to continue this line of research, including a system and user evaluation of the M2M terminology mapping server as embedded within a client service. Results gleaned from this evaluative work are expected to be disseminated in a separate research paper.

A continuation of this research will demand that several areas undergo further study or investigation. In particular, our work has drawn to attention potential issues within the current draft of the SKOS Core MVS. The definitions of the MVS match types are based on the principles of set theory. Such an abstract paradigm can be useful as an arbitrary means of assessing equivalence in a variety of terminological scenarios; however, their abstractness can also cause uncertainty in practical application. For example, the use of majorMatch (`map:majorMatch`) and minorMatch (`map:minorMatch`) within our

```
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#"
xmlns:map="http://www.w3.org/2004/02/skos/mapping#"
xml:base="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/concepts.php">
<skos:Concept rdf:about="#363.34">
        <skos:prefLabel xml:lang="zxx">363.34</skos:prefLabel>
        <skos:altLabel xml:lang="en">Disasters</skos:altLabel>
        <skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/schemes/DDC.xml"/>
        <map:narrowMatch>
                    <skos:Concept rdf:about="#sh 91000441"/>
        </map:narrowMatch>
        <map:exactMatch>
                <skos:Concept rdf:about="#2256"/>
        </map:exactMatch>
        <map:exactMatch>
                <skos:Concept rdf:about="#762"/>
        </map:exactMatch>
        <map:exactMatch>
                    <skos:Concept rdf:about="#143"/>
        </map:exactMatch>
</skos:Concept>
<skos:Concept rdf:about="#sh 91000441 ">
<skos:prefLabel xml:lang="en">Emergency management</skos:prefLabel>
<skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/schemes/LCSH.xml"/>
</skos:Concept>
<skos:Concept rdf:about="#2256">
<skos:prefLabel xml:lang="en">Natural disasters</skos:prefLabel>
<skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/schemes/UNESCO.xml"/>
</skos:Concept>
<skos:Concept rdf:about="#762">
<skos:prefLabel xml:lang="en">Natural hazards</skos:prefLabel>
<skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/schemes/GCMD.xml"/>
</skos:Concept>
<skos:Concept rdf:about="#143">
<skos:prefLabel xml:lang="en">Civil emergencies</skos:prefLabel>
<skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/schemes/IPSV.xml"/>
</skos:Concept>
</rdf:RDF>
```

Figure 4: Example of M2M pilot terminology server response to Get_non_DDC_records in SKOS Core with the Mapping Vocabulary Specification. Note: example has been truncated for publication purposes and therefore does not show full response or XML SOAP envelope.

model can be difficult to apply as defined and it is unclear how a third party terminology service would incorporate such properties since the current definitions might be interpreted as implying knowledge of database content and indexes. It is noteworthy that similar application difficulties have been encountered by other researchers within different contexts (Liang & Sini, 2006). Current analyses indicate that our approach will require additional match types to those specified by the MVS. Although a pilot service could be implemented using the MVS alone, the lack of detail afforded in the

Specification would, we hypothesise, impose unnecessary cognitive load on the user since only minimal match type feedback would be provided. Any use of match types to assist in the ranking of results (according to the degree of concordance with users' preferred terminology) would also be limited. Future research will therefore aim to test this hypothesis by comparing the relative benefits of each approach for the purposes of user disambiguation.

Future work will also aim to optimise the way in which terminological data is modelled in SKOS Core. Since our system is accommodating numerous terminologies from different KOS, a generic approach has been necessary in the treatment of terminologies and it has therefore not been possible to model the nuances of every particular scheme. For example, LCSH structured headings use a delimiter (--) to denote the use of structured headings (e.g. 'Beach erosion--Monitoring'). Within our current system such a heading is not considered to represent two concepts and is therefore mapped as if it were one concept. Future work will aim to investigate the use of the MVS classes (e.g. AND, NOT, and OR) to optimise the way in which some terminological data is represented and to better accommodate compound concept searching.

6 Acknowledgements

This research is funded by the UK Joint Information Systems Committee (JISC) Shared Services Programme.

7 References

[1] Binding, C. & Tudhope, D. (2004). KOS at your Service: Programmatic Access to Knowledge Organisation Systems, *Journal of Digital Information 4*(4). Retrieved September 19, 2006 from http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Binding/

[2] Chan, L., M. & Zeng, M., L. (2002). Ensuring interoperability among subject vocabularies and Knowledge Organization Schemes: a methodological analysis, *IFLA Journal 28*(5/6) 323-327.

[3] Doerr, M. (2001). Semantic problems of thesaurus mapping, *Journal of Digital Information 1*(8). Retrieved September 19, 2006 from http://jodi.tamu.edu/Articles/v01/i08/Doerr/

[4] Efthimiadis, E., N. (1996). Query expansion, *Annual Review of Information Systems and Technology (ARIST)*, 31, 121-187.

[5] JISC. (2003). *Strategic activities: Information Environment.* Retrieved September 19, 2006 from http://www.jisc.ac.uk/index.cfm?name=about_info_env

[6] Liang, A., C. & Sini, M. (2006). Mapping AGROVOC and the Chinese Agricultural Thesaurus: definitions, tools, procedures, *New Review of Hypermedia and Multimedia 12*(1), 51-62.

[7] McCulloch, E., Shiri, A. & Nicholson, D. (2005). Challenges and issues in terminology mapping: a digital library perspective, *Electronic Library 23*(6), 671-677.

[8] Miles, A. & Brickley, D. (Eds.). (2004). *SKOS Mapping Vocabulary Specification*. Retrieved September 19, 2006 from http://www.w3.org/2004/02/skos/mapping/spec/

[9] Miles, A. & Brickley, D. (Eds.). (2005). *SKOS Core Guide: W3C Working Draft 2 November*. Retrieved September 19, 2006 from http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/

[10]Miles, A., Matthews, B., Wilson, M. & Brickley, D. (2005). SKOS Core: simple knowledge organisation for the Web, *International Conference on Dublin Core and Metadata Applications (DC-2005): Vocabularies in Practice*, September 12-15, 2005, Madrid, Spain. Retrieved September 19, 2006 from http://epubs.cclrc.ac.uk/bitstream/675/dc2005skospapersubmission1.pdf

[11]Nicholson, D. (2006). *HILT M2M pilot requirements document*, v6.0. Retrieved September 19, 2006 from http://hilt.cdlr.strath.ac.uk/hilt3web/reports/h3requirementsv6.pdf

[12]Nicholson, D., Dawson, A. & Shiri, A. (2006). HILT: A pilot terminology mapping service with a DDC spine, *Cataloging & Classification Quarterly 42*(3/4), 187-200.

[13]Shiri, A., Nicholson, D. & McCulloch, E. (2004). User evaluation of a pilot terminologies server for a distributed multi-scheme environment, O*nline Information Review 28*(4) 273-283.

[14]Vizine-Goetz, D., Hickey, C., Houghton, A. & Thompson, R. (2004). Vocabulary mapping for terminology services, *Journal of Digital Information 4*(4). Retrieved September 19, 2006 from http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/

[15]W3C. (2006). *Semantic Web Best Practices and Deployment Working Group*. Retrieved September 19, 2006 from http://www.w3.org/2001/sw/BestPractices/

[16]Zeng, M. L. Chan, L: M. (2004). Trends and issues in establishing interoperability among Knowledge Organization Systems, *Journal of*

the American Society for Information Sciences and Technology 55(5), 377-395.

Terminology server for improved resource discovery: analysis of model and functions

George Macgregor[1], Emma McCulloch[2] and Dennis Nicholson[2]

[1] Information Strategy Group, Liverpool Business School, Liverpool John Moores University, Liverpool, UK
[2] Centre for Digital Library Research, University of Strathclyde, Glasgow, UK
G.R.Macgregor@ljmu.ac.uk; e.mcculloch@strath.ac.uk; d.m.nicholson@strath.ac.uk

Abstract. This paper considers the potential to improve distributed information retrieval via a terminologies server. The restriction upon effective resource discovery caused by the use of disparate terminologies across services and collections is outlined, before considering a DDC spine based approach involving inter-scheme mapping as a possible solution. The developing HILT model is discussed alongside other existing models and alternative approaches to solving the terminologies problem. Results from the current HILT pilot are presented to illustrate functionality and suggestions are made for further research and development.

1 Introduction: Subject Interoperability Problem

One impediment to searching distributed digital collections is the difference in metadata standards used, particularly within subject or keyword fields [1]. By adopting different subject schemes, information providers may unwittingly prevent the widespread discovery of, and therefore access to, their resources. Unless the terminology employed by online collections and services is widely used and/or known to the user, search terms may not match those embedded within resource metadata. The likelihood of this depends on a variety of factors, including the knowledge of the user and the specificity of the resource. Figure 1 illustrates the problem simplistically. A indicates the subject(s) of retrieved documents and B indicates the subject(s) of those that may remain undiscovered in response to a user query for 'Lung disease' within a traditional information retrieval (IR) system. (There are a great many more terms, and from a wider range of schemes, that may feature in either A or B; Fig. 1 shows a selection of these only.)

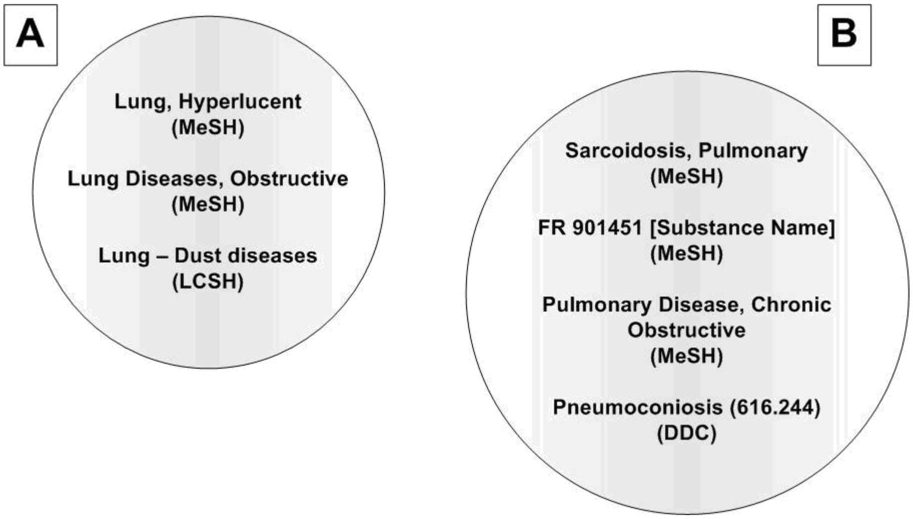


Fig. 1. Examples of documents retrieved in response to Lung disease (A), via assigned subject metadata, together with scheme information, and documents not retrieved (B).

Figure 1 shows that the user query will not retrieve documents indexed using specific terms, which may be conceptually equivalent to the user's search term 'Lung disease'. Depending on the user's perspective on any given topic therefore, vital documents may be missed. For example, amongst the potentially relevant material not retrieved are resources concerned with various aspects of lung disease including specific manifestations and treatments.

This 'translation' problem between subject schemes creates a barrier to discovery and access, and various methodologies to address this well-documented problem have been proposed over the years [2][3][4][5]. This paper will focus on the model adopted by the HILT project (http://hilt.cdlr.strath.ac.uk) and will discuss the potential of such a system to overcome, or at least minimise, the lack of interoperability afforded by collections and services' adoption of different schemes.

The paper describes and discusses a pilot terminologies service designed to facilitate resource discovery and access across distributed heterogeneous services by improving interoperability via inter-scheme mapping. Section 2 provides a general description of the HILT model. Section 3 reviews alternative models and their features, while section 4 pays particular attention to the use of SKOS Core. Section 5 presents HILT results sets and considers their ability to improve distributed information retrieval. Section 6 discusses the value of each of the functions in relation to the aim of improving resource discovery and section 7 presents conclusions and suggestions for further research.

2 The HILT Solution

The current instantiation of HILT [Fig. 2] demonstrates the model's functionality via the use of two (or more) independent SRW clients, a central SRW server and a SOAP server, described in Fig. 2 as the 'HILT pilot requests handler: SOAP server'. Non-proprietary standards including SRW [6] have been adopted enabling services to develop their own local user interfaces, capable of connecting to the HILT SRW server and employing HILT mappings within their local environment(s). Completing the model are two databases; one holding records of collections and services within the JISC (Joint Information Systems Committee) Information Environment [7] and the other holding terminologies data including mappings from satellite schemes to the central DDC spine. The response to a user query is wrapped in SKOS (Simple Knowledge Organization System) Core [8].

HILT's model involves inter-scheme mapping, whereby concepts/terms from a range of different schemes are mapped to a Dewey Decimal Classification (DDC) Scheme [9] [10] spine, which acts as a switching language [11][12]. The mapping of subject schemes is not problem free [4][1]. Schemes typically illustrate "'theoretical, conceptual, cultural and practical" [13] variations, often making the mapping process difficult, particularly if implemented via an intermediary switching language. The process has also been documented as costly and time consuming [1], as well as highly variable in its success according to subject area [3] due to differing structures, levels of specificity and, particularly, the varying proportion of single and compound terms within domain-specific schemes [14].

Despite its various drawbacks, the mapping approach does offer a practical solution to the interoperability issue, provided sound methodologies are adopted and that 'complete' mappings are implemented. Complete in this sense refers to the extent of mappings implemented between a term or concept in one scheme and any number of possibly equivalent terms or concepts in another. It is highly probable that "one-to-one relationships are certainly not sufficient" [13] for the purposes of an effective terminology server in a distributed information retrieval environment. HILT is piloting a mapping based system, investigating the value of high level mapping and more granular, complete mapping within specific subject areas. It is worth noting that the model also provides some generic terminological functionality, such as the provision of broader and narrower terms, related terms, non-preferred terms and so forth. Such terminological data can be used by services to implement retrieval tools such as interactive query expansion or hierarchical browsing of scheme data [15].

HILT currently holds XML versions of DDC 22 [9], AAT (Art and Architecture Thesaurus) [16], GCMD (Global Change Master Directory) [17], HASSET (Humanities and Social Science Electronic Thesaurus) [18], IPSV (Integrated Public Sector Vocabulary) [19], JACS (Joint Academic Coding System) [20] JITA (Classification Scheme used within E-LIS repository) [21], LCSH (Library of Congress Subject Headings) [22], MeSH (Medical Subject Headings [23], NMR (National Monuments Record Thesaurus) [24], SCAS (Standard Classification of

Academic Subjects)[25] and UNESCO Thesaurus [26]. The adoption of further schemes is currently under consideration due to the need to satisfy the requirements of two JISC services/collections within the remit of further research. An example of a scheme to be added in the near future is CAB Thesaurus [27]. By incorporating all schemes used by services and collections within the JISC Information Environment (or, indeed, any given realm) it is envisaged that individual services will be able to implement their own mappings between local collections and the centrally available HILT DDC spine. Appropriate documentation would be provided by the HILT project to facilitate this process and to ensure standardisation and consistency throughout.

Like the selection of individual schemes, the adoption of a DDC spine has been purposive. Not only is DDC a universal scheme covering most subject areas, it is also available in many languages, thus potentially facilitating multi-lingual as well as multi-KOS interoperability. Another advantage of adopting DDC as a spine is that there already exist many mappings to it from other schemes such as LCSH [22] and MSC (Mathematics Subject Classification) [28].

Preliminary research has been conducted [29] into the various types of mapping required within a system such as HILT. It is thought necessary to characterise the range of different types of equivalence imposed between terms/concepts from disparate schemes, partially to provide users with detailed relevance feedback but also as a basis for ranking results returned in response to any given search. For example, a plural version of a user's singular search term may, in some cases, be more valuable than a narrower term.

Based on an earlier study by Chaplan [30], McCulloch and Macgregor [29] determined a need for at least nine types of equivalence relationship and consider it necessary that mapped terms be encoded accordingly, in order to provide the user with information on whether or not a search term returned by the system is, for example, a synonym (i.e. concept match), a plural version or a broader or narrower term of that originally sought by the user. Dolin et al [31] have noted that "Because the relationship between two concepts can differ depending on the use case, it is possible that different cross map sets will contain the same source and target concept, but with a different relationship". This may suggest that a single mapping requires to be encoded to reflect multiple types of equivalence.

The SKOS MVS (Mapping Vocabulary Specification) has been proposed as a means of categorising the various types of relationship evident between mapped terms [32]. This has proven insufficient at its current stage of development and suggestions for extending the MVS have been submitted [29] [33].

Alternative models proposed for terminology services and as potential solutions to the interoperability problem will now be briefly presented. The HILT model will thereafter be described in further detail in relation to its functionality, with discussion of how such an intermediary system could be exploited within the distributed information environment to improve resource discovery.

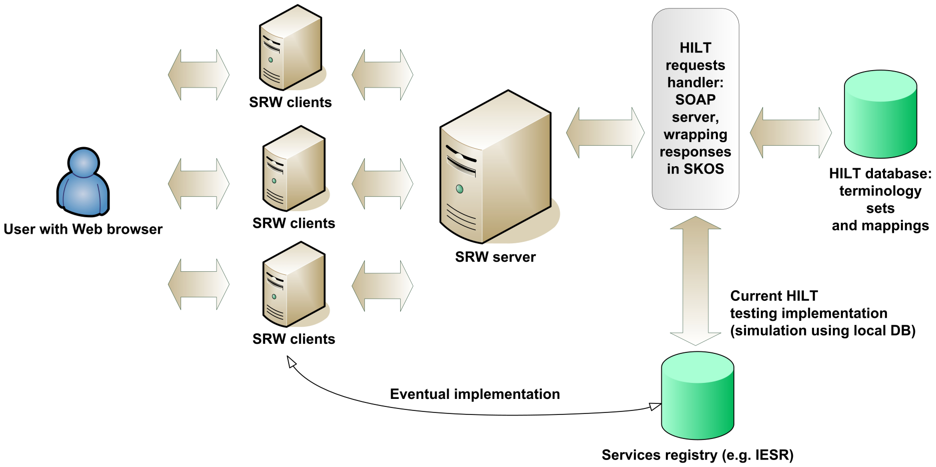

SRW clients
SRW clients
SRW clients
User with Web browser
SRW server
HILT requests handler: SOAP server, wrapping responses in SKOS
HILT database: terminology sets and mappings
Current HILT testing implementation (simulation using local DB)
Eventual implementation
Services registry (e.g. IESR)

Fig. 2. HILT pilot architecture.

3 Alternative approaches

Although there are many different approaches to solving the interoperability problem, for the purpose of this paper we will limit ourselves to reviewing those developing terminology servers. Many different examples of terminology servers and services have been proposed [34] [35], too numerous to review here. We will therefore further limit ourselves to discussing those that adopt mapping methodologies. We will consider one general model as well as looking at one within a specific subject domain - medicine, a domain in which much research and development has been conducted into the merging of, and switching between, standard terminologies in use.

The Aquarelle terminology service [36] exhibits the same basic components as HILT, namely ”vocabularies in local databases, local thesaurus management systems of wider use and central Term Servers for retrieval”. Although currently HILT holds terminologies centrally within the same site as the main terminology server the vision is that this element of the model will become distributed in due course, with individual collections and services able to plug their own local terminologies into the central model. The overview of the Aquarelle service shown in Figure 3 indicates a significant degree of similarity to the centralised HILT model.

The Aquarelle service was developed in the 1990s [36][37] but is no longer in operation. It is unclear whether the project was discontinued due to the viability of the model or for other reasons.

A second initiative worth noting is the GALEN programme, one component of which is the GALEN terminology server [38]. GALEN is an operational terminology service active within the clinical area. It offers the ability to provide clarification of concepts, e.g. do you know about the ”leg”?; concept manage-

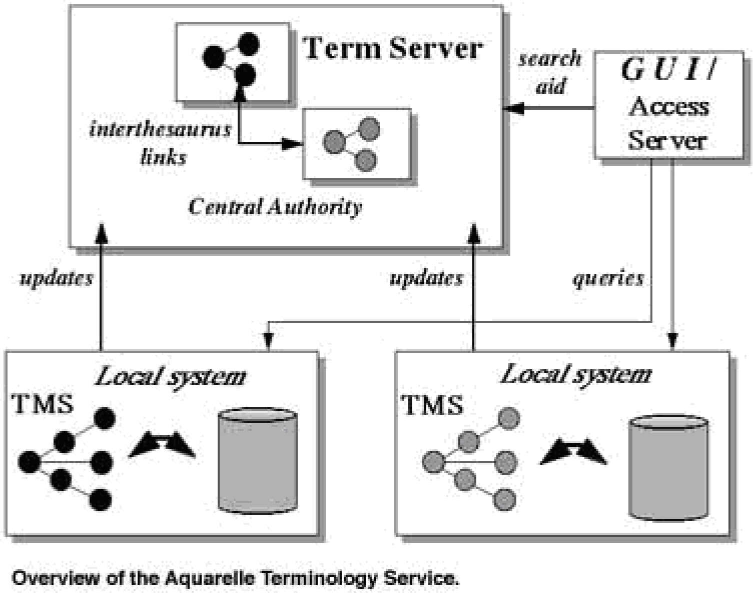


Fig. 3. Aquarelle Terminology Service architecture.

ment and specialisation, e.g. what is known about the leg? What bones does it contain? if they are broken, how might they be clinically described?; translation functionality, e.g. "what is a French language phrase for the combination of a severe fracture of the neck of the left femur?"; identification of the preferred term for a particular concept; coding e.g. "what is the closest ICD code for this concept?"; and extrinsic information e.g. "is there any relevant literature known about this condition?". Providing this range of functionality is an architectural model that fits "very comfortably with the notion of client-server computing, and commercial implementations now use standard object component technologies to deliver their services" [38].

Contrary to the primary function of the HILT model (i.e. to switch between several different terminologies via inter-scheme mappings), the GALEN model is optimised for the answering of clinical questions and appears to provide a broad databank relating to various aspects of conditions and treatments and so on, as opposed to acting as an intermediary between the user and services or collections. In this respect it appears more closely related to the notion of an expert system. Although architecturally similar, the functionality of GALEN is very different to that of HILT. GALEN does map natural language to concepts and concept to classification schemes, but the purpose of doing so is more extensive than the provision of a switching mechanism.

It has been documented [39] that the key desiderata for a clinical terminology server are 1) word normalization, 2) word completion, 3) target terminology specification, 4) spelling correction, 5) lexical matching, 6) term completion, 7) semantic locality, 8) term composition and 9) decomposition [39]. These functions echo those identified as desirable by HILT. However, the purpose of such a clinical server is mainly to enable "clinicians to enter patient observations, findings, and events, such as procedures. It does not need to carry the weight of terminology updates, maintenance, or development and thus might be regarded as a server "lite"." Quite distinct from HILT's aim to improve mediated resource discovery and retrieval for the end user, it seems that the primary users of this type of model are professionals, who are likely to have a substantial degree of knowledge about the terminology and conditions being queried.

It seems therefore that although much of the functionality desired by HILT is also desirable in other domain specific terminology servers. HILT represents a novel implementation in that it aims to cover all areas of knowledge, by incorporating and mapping together schemes from all disciplines and (eventually) languages. It follows that HILT has a wider remit than other servers currently implemented. Although the Aquarelle service is similar to HILT in terms of architecture and functionality, its stage of development remains unclear.

Although dissimilar in architectural terms, Renardus [40] is similar to HILT in that it employs DDC as its central terminology. This service enables users to search by title, subject, description, creator, document type or DDC classification. In contrast to HILT, Renardus retrieves item level resources in response to the entry of a DDC number, without first clarifying what the user is intending to search for. This aspect of the model is not conducive to user interrogation since the average user is unfamiliar with DDC notation and is likely to experience difficulties in expressing an information need in this way. HILT, on the other hand, provides the user with DDC captions relating to a specific numerical notation, providing relevance feedback throughout the search process. The user is able to ensure he/she is within the correct discipline, determining the relevant focus of a given subject, since different aspects of the same basic concept may be located in various disciplines of a classification system. When browsing the DDC hierarchy for a subject in Renardus - thus accessing the more meaningful captions of the scheme - the service intends to link the user into gateways holding records on the subject of interest. At the time of writing it was noted that few gateway services have retained collaboration with Renardus, resulting in 'dead ends' for many of the browse trees.

Should the HILT architecture and general model prove effective, it may be that elements of the HILT model could be tackled in different ways. For example, is a DDC spine the best option in this context? The very nature of DDC (and indeed library classifications) has been questioned and undoubtedly causes problems relating to the mapping of schemes [41]; most obviously because the majority of schemes contain terms and/or concepts whereas the unique identifier conveying a concept in DDC is a numerical notation. Further difficulties stem from the analytico-synthetic properties of DDC, requiring a subject to be analysed before undertaking the synthesis of an appropriate notation by which it can be expressed. This means that all notations to which terms from a satellite scheme may require to be mapped will not necessarily be pre-coordinated; that is, the mapping process may also require an extensive process of number building to express concepts accurately. In conducting such number building it is common to add standard subdivisions to a basic concept, where rules tend to vary according to circumstance. For example, where a three digit notation ends in 0 e.g. 370, the 0 added to indicate the addition of a standard subdivision is omitted; in other circumstances there may be an instruction to add an extra 0. These types of practice are likely to have implications for the truncation process adopted by HILT, described in section 5.1. Standard subdivisions can only be added once, which means that subjects referring to multiple locations or dates cannot be expressed adequately. So, for instance, France and Belgium cannot be incorporated into a single notation to express, for example, French language usage in these two countries. One final difficulty worth mentioning is that not all areas of DDC reflect the superordinate or subordinate nature typical of hierarchical schemes. An example of this can be seen in the 900 section, where 900 denotes History, geography, and auxiliary disciplines [42]. One level down the hierarchy lies 970 denotes History of North America, while 973 relates to United States. Although, therefore, United States is subordinate to History of North America, this is not reflected in the DDC notation, with each number being of equal length.

Such limitations seem to warrant the investigation of alternative schemes, bearing in mind that an effective spine must be universal in nature since it should encompass all concepts expressed within all other schemes being mapped [12]. Although much work in the area features a central DDC spine [40] or mappings of individual schemes to DDC [28], several other projects have employed a central terminology other than DDC. UDC (Universal Decimal Classification) [43] has been adopted in this context due to its ability to offer ”international notation, depth documentation, retrieval and mechanization facilities” [44] [45]. Other initiatives have implemented direct mappings between two disparate schemes [30] [33] devoid of the switching model favoured by HILT. Although clearly valid and likely to improve retrieval within a given subject discipline, it is unlikely that such an approach would prove universally effective or scalable.

4 SKOS: Modelling Terminological Data

SKOS Core [8] is a useful development within the context of M2M terminology service architectures. SKOS Core is an application of the Resource Description Framework (RDF) proposed by the W3C Semantic Web Deployment Working Group [46] and provides a flexible framework for representing the structure and content of KOS (or ’concept scheme’) on the Web. SKOS Core essentially comprises a series of RDF properties and RDF Schema (RDFS) classes to encode the content and structural characteristics of KOS. As an application of RDF, SKOS data remains inherently adaptable and can be integrated with other RDF data on the Web using Semantic Web applications. A draft mapping specification has also been proposed by Miles et al [32] enabling the mapping of concepts between different KOS within the SKOS framework.

Although the primary objective of SKOS Core is to provide a means of publishing KOS for the Semantic Web, use of the specification for dynamic client-server interactions has attracted attention from those active in terminology service research and development [47] [48] [49] [50]. SKOS Core can prove particularly advantageous in such contexts since terminological data can be richly modelled and data structures can be maintained when communicating with clients, particularly when using web service protocols such as SOAP [15]. This can facilitate reliable, flexible and simple multipurpose reuse by client services.

Alternative frameworks are available to facilitate the aforementioned functionality. These can occasionally be inappropriate or less flexible, thus increasing the potential for low adoption among client services. Despite increased complexity, OWL [51] has been demonstrated as effective within similar technical architectures [52]. It also continues to be used successfully to represent some terminological data [53]; however, it remains unsuitable for other schemes [54]. For example, the OWL class-instance does not reflect the structure of all KOS, resulting in the need for unnecessary KOS reengineering [55].

Zthes [56] provides an abstract model and an XML schema for relational vocabulary representation (particularly thesauri) and is suitable for 'storing and transmitting' such terminological data. Use of Zthes can be advantageous as the specification also defines how queries to Zthes-compliant terminologies can be implemented using Z39.50 and/or SRU/W. Further experimentation with this approach has been undertaken by Vizine-Goetz et al [57]. However, Zthes remains less suited to handling disparate terminological data [58]. The flexibility of SKOS and its increased suitability with Web services and the Semantic Web community make it more conducive to the system we demonstrate here [59].

5 Functionality

Within the third phase of the project five distinct functions were implemented to simulate ways in which users may interact with HILT, based on a set of use cases [60]. It was deemed desirable to build a system which could, for example, 1) provide terminological data on any given term within a scheme held; 2) return all instances of a given search term within DDC, together with the appropriate hierarchical data and DDC notation; 3) return all terms across schemes related (predetermined via mapping) to the DDC notation matched to a given search term; 4) return combinations of 1), 2) and 3) as specified by the user.

Each of the functions developed (get_collections, get_all_records, get_ddc_records, get_non_ddc_records, get_filtered_set) will be discussed in turn, to help contextualise their purposes, with a view to aiding discovery and access across distributed digital collections. The purpose and mechanism of each function will be documented, before illustrating its value, or otherwise, by presenting HILT output in response to an example query. This will better explain the strengths and weaknesses of the system in its current instantiation. Examples will be given for queries sent to the HILT pilot requests handler: SOAP server via the test HILT SRW client [61].

5.1 get_collections

The get_collections function aims to provide the user with collection information relevant to the area of a subject query. It will return information and/or a link to and/or dynamic searching of any collection(s) classified under a specified DDC number or its stem. The process is carried out as follows:

1. A DDC number relating to a caption/hierarchy identified during the disambiguation stage (user enters term prior to this stage; this is then matched to appropriate notation(s)) is sent from the SRW client service to the SRW server.
2. The SRW server sends an appropriate request for get_collections via the SOAP server.
3. The get_collections function queries the database using successive truncations of the DDC number sent.
4. The SOAP requests handler receives back collections' connection details and scheme information.
5. The SOAP requests handler wraps the results in Dublin Core Collection Description Application Profile (DC CD AP) and sends the results back to the SRW server.
6. The SRW server sends the results back to the client service.
7. The client service processes the results to offer the user a set of collections relevant to their query.

On entering the query '371.07' (Education - Schools and their activities; special education - Religious schools) to the pilot demonstrator search box [61] (which simulates the processes of stages 1 and 2 above) the following result is returned. The result is expressed in DC CD AP within a SOAP envelope (envelopes have been edited out in all examples given). On development of a more advanced end-user oriented system, the result will be parsed by a client and presented to the end-user in a human readable format, dependent on how a given local service decides to present the information being returned.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-SOAP envelope -->
<metadata
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:dcmitype="http://purl.org/dc/dcmitype/"
xmlns:iesr="http://iesr.ac.uk/terms/#usesControlledList"
xmlns:cld="http://purl.org/cld/terms/">
```

```
<dcmitype:Collection>
<dc:title>BUBL LINK: Education</dc:title>
<dc:identifier xsi:type="dcterms:URI">http://bubl.ac.uk/link/
</dc:identifier>
<dcterms:abstract>Catalogue of selected Internet resources.
</dcterms:abstract>
<dc:creator>BUBL Information Service</dc:creator>
<dc:type xsi:type="dcterms:DCMIType">Collection</dc:type>
<dc:subject xsi:type="dcterms:DDC">370</dc:subject>
<cld:isAccessedVia>http://hilt.cdlr.strath.ac.uk/bublsearch/
bubl.cfm?queryString=</cld:isAccessedVia>
</dcmitype:Collection>
<dcmitype:Collection>
<dc:title>Education-line</dc:title>
<dc:identifier xsi:type="dcterms:URI">http://www.leeds.ac.uk/
educol/</dc:identifier>
<dcterms:abstract>Project funded under the Electronic Libraries
programme to gather an electronic archive of preprints, grey
literature and texts in education and training. </dcterms:abstract>
<dc:creator>Leeds University</dc:creator>
<dc:type xsi:type="dcterms:DCMIType">Collection</dc:type>
<dc:subject xsi:type="dcterms:DDC">370</dc:subject>
</dcmitype:Collection>
</metadata>
```

Fig. 4. Result for get_collections function using query '371.07' (Education - Schools and their activities; special education - Religious schools).

Figure 4 shows two collections being returned in response to the query '371.07': BUBL LINK: Education and Education-line. The value of this function is illustrated by its flexibility. For example, Figure 4 above shows that both collections returned have been classified in the system's collections database at DDC 370. This is due to the ability of the system to truncate a DDC number successively in the event of no direct matches in response to a query. Since no match was found for 371.07, the system has searched upwards through the DDC hierarchy until a match was found at 370. This means that however specific the DDC number sent via point 1 above is, collections should always be returned, even if broadly classified at one of the ten main classes (i.e. 000 - 900). Once collections have been identified at any given point via the process of truncation, no further truncation will be invoked. This means that a query for 371.07 will return the two collections above classified at 370, but will not present more general collections relating to education, classified at 300.

For research purposes, experimentation for get_collections has been with a local collections database containing test data; however, the model has been designed to interact with distributed service registries as a source of accurate collection and service descriptions. To this end research testing HILT interaction with the Information Environment Services Registry (IESR) [62] is currently being pursued.

5.2 get_all_records

The get_all_records function retrieves records that include - or are mapped to records that include - the term or term phrase specified within a given query. This function operates as follows:

1. User enters term via the embedded SRW client service, and a resultant request is sent to the SRW server.
2. The SRW server parses the request to obtain search terms and uses these to call the SOAP get_all_records function.
3. The get_all_records function queries the database to find (1) all DDC records that either include the user term or that are mapped to from other non-DDC records that include the term (2) all non-DDC records mapped from the DDC records retrieved under (1) and returns these records to the SOAP server.
4. The SOAP requests handler wraps the results in SKOS Core with the SKOS Mapping Vocabulary Specification (MVS) and sends the results to the SRW server.
5. The SRW server sends the results back to the client service.
6. The client service processes the results to offer DDC and non-DDC records to the user.

The result of a query entered selecting the get_all_records function should contain DDC numbers, mapped terms and details of what scheme such terms belong to, and mapping match type information denoting the nature of the equivalence relationship imposed. The following code (Figure 5), embedded within a SOAP envelope, illustrates the result returned in response to a query for 'Natural hazards':

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-SOAP envelope -->
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core.rdf#"
xmlns:map="http://www.w3.org/2004/02/skos/mapping#"
xml:base="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/concepts.php">
<skos:Concept rdf:about="#363.34">
<skos:prefLabel xml:lang="zxx">363.34</skos:prefLabel>
<skos:altLabel xml:lang="en">Disasters</skos:altLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/DDC.rdf"/>
<map:exactMatch>
<skos:Concept rdf:about="#16117"/>
</map:exactMatch>
<map:exactMatch>
```

```
<skos:Concept rdf:about="#16118"/>
</map:exactMatch>
<map:narrowMatch>
<skos:Concept rdf:about="#16119"/>
</map:narrowMatch>
<map:narrowMatch>
<skos:Concept rdf:about="#2256"/>
</map:narrowMatch>
<map:narrowMatch>
<skos:Concept rdf:about="#762"/>
</map:narrowMatch>
<map:exactMatch>
<skos:Concept rdf:about="#2696"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#143"/>
</map:exactMatch>
</skos:Concept>
<skos:Concept rdf:about="#16117">
<skos:prefLabel xml:lang="en">Disasters</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#16118">
<skos:prefLabel xml:lang="en">Emergency management</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#16119">
<skos:prefLabel xml:lang="en">Natural disasters</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#2256">
<skos:prefLabel xml:lang="en">Natural disasters</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/UNESCO.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#762">
<skos:prefLabel xml:lang="en">Natural Hazards</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/GCMD.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#2696">
<skos:prefLabel xml:lang="en">HAZARDS, ACCIDENTS AND DISASTERS
```

```
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#143">
<skos:prefLabel xml:lang="en">Civil emergencies</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/IPSV.rdf"/>
</skos:Concept>
</rdf:RDF>
```

Fig. 5. Result for get_all_records function in SKOS RDF/XML, using the query 'Natural hazards'.

Figure 5 shows that following the initial query for 'Natural hazards', DDC 363.34 (Disasters) was selected as an appropriate match. In addition to the DDC record returned, a number of mappings to DDC 363.34 from other satellite schemes were returned as shown in Table 1.

Term	Source Scheme	Type of Equivalence
Disasters	DDC	Exact match*
Disasters	LCSH	Exact match
Emergency management	LCSH	Exact match
Natural disasters	LCSH	Narrow match
Natural disasters	UNESCO	Narrow match
Natural hazards	GCMD	Narrow match
Hazards, accidents and disasters	HASSET	Exact match
Civil emergencies	IPSV	Exact match

Table 1. Summary of results for 'Natural hazards', selecting get_all_records function *note that exact match in this sense (in line with SKOS MVS) encompasses a concept match.

The encoded result and Table 1 indicate the range of related terms available within the loaded terminologies. These enjoy some form of equivalence relationship with the original query. By offering synonymous and narrower terms to the user query, HILT is providing the opportunity to explore matched concepts in other schemes and by extension interrogate alternative repositories using the correct query to match local indexes. It also allows users to conduct a more specific search by opting to use those terms returned as having a narrower foci than the original query.

5.3 get_ddc_records

The get_ddc_records function retrieves any DDC record that includes the term(s) specified, or that is mapped to by a record from another scheme that includes the term(s) specified. This function is handled as follows:

1. User enters term via embedded SRW client service, and a resultant request is sent to the SRW server.
2. The SRW server parses the request to obtain search terms and uses these in a call to the SOAP get_ddc_records function.
3. The get_ddc_records function queries the database for DDC records that include the user term entered or that are mapped to by non DDC records that include the term.
4. The SOAP requests handler receives DDC numbers and associated DDC captions, wraps the results in SKOS Core, and sends them back to the SRW server.
5. The SRW server sends the results back to the client service.
6. The client service processes the results to offer the user terms possibly relevant to their query from DDC with corresponding DDC numbers.

Figure 6 illustrates functionality in response to a search for a DDC caption, 'Shore protection'.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-SOAP envelope -->
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core.rdf#"
xml:base="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/concepts.php">
<skos:ConceptScheme rdf:about="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/DDC.rdf"/>
<skos:Concept rdf:about="#627.58">
<skos:prefLabel xml:lang="zxx">627.58</skos:prefLabel>
<skos:altLabel xml:lang="en">Shore protection</skos:altLabel>
</skos:Concept>
<skos:Concept rdf:about="#333.91716">
<skos:prefLabel xml:lang="zxx">333.91716</skos:prefLabel>
<skos:altLabel xml:lang="en">Shore protection, . . .
</skos:altLabel>
</skos:Concept>
</rdf:RDF>
```

Fig. 6. Result for get_ddc_records in SKOS RDF/XML for the query, 'Shore protection'.

The result shows two distinct incidences of the caption 'Shore protection' within the DDC schedules; one instance resides in the 600 section (Technology) with the other dealing with social aspects of 'Shore protection' in the 300 section (Social sciences). No results are returned from any scheme other than DDC in response to this function. Part of the added value offered as a result of the mapping based methodology adopted by HILT in relation to the get_ddc_records function is that DDC records will be returned following matches to terms in other schemes, which are mapped to DDC. An example whereby ’Plant genetics’, a known term from the HASSET scheme, was searched for using the get_ddc_records follows (Figure 7):

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-SOAP envelope -->
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core.rdf#"
xml:base="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/concepts.php">
<skos:ConceptScheme rdf:about="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/DDC.rdf"/>
<skos:Concept rdf:about="#631.5233">
<skos:prefLabel xml:lang="zxx">631.5233</skos:prefLabel>
<skos:altLabel xml:lang="en">Agricultural genetics</skos:altLabel>
</skos:Concept>
<skos:Concept rdf:about="#581.35">
<skos:prefLabel xml:lang="zxx">581.35</skos:prefLabel>
<skos:altLabel xml:lang="en">Genetics</skos:altLabel>
</skos:Concept>
<skos:Concept rdf:about="#631.53">
<skos:prefLabel xml:lang="zxx">631.53</skos:prefLabel>
<skos:altLabel xml:lang="en">Plant propagation</skos:altLabel>
</skos:Concept>
</rdf:RDF>
```

Fig. 7. Result for get_ddc_records in SKOS RDF/XML for the query, ’Plant genetics’.

Figure 7 shows the DDC notation, and corresponding captions, to which the HASSET term ’Plant genetics’ is mapped. Three mappings have been implemented; one to DDC 631.5233 ’Agricultural genetics’; one to DDC 581.35 ’Genetics’ and a third to DDC 631.53 ’Plant propagation’. Clearly the value of such results is user dependent, and reliant on the completeness of mappings implemented.

5.4 get_non_ddc_records

The get_non_ddc_records function retrieves any non-DDC record that includes a mapping to the DDC number sent. That is, the system retrieves records from other schemes (non-DDC) that have been mapped to an input DDC number. Only the non-DDC records mapped to the DDC number sent are retrieved, as follows:

1. User chooses DDC number on screen and embedded SRW client service sends an appropriate request to the SRW server.
2. The SRW server parses the request and sends an appropriate query to the SOAP get_non_ddc_records function.

3. The get_non_ddc_records function searches the database to find non-DDC records containing a mapping to the DDC number sent and returns the results to the SOAP server.
4. The SOAP server wraps the results in SKOS Core and SKOS MVS and returns them to the SRW server.
5. The SRW server sends the results back to the client service; results comprise DDC number entered, terms from other schemes mapped to that DDC number, with the name of the scheme and match type information defining the relationship between a scheme's term and the DDC number entered.
6. The client service processes the results and provides the user (via the service interface) with information on which term to use for individual schemes used by individual JISC collections.

The DDC notation 631.53 will form the search query to illustrate the get_non_ddc_records function. We saw from the get_ddc_records result above that this notation relates to 'Plant propagation'. The result for this query is presented below (Figure 8):

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-SOAP envelope -->
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core.rdf#"
xmlns:map="http://www.w3.org/2004/02/skos/mapping#"
xml:base="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/concepts.php">
<skos:Concept rdf:about="#631.53">
<skos:prefLabel xml:lang="zxx">631.53</skos:prefLabel>
<skos:altLabel xml:lang="en">Plant propagation</skos:altLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/DDC.rdf"/>
<map:exactMatch>
<skos:Concept rdf:about="#36011"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#36012"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#36013"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#36014"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#36015"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#36016"/>
```

```
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#2539"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#17"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#4712"/>
</map:exactMatch>
</skos:Concept>
<skos:Concept rdf:about="#36011">
<skos:prefLabel xml:lang="en">Plant breeding</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#36012">
<skos:prefLabel xml:lang="en">Plant cell culture</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#36013">
<skos:prefLabel xml:lang="en">Plant micropropagation
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#36014">
<skos:prefLabel xml:lang="en">Plant mutation breeding
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#36015">
<skos:prefLabel xml:lang="en">Plant propagation</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#36016">
<skos:prefLabel xml:lang="en">Vegetative propagation
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#2539">
```

```
<skos:prefLabel xml:lang="en">Plant genetics</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/UNESCO.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#17">
<skos:prefLabel xml:lang="en">Plant Breeding and Genetics
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/GCMD.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#4712">
<skos:prefLabel xml:lang="en">PLANT GENETICS</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
</rdf:RDF>
```

Fig. 8. Result for a get_non_ddc_records query for DDC 631.53 (Plant propagation), in SKOS RDF/XML.

The system has retrieved nine results, summarised in Table 2:

Term	Source Scheme	Type of Equivalence
Plant breeding	LCSH	Exact match
Plant cell culture	LCSH	Exact match
Plant micropropagation	LCSH	Exact match
Plant mutation breeding	LCSH	Exact match
Plant propagation	LCSH	Exact match
Vegetative propagation	LCSH	Exact match
Plant genetics	UNESCO	Exact match
Plant breeding and genetics	GCMD	Exact match
Plant genetics	HASSET	Exact match

Table 2. Summary of results for get_non_ddc_records.

Table 2 indicates that terms have been retrieved from a total of four distinct schemes, relating to the search for DDC 631.53. This notation and corresponding caption is shown at the beginning of the result set, before listing all terms mapped to this notation from other schemes. As mentioned before, work continues into establishing mapping types and appropriate coding of such equivalence relationships. The indication that all terms are 'exact matches' to the original query is therefore misleading. Where explicit relationships have not yet been established within the HILT research programme, the default is to express any relationship as an exact match; this will be rectified as the project progresses.

5.5 get_filtered_set

get_filtered_set is a more generic terminological function, not employing the use of mappings. get_filtered_set retrieves records that meet the specified parameters;

that is, the search term entered but 'filtered' by scheme name(s) and /or field name(s). Functionality to filter a search by scheme, and/or to search preferred and non-preferred terms will be in-built. This enables a user to search one scheme directly, or to incorporate multiple schemes in the scope of his/her search. The get_filtered_set function operates as described below:

1. User enters term via embedded SRW client service, and a resultant request is sent to the SRW server.
2. The SRW server parses the request and uses the results to send an appropriate query to the SOAP get_filtered_set function.
3. The get_filtered_set function queries the database for records that match the terms and the specified filters and the results are sent back to the SOAP server.
4. The SOAP server wraps the results in SKOS Core and returns them to the SRW server.
5. The SRW server sends the results back to the client service; results comprise terms together with information about each term's source scheme, notation (DDC) or ID (other schemes), and broader, narrower and related terms, where applicable.
6. The client service processes the results to provide the service interface with terms from specific schemes relevant to the query and with any relevant additional data on the terms (e.g. related terms).

To illustrate the functionality of the get_filtered_set function, 'Plant genetics' will be searched for, selecting HASSET as the preferred scheme to be searched. Results are detailed in Figure 9:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-SOAP envelope -->
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core.rdf#"
xml:base="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/concepts.php">
<skos:Concept rdf:about="#2465">
<skos:prefLabel xml:lang="en">GENETICALLY MODIFIED CROPS
</skos:prefLabel>
<skos:broader rdf:resource="#1389"/>
<skos:broader rdf:resource="#2463"/>
<skos:related rdf:resource="#110"/>
<skos:related rdf:resource="#2466"/>
<skos:related rdf:resource="#4712"/>
<skos:altLabel xml:lang="en">GM CROPS</skos:altLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#4712">
```

```
<skos:prefLabel xml:lang="en">PLANT GENETICS</skos:prefLabel>
<skos:broader rdf:resource="#624"/>
<skos:broader rdf:resource="#2467"/>
<skos:related rdf:resource="#2465"/>
<skos:altLabel xml:lang="en">PLANT BREEDING</skos:altLabel>
<skos:altLabel xml:lang="en">PLANT REPRODUCTION</skos:altLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#1389">
<skos:prefLabel xml:lang="en">CROPS</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#2463">
<skos:prefLabel xml:lang="en">GENETIC ENGINEERING
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#110">
<skos:prefLabel xml:lang="en">AGRICULTURAL PRODUCTION
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#2466">
<skos:prefLabel xml:lang="en">GENETICALLY MODIFIED FOOD
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#4712">
<skos:prefLabel xml:lang="en">PLANT GENETICS</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#624">
<skos:prefLabel xml:lang="en">BOTANY</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#2467">
<skos:prefLabel xml:lang="en">GENETICS</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
```

```
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
</rdf:RDF>
```

Fig. 9. Results for a get_filtered_set query (set to HASSET) for the term 'Plant genetics'. Result in SKOS RDF/XML.

Figure 9 shows how HILT can provide extremely specialised terminological data, in this case from a single scheme selected using the get_filtered_set function. Any individual scheme or any combination of schemes within the system can be accessed in this way. Further, a user / client service can specify whether they wish to retrieve preferred, non-preferred or related terms within a search for terms in any scheme(s). Such terminological data can be used in a variety of ways; however, it is expected that get_filtered_set will be used most by those services wishing to extend the retrieval tools available to users. For example, using get_filtered_set to implement forms of interactive query expansion or hierarchical scheme browsing to improve local repository interrogation or to aid query formulation. GoGeo! has implemented a keyword search demonstrator employing HILT get_filtered_set functionality [63]. This provides a real-life example of how HILT could be integrated within an existing service in order to mediate searching of associated collections and as a means of providing query expansion opportunities for users.

The SKOS result in Figure 9 indicates that searching for 'Plant genetics' within HASSET retrieves terms including 'Genetically modified crops', 'GM crops', 'Plant genetics', 'Plant breeding', 'Plant reproduction', 'Crops', 'Genetic engineering', 'Agricultural production', 'Genetically modified food', 'Botany', 'Genetics'. Not all of these are likely to be directly relevant to a user requesting information on 'Plant genetics'. The current search parameters within HILT first search for an exact phrase match i.e. Boolean AND; thereafter conducting further searches in line with the Boolean OR principle. It follows that single terms within the search query 'Plant' and 'Genetics' are retrieved individually, which may or may not prove relevant in every instance.

5.6 Function summary

HILT currently enables users / client services to retrieve DDC only terms, non-DDC only terms, a combination of both DDC and non-DDC terms, or to specify an individual scheme or a selection of schemes, which they wish to search. In the latter case, functionality also extends to the switching on or off of preferred, non-preferred or related terms, enabling yet greater search specificity.

The perceived effectiveness or otherwise of four out of the five functions (excluding get_collections) is heavily reliant on inter-scheme mapping. Results for get_ddc_records, get_non_ddc_records, get_all_records and get_filtered_set, where more than one scheme is selected, is dependent upon an effective mapping infrastructure. It is therefore necessary to ensure valid and robust mappings are implemented. Such mappings should also be complete. That is, one-to-one mappings are likely to be insufficient for the types of scenarios presented above, even between one individual scheme and another.

6 Discussion

The preceding examples relating to each of HILT's five functions currently implemented indicate that the system does indeed have the potential to improve distributed information retrieval where different services/collections employ disparate terminologies.

The classification of services/collections by DDC enables the get_collections function to retrieve details of services holding resources covering the user's chosen subject area. One limitation of this function in its current instantiation is that within the local collections database searched by HILT, each service/collection is only assigned one DDC number. For general and multidisciplinary collections it would be pertinent to extend this to as many DDC numbers as required to convey subject coverage adequately. This would facilitate the retrieval of collections, with only a subset of items relevant to a user's needs. It is thought that the assignation of multiple class numbers in this way would greatly enhance the get_collections function by opening up more potentially relevant information sources to the user. It should be noted that IESR [62] already offers multiple DDC numbers for any given collection.

A further limitation relates to the process of truncation implemented. The example in 5.1 above shows that a search for 371.07 will retrieve collections classified at 370 but nothing beyond that. It is proposed to extend the process of truncation beyond the decimal point so that a general collection will be returned if nothing more specifically relevant is returned. The retrieval of a general social science collection classified at 300, for example, is considered to have greater value to the user than a scenario where they retrieve no hits. By extending truncation beyond the decimal point users will retrieve collections classified at one of the ten main DDC classes.

In some of the current examples, scheme information is missing from the DC CD AP result returned. It should be noted that this is due to incomplete information within the collections database. This issue should be ameliorated with the incorporation of relevant collection and service registries to the HILT model, as noted in 5.1. In line with the architecture of the JISC Information Environment, it is intended that the collections database ultimately be maintained externally and independently by the IESR [62].

The additional four functions described in section 5 illustrate how users can retrieve exact matches for terms across schemes, synonyms or concept matches, along with broader or narrower terms. Such functionality will aid improved retrieval performance for users by lowering the cognitive load experienced by the user during query formulation [64]. Where in general search engines a user may retrieve no directly relevant hits, or relevant hits may be buried a considerable way down a long results list, HILT provides alternative search terms with a view to expanding users' queries, and where no exact or concept matches exist, related terms in the form of more general, more specific and so on will be presented.

The dynamic element of the system, whereby selected terms trigger a search within a relevant collection, further improves the level of information retrieval for the user. This process miminises the number of clicks and limits the need for the user to re-enter search terms into a number of different services' search boxes.

The success of these types of functions is heavily reliant on the appropriateness of mappings implemented, as well as the accuracy of repository resource indexing (particularly in distributed subject resource discovery contexts). Users will only benefit from the retrieval of synonyms and the like if they have been correctly identified and encoded as such within the HILT model. The cost and time consuming nature of implementing mappings has already been discussed. Due to such constraints, HILT proposes to first consider a fairly broad set of mappings, likely to be imposed between satellite schemes and DDC's top 1000 captions, or most frequently used numbers, before piloting an area of more in-depth mapping within a more detailed subject area. This work is likely to inform how to proceed with fuller-scale mapping exercises. It is hoped that patterns will emerge to enable some degree of automation to be implemented, although manual verification of the appropriateness or otherwise of relationships will still be required. It will also be necessary to review existing mappings within the current instantiation. OCLC provided an XML version of DDC 22 with mappings to LCSH, many of which appear inappropriate for the purpose of HILT. Function testing has revealed that many of the DDC-LCSH mappings are not considered of potential benefit to users retrieving information from distributed sources. This may be a result of such mappings having been derived statistically.

Progression towards a more precise system depends on refinement of search parameters. Results sets presented in section 5, particularly that for the get_filtered_set function, indicate that fairly imprecise results are currently being retrieved due to the broad nature of the current search parameter. It is thought likely that this will require refinement, perhaps to only search using Boolean AND in the first instance. The OR operator could potentially be invoked if requested by the user. This will maintain transparency enabling the user to keep track of the results provided. Otherwise, some of the terms returned may not appear directly relevant to the user's search, giving the impression of an ineffective system.

It is considered of interest to investigate the suitability of other universal schemes with a view to replacing DDC as a spine, although the full extent of the advantages of using DDC have not yet been fully explored. HILT will continue to work with DDC, whilst considering how alternatives may improve or degrade the level of success for the user in relation to the functions implemented.

The range of schemes incorporated into the current HILT model should clearly be reviewed and extended as necessary. The selection of schemes was originally purposive since the project largely depended on those schemes it could obtain free of charge for research purposes and in a suitable format for uploading into a terminologies database with minimal intervention. Depending on the nature of HILT's growth, and the community it requires to serve, the inclusion of schemes will be heavily modified. It is also of interest to incorporate folksonomies into the HILT model. The inclusion of folksonomies, or folksonomy-type terms is likely to create a range of additional access points for users unfamiliar with formal terminology used to express certain concepts. Less formal terms in everyday usage could be mapped to the DDC spine in the same way as standard schemes and it is possible that tag clouds characteristic of Web 2.0 folksonomy driven services could have a role to play in the expression of synonymous concepts, as well as broader and narrower equivalence relationships. HILT has done some preliminary work in incorporating user terms taken from search logs, to ascertain whether or not this improves the hit rate for users following the translation process afforded by mapping such terms to DDC, which can then, in turn, be translated to any other scheme providing relevant subject coverage. Folksonomies or folksonomy-type terms are likely to be incorporated as research proceeds, in addition to mappings being established from the standards schemes included.

The validity of an ontological approach to developing a terminology server is also of interest. Sanchez-Alonso and Garcia-Barriocanal [65] investigated the feasibility of mapping SKOS Core metadata to an upper ontology. Various difficulties were encountered as a result of the lack of formalisation in the current instantiation of SKOS and the need for mapping criteria to promote semantic interoperability. The authors endeavour to find a way "to map a concept in a SKOS scheme to a term in an upper ontology that provides a formal definition". Their investigation found that an intermediate model was required to do so. At present, there is no immediate remit to pursue this type of approach within HILT, although the progress of others working in the area will be followed with interest.

For the purpose of creating further and more advanced functionality within the system, it will first be necessary to survey the JISC community to determine the types of features they would find useful in a system such as HILT. Such a survey is planned for the current phase of the project and is likely to inform the design of additional functions. User evaluation is also necessary to assess the appropriateness and usefulness of such functions. The functions already described in the current paper will also be assessed by users in the near future.

7 Conclusion and further research

Some areas for future research were discussed in the previous section. In addition to these, further research into match types should be conducted to establish how best to express the nature of equivalence relationships between terms. Currently, five mapping types are in use, in line with the SKOS MVS. These are exact match, narrow match, broad match, major match and minor match. It is thought likely that further match types may prove useful although this theory must be considered in the context of user testing.

It is considered likely that a range of additional use cases, and therefore functions, will prove valuable within the HILT service. A survey of potential users of HILT (both services/collections and individuals) should be undertaken to inform the HILT team on what these use cases might be. Appropriate functionality can then be designed and built in to the system.

To assess the more robust measures of retrieval, precision and recall, precision being the proportion of relevant documents retrieved within the retrieved set and recall being the proportion of relevant documents retrieved from the total number of relevant documents available, rigorous testing is required within a controlled environment. It is necessary to build a document collection and run robust tests in order to assess such measures of success.

In conclusion, effective resource discovery can only be realised if the means of access becomes more transparent. If users are unable to locate relevant resources on the web due to lack of awareness and openness, the success of digital publishing is compromised. Users require to be made aware of the existence of resources relevant to their needs and require metadata to be sufficiently penetrable to conduct effective and efficient information retrieval. In an environment where subject metadata varies from collection to collection or service to service, in an increasingly fragmented digital world, such efficiency cannot be realised. Terminologies need to be brought together to improve interoperability between services, thus making disparate collections cross-searchable. It is the authors' belief that a system like HILT can go some way to improving the openness of resources and therefore widening access to material held in heterogeneous collections across the web, which would otherwise be hidden, and that HILT's architecture and mapping based infrastructure will, in time, prove an efficient means of reaching this goal.

References

1. McCulloch, E., Shiri, A., Nicholson, D.: Challenges and issues in terminology mapping: a digital library perspective. The Electronic Library. Vol. 23 No. 6 (2005) 671-677
2. Open Archives Forum: Breakout Session [online]. Lisbon. (2002) 19 Available at: http://www.oaforum.org/otherfiles/oaf_d43_workshop2.pdf [cited 31 August 2007]
3. Chan, L., Zeng, M.: Ensuring Interoperability among Subject Vocabularies and Knowledge Organization Schemes: a Methodological Analysis [online]. 68th IFLA Council and General Conference, (2002) Glasgow. Available at: http://www.ifla.org/IV/ifla68/papers/008-122e.pdf [cited 31 August 2007]
4. Doerr, M.: Semantic problems of thesaurus mapping [online]. Journal of Digital Information. Vol. 1 No. 8 (2001) Available at: http://jodi.tamu.edu/Articles/v01/i08/Doerr/ [cited 31 August 2007]
5. Koch, T., Neuroth, H., Day, M.: Renardus: Cross-browsing European subject gateway via a common classification system (DDC). In: McIlwaine, I., C. (ed.):

Proceedings of the IFLA Satellite Meeting Held in Dublin, Ohio, 14-16 August 2001 and Sponsored by the IFLA Classification and Indexing Section, the IFLA Information Technology Section and OCLC. K. G. Saur, Munchen (2003) 25-33

6. SRW/SRU. Information available at: http://www.loc.gov/standards/sru/ [cited 31 August 2007]
7. JISC. JISC Information Environment. Information available at: http://www.jisc.ac.uk/whatwedo/themes/information_environment.aspx [cited 31 August 2007]
8. Miles, A., Brickley, D.: (eds). SKOS Core guide: W3C working draft 2 November. (2005) Available at: http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/ [cited 29 August 2007]
9. OCLC. DDC 22 [online]. Available via OCLC Connexion: http://connexion.oclc.org [cited 31 August 2007]
10. Nicholson, D., Dawson, A., Shiri, A.: HILT: A pilot terminology mapping service with a DDC spine. Cataloging & Classification Quarterly. Vol. 42 No. 3/4 (2006) 187-200
11. Coates, E. J.: Switching languages for indexing. Journal of Documentation. Vol. 26 No. 2 (1970) 102-110
12. Horsnell, V.: The Intermediate Lexicon: an aid to international co-operation. Aslib Proceedings. Vol. 27 No. 2 (1975) 57-66
13. Koch, T.: Desire project handbook: 2, 5 subject classification, browsing and searching [online]. Available at: http://www.desire.org/handbook/2-5.html [cited 29 August 2007]
14. NISO (National Information Standards Organization), Report on the Workshop on Electronic Thesauri, November 4-5, 1999. Available at http://www.niso.org/news/events-workshops/thes99rprt.html [cited 20 July 2007]
15. Macgregor, G., Joseph, A., Nicholson, D.: A SKOS Core approach to implementing an M2M terminology mapping server. International Conference on Semantic Web and Digital Libraries (ICSD-2007 Proceedings of the), 21-23 February. Bangalore, India. Bangalore: Documentation Research & Training Centre, Indian Statistical Institute (2007) 109-120 Available at: http://eprints.cdlr.strath.ac.uk/2970/ [cited 29 August 2007]
16. J. Paul Getty Trust, Art and Architecture Thesaurus Online. Available at: http://www.getty.edu/research/conducting_research/vocabularies/aat/ [cited 31 August 2007]
17. GCMD, Global Change Master Directory. Available at: http://gcmd.nasa.gov/ [cited 31 August 2007]
18. HASSET, Humanities and Social Science Electronic Thesaurus. Available at: http://www.data-archive.ac.uk/search/hassetSearch.asp [cited 31 August 2007]
19. IPSV, Integrated Public Sector Vocabulary. Available at: http://www.esd.org.uk/standards/ipsv/ [cited 31 August 2007]
20. JACS, Joint Academic Coding System. Available at: http://www.hesa.ac.uk/jacs/jacs.htm [cited 31 August 2007]
21. JITA Available at: http://eprints.rclis.org/jita.html [cited 31 August 2007]
22. LCSH, Library of Congress Subject Headings. Available at: http://www.loc.gov/cds/lcsh.html [cited 31 August 2007]
23. MeSH, Medical Subject Headings, Available at: http://www.nlm.nih.gov/mesh/ [cited 31 August 2007]

24. NMR, National Monuments Record, Available at: http://thesaurus.english-heritage.org.uk/ [cited 31 August 2007]
25. SCAS, Standard Classification of Academic Subjects. Available at: http://www.ucas.com/higher/courses/scascode.pdf [cited 29 August 2007]
26. UNESCO Thesaurus, Available at: http://www2.ulcc.ac.uk/unesco/ [cited 20 July 2007]
27. CAB Thesaurus, Available at: http://www.cabi.org/DatabaseSearchTools.asp?PID=277 [cited 20 July 2007]
28. Iyer, H. and Giguere, M.: Towards designing an expert system to map mathematics classificatory structure. Knowledge Organization. Vol. 22 No. 3/4 (1995) 141-147
29. McCulloch, E., Macgregor, G.: Analysis of equivalence mapping for terminology services. Journal of Information Science. Vol. 33 No. 5 (2007)
30. Chaplan, M. A.: Mapping Laborline Thesaurus terms to Library of Congress Subject Headings: implications for vocabulary switching. Library Quarterly. Vol. 56 No. 1 (1995) 39-61
31. Dolin, Robert, H., Mattison, John, E., Cohn, S., Campbell, Keith, E., Wiesenthal, Andrew, M., Hochhalter, B., LaBerge, D., Barsoum, R., Shalby, J., Abilla, A., Clements, Robert, J., Correia, Carol, M., Esteva, D., Fedack, John, M., Goldbert, Bruce, J., Gopalarao, S., Hafeza, E., Hendler, P., Hernandez, E., Kamangar, R., Khan, Rafique, A., Kurtovich, G., Lazzareschi, G., Lee, Moon, H., Lee, T., Levy, D., Lukoff, Jonathan, Y., Lundbert, C., Madden, Michael, P., Ngo, Trongtu, L., Nguyen, Ben, T., Patel, Nikhilkumar, P., Resneck, J., Ross, David, E., Schwarz, Kathleen, M., Selhorst, Charles, C., Snyder, A., Umarji, Mohamed, I., Vilner, M., Zer-Chen, R., Zingo, C.: Kaiser Permanente's Convergent Medical Terminology. MEDINFO. Vol. 11 No. 1 (2004) 346-50 Available at: http://square.umin.ac.jp/DMIESemi/y2004/20041129_3.pdf [cited 31 August 2007]
32. Miles, A., Brickley, D.: (eds). SKOS Mapping Vocabulary Specification. (2004). Available at: http://www.w3.org/2004/02/skos/mapping/spec/ [cited 29 August 2007]
33. Liang, A., Sini, M., Chun, C., Li, S. J., Lu, W. L., He, C. P., Keizer, J.: The mapping schema from Chinese Agricultural Thesaurus to AGROVOC, 6th Agricultural Ontology Service (AOS) Workshop on Ontologies: the more practical issues and experiences, July 25-28, Vila Real, Portugal, 2005 (Food and Agriculture Organization, Rome, 2005). Available at: ftp://ftp.fao.org/docrep/fao/008/af241e/af241e00.pdf [cited 31 August 2007]
34. LexGrid, The Lexical Grid: Shared Terminology Resources. Available at: http://informatics.mayo.edu/LexGrid/index.php?page=aboutlg [cited 31 August 2007]
35. OCLC Terminologies service. Available at: http://www.oclc.org/terminologies/default.htm [cited 31 August 2007]
36. Doerr, M., Fundulaki, I.,: The Aquarelle Terminology Service, ERCIM News, 1998, no. 33. Available at: http://www.ercim.org/publication/Ercim_News/enw33/doerr2.html [cited 31 August 2007]
37. Christophides, V., Doerr, M., Fundulaki, I.: The Aquarelle Folder Server, ERCIM News, no. 33. Available at: http://www.ercim.org/publication/Ercim_News/enw33/doerr1.html[cited 31 August 2007]

38. OpenGALEN, Information available at: http://www.opengalen.org/faq/faq5.html [cited 31 August 2007]
39. Chute, C. G., Elkin, P. L., Sheretz, D. D., Tuttle, M., S.: Desiderata for a Clinical Terminology Server. American Medical Informatics Association. Available at: http://www.amia.org/pubs/symposia/D005782.PDF [cited 31 August 2007]
40. Renardus, Available at: http://www.renardus.org/ [cited 20 July 2007]
41. Svenonius, E.: Use of Classification in Online Retrieval. Library Resources and Technical Services. Vol. 27 No. 1 (1983) 76-80
42. Bowman, J., H.: Essential Dewey. Facet Publishing London (2005) 15
43. UDC Consortium, Available at: http://www.udcc.org/ [cited 31 August 2007]
44. Lloyd, G., A.: The Universal Decimal Classification as an International Switching Language. International symposium on UDC in relation to other indexing languages. Herceg Novi, Yugoslavia, June 28-July 1 (1971)
45. Balikova, M.: Multilingual Subject Access to catalogues of National Libraries (MSAC): Czech Republic's collaborations with Slovakia, Slovenia, Croatia, Macedonia, Lithuania and Latvia. In: Proceedings of the World Library and Information Congress: 71st IFLA General Conference and Council - Classification and indexing with cataloguing, Oslo, Norway, August 14-18 (2005) (IFLA, The Hague, 2005) Available at: http://www.ifla.org/IV/ifla71/papers/044e-Balikova.pdf [cited 30 August 2007]
46. W3C.: Semantic Web Deployment Working Group. (2007) Available at: http://www.w3.org/2006/07/SWD/ [cited 29 August 2007]
47. Vizine-Goetz, D., Houghton, A., Childress, E.,: Web services for controlled vocabularies. Bulletin of the American Society for Information Science and Technology, Vol. 32 No. 5 (2006) Available at: http://www.asis.org/Bulletin/Jun-06/vizine-goetz_houghton_childress.html [cited 29 August 2007]
48. Tudhope, D., Binding, C.: Toward terminology services: experiences with a pilot web service thesaurus browser. Bulletin of the American Society for Information Science and Technology, Vol. 32 No. 5 (2006) Available at: http://www.asis.org/Bulletin/Jun-06/tudhope_binding.html [cited 29 August 2007]
49. Nicholson, D., McCulloch, E.: Investigating the feasibility of a distributed, mapping-based, approach to solving subject interoperability problems in a multi-scheme, cross-service, retrieval environment. Proceedings of International Conference on Digital Libraries, 5-8 December, New Delhi, India. (2006) Available at: http://eprints.cdlr.strath.ac.uk/2875/ [cited 29 August 2007]
50. Svensson, Lars. G.: National libraries and the Semantic Web: requirements and applications. Proceedings of the International Conference on Semantic Web and Digital Libraries, Documentation Research and Training Centre, Bangalore, India. (2007) 101-108
51. W3C.: OWL Web ontology language guide, W3C, Massachusetts Institute of Technology, European Research Consortium for Informatics and Mathematics, Keio University. (2004) Available at: http://www.w3.org/TR/owl-guide/ [cited 29 August 2007]
52. Zhao, Y.: Combining RDF and OWL with SOAP for Semantic Web Services. Proceedings of the 3rd annual Nordic Conference on Web Services (NCWS'04), Vxj, Sweden 22-23 Nov (2004) 31-45 Available at: http://www.ida.liu.se/ yuxzh/doc/ncws-041002.pdf [cited 29 August 2007]

53. Liang, A., C., Sini, M.: Mapping AGROVOC and the Chinese Agricultural Thesaurus: Definitions, tools, procedures. New Review in Hypermedia and Multimedia. Vol. 12 No. 1 (2006) 51-62
54. Miles, A., Matthews, B., Wilson, M., Brickley, D.: SKOS Core: Simple knowledge organization for the Web. Proceedings of the International Conference on Dublin Core and Metadata Applications (DC-2005), Madrid, Spain, 12-15 Sep (2005) Available at: http://isegserv.itd.rl.ac.uk/public/skos/press/dc2005/dc2005skospaper.pdf [cited 29 August 2007]
55. Miles, A., Matthews, B., Beckett, D., Brickley, D., Wilson, M., Rogers, N.: SKOS: A language to describe simple knowledge structures for the web. Proceedings of XTech 2005: XML, the Web and beyond, Idealliance, Amsterdam, Netherlands (2005) Available at: http://www.idealliance.org/proceedings/xtech05/papers/03-04-01/ [cited 29 August 2007]
56. Zthes.: The Zthes specifications for thesaurus representation, access and navigation (2006) Available at: http://zthes.z3950.org/ [cited 29 August 2007]
57. Vizine-Goetz, D., Hickey, C., Houghton, A., Thompson, R.: Vocabulary Mapping for Terminology Services. Journal of Digital Information, (2004) Vol. 4 No. 4 Available at: http://jodi.tamu.edu/Articles/v04/i04/Vizine-Goetz/ [cited 31 August 2007]
58. Will, Leonard. RE: Zthes and DDC. Zthes - Development of the Zthes model for thesauri (mailing list), Index Data, Denmark (2005) Available at: http://lists.indexdata.dk/pipermail/zthes/2005-February/000020.html [cited 29 August 2007]
59. Nicholson, D., McCulloch, E.: Interoperable subject retrieval in a distributed multi-scheme environment: new developments in the HILT project. Ibersid, Zaragoza, Spain 2-4 Nov (2005) Available at: http://eprints.cdlr.strath.ac.uk/2317/01/Nicholson_ZaragosaPaperFinal.pdf [cited 31 August 2007]
60. HILT: M2M Final Report. Appendix D: Assessment: Use Cases, Protocols and Mark-ups. (2005) Available at: http://hilt.cdlr.strath.ac.uk/hiltm2mfs/0HILTM2MFinalReportRepV3.1.pdf [cited 29 August 2007]
61. HILT: Demonstrator. (2006) Available at: http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/hiltsoapclient.php [cited 29 August 2007]
62. IESR: Internet Environment Services Registry. Available at: http://iesr.ac.uk/ [cited 31 August 2007]
63. Go-Geo! Demonstrator. (2006) Available at: http://nevis.ed.ac.uk:9200/gogeo-hilt2.html [cited 20 July 2007]
64. Ethimiadis, E.,N.: Interactive query expansion: a user-based evaluation in a relevance feedback environment. Journal of the American Society for Information Science. Vol. 51 No. 11 (2000) 989-1003
65. Sanchez-Alonso, S., Garcia, E.: Making use of upper ontologies to foster interoperability between SKOS concept schemes. Online Information Review. Vol. 30 No. 3 (2006) 263-277

Analysis of equivalence mapping for terminology services

Emma McCulloch and George Macgregor

Centre for Digital Library Research, Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK

Abstract.

This paper assesses the range of equivalence or mapping types required to facilitate interoperability in the context of a distributed terminology server. A detailed set of mapping types were examined, with a view to determining their validity for characterizing relationships between mappings from selected terminologies (AAT, LCSH, MeSH, and UNESCO) to the Dewey Decimal Classification (DDC) scheme. It was hypothesized that the detailed set of 19 match types proposed by Chaplan in 1995 is unnecessary in this context and that they could be reduced to a less detailed conceptually-based set. Results from an extensive mapping exercise support the main hypothesis and a generic suite of match types are proposed, although doubt remains over the current adequacy of the developing Simple Knowledge Organization System (SKOS) Core Mapping Vocabulary Specification (MVS) for inter-terminology mapping.

Keywords: classification; interoperability; knowledge organization systems; SKOS Core; term equivalence; terminologies; vocabulary mapping

1. Introduction

The recent growth in distributed digital libraries and repositories has restored interest in the interoperability of knowledge organization systems (KOS) to facilitate user access to discrete heterogeneous digital objects [1]. KOS employ a variety of disparate *terminologies* in the form of term lists (e.g. authority files, glossaries, gazetteers, dictionaries), classifications and categorization schemes (e.g. bibliographic classifications, taxonomies) and relational vocabularies (e.g. thesauri, subject heading lists, semantic networks, ontologies) [2].

Within the growing number of repositories, digital objects are indexed and organized in accordance with a variety of different schemes. Since it is unrealistic to expect users to search each

Correspondence to: Emma McCulloch, Centre for Digital Library Research, Department of Computer and Information Sciences, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow G1 1XH. UK. Email: e.mcculloch@strath.ac.uk

repository separately or to familiarize themselves with the numerous terminologies deployed, it is increasingly important that users are able to search or browse multiple distributed repositories simultaneously. Currently, however, the effectiveness of such systems tends to remain dependent upon the degree of interoperability afforded between the terminologies in use. Technical approaches seeking to artificially or intellectually optimize interoperability therefore continue to form a key area of research [e.g. 2–5]. One such approach that has attracted significant attention is *terminology mapping* (or vocabulary mapping).

Terminology mapping is evident in a variety of KOS interoperability approaches and essentially involves imposing equivalence, conceptual and hierarchical relationships between terms in different schemes [4]. The assumption underpinning mapping is that equivalence can exist between disparate KOS and their respective terminologies [4]; however, exact equivalence is rarely attainable [6] due to the complexities inherent in natural language.

Whilst recent research into the application of automated techniques has aided in the management of large terminology sets and even assisted in mapping implementation itself [7], the process of terminology mapping remains largely intellectual, and therefore heavily dependent on human intervention. One continuing problem inherent in the terminology mapping process – whether intellectual or automated – is accurately characterizing the type of mapping match between terms. The existence of linguistic inconsistencies across terminologies (e.g. synonyms, homonyms, antonyms, etc.), grammatical variations (e.g. singular/plural forms, alternative spellings or punctuation, verb tenses, etc.), variations in subject coverage, and the relative specificity or level of granularity with which terminologies accommodate like concepts, limits their potential for exact equivalence. Differing structures of the terminologies being mapped can also prove problematic for mapping across different KOS; for example, classification schemes can have radically different structures to relational vocabularies. Consequently, mapped terms may only exemplify partial equivalence.

Given that exact equivalence between terminologies will be rare, it is necessary to accurately characterize the degree of equivalence by assigning match types during the mapping process. This is considered necessary to:

- Enable the ranking of results according to the degree of concordance with users' preferred terminology. For example, it is considered likely that an exact match will be more relevant to a user query than an inexact match, whether due to a spelling variation, part-of-speech difference and so on, since it is closer to the term originally sought by the user. A user may search for 'tooth'; it is likely that matches for 'tooth' will be more relevant to the user than those for 'teeth', and so should be ranked more highly in the results set presented.
- Provide users with details of the precise nature of the relationship(s) between their entered query and their retrieved result set (which will invariably include mapped terms from other terminologies, or comprise resources retrieved using terms derived from mapped terminologies).
- Impart sufficient information during subject hierarchy browsing to enable users to make informed decisions about the relevance of mapped terms.
- Provide users with mappings that can be used to generate relevance feedback.
- Help identify mapping regularities between specific terminologies, thus facilitating the research and development of improved automated routines to assist in large-scale terminology mapping.

Various match types have been proposed, e.g. [3, 4, 7–10]. In this paper we examine terminology mapping match types in relation to a Dewey Decimal Classification (DDC)-based terminology server. In particular, we assess the suitability of Chaplan's 19 match types [8] as forming the basis of a *generic suite* of equivalence matches to be used by services employing terminology mapping. Chaplan's investigation constitutes one of the most thorough pieces of research in this area and, as such, the 19 mapping types presented in her paper provide a concrete basis for the current study.

The remainder of this paper is organized as follows: Sections 2, 3 and 4 will review related literature and establish the aims of the study. Section 5 describes the methodology used to test the Chaplan match types in relation to our data set. The crux of the paper (Sections 6 and 7) deals with

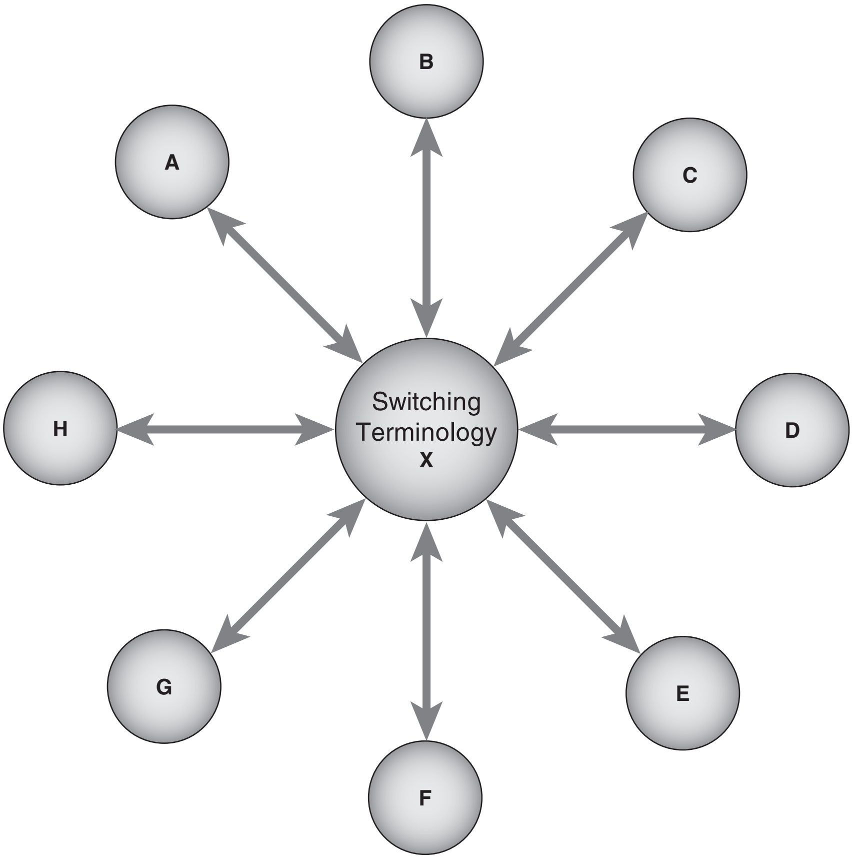


Fig. 1. Typical terminology switching model.

the results, analysis and subsequent discussion. Conclusions and suggestions for further research are provided in Section 8.

2. Terminology mapping

Interest in mapping as a means of facilitating terminology interoperability for improved distributed searching is not new. The rapid development of distributed online databases in the 1960s and 1970s, and the associated rise in domain-specific terminologies, forced researchers to address the issue of terminology compatibility and related system-based solutions, with terminology mapping featuring significantly in many of the proposed solutions. Although so-called 'direct mapping' was popular and continues to be for some recent solutions [11, 12], it generally requires considerable intellectual effort and resourcing [2, 13]. Mapping work has consequently focussed on the use of terminology *switching* (or 'switching languages') to simplify the management of multiple terminological mappings and to minimize the intellectual demands normally associated with direct mapping.

The switching model entails the use of a single terminology as an intermediary (Figure 1). Each of the terminologies to be used in the retrieval system (*A*–*H*) is mapped to a common terminology (*X*). This allows user queries entered using terminology *A* to be translated to *X* and then switched to the equivalent terms in terminology *G*, for example. Switching was often the chosen model for early mapping research [14–16] and has recently been revisited [3, 17].

Any terminology can theoretically be used for switching although it is generally acknowledged that the coverage of the selected switching terminology must be sufficiently broad to include most, if not all, disciplines encompassed by the schemes with which it is to be used [15, 18]. Failure to select such a broad terminology will result in the switching language degrading any requests sent from a detailed terminology (e.g. a domain-specific terminology). The use of universal classification schemes for switching has therefore attracted much attention, particularly those schemes that are notation-based. The theoretical foundation for using such terminologies was established early on [19, 20] and has subsequently been explored by a variety of recent research projects, e.g. [3, 21, 22]. Schemes such as the Universal Decimal Classification (UDC) and the Dewey Decimal Classification (DDC) exemplify wide subject coverage and are suitable for multi-disciplinary user groups. They also experience global use; DDC, in particular, is available via the subscription-based OCLC Connexion service, providing quick access to frequently updated schedules and thus aiding the mapping and management of terminologies [3]. The analytico-synthetic features, although more prominent in UDC than DDC, allow the construction of diverse and detailed concepts which can then be expressed in hierarchical notation [23]. More generally, such schemes are conducive to hierarchical browsing and facilitate the display of associated thesaurus terms [2].

The match type work documented in this paper is based within the context of a DDC spine-based terminology server [24]. The spine-based approach is conceptually similar to switching but differs in that DDC does not always assume a passive role in the switching process. Rather, DDC remains central to users' *disambiguation* processes [25]. Since DDC notation is generally indicative of the taxonomic hierarchy, the truncation of DDC numbers is initiated if no hits are found using mapped terminologies. This truncation occurs successively until one or more hits are identified. For example, if a user query for 'greenhouse gases' was matched to 363.73874 and no hits were found, the system would truncate the DDC number as follows:

363.7387 (Fumes, gases, smoke)

363.738 (Pollutants)

363.73 (Pollution)

The spine-based approach also facilitates hierarchical browsing and the discovery of like terms within other terminologies. The current machine to machine (M2M) web-service implementation of this server provides a variety of terminological functions, such as the enrichment of users' search queries by providing (where applicable) related terms (RT), broader terms (BT), narrower terms (NT), scope notes, etc. associated with specific, named, terminologies (which are then marked up in the Simple Knowledge Organization System (SKOS) Core and sent to local systems for use) [26]. However, the primary function of the server remains mapping between disparate terminologies within a variety of different user searching scenarios.

3. Match types: related work

Several investigations into mapping match types have been conducted over many years and have largely arisen as a result of research into terminology mapping; however, investigations into terminology compatibility and integration have also been successful in defining degrees of equivalence. For example, Neville [9] studied the types of incompatibility between keywords in different thesauri pertaining to the same subject area for possible thesauri reconciliation based on a source thesaurus using concept code numbering. Although he identified numerous types of relationships and proposed some solutions for accommodating those which were complex, many have limited applicability within an operational mapping system.

While researching and developing an operational terminology switching system, Silvester and Klingbiel [10] developed a series of rules to accommodate switching between the Defense Technical Information Center (DTIC) subject terms and the National Aeronautics and Space Administration (NASA) thesaurus by using a so-called 'lexical dictionary'. These rules were established as system commands, but characterized the degree of term equivalence between terminologies: with delete, identity, simple change, list, and table. 'Delete' indicates that there are no conceptually equivalent terms in the NASA thesaurus (i.e. no match between input and output terms). 'Identity' indicates that the input is identical or equal to the output (i.e. an exact match between input and output terms). 'Simple change' indicates a 'minor' change and characterizes instances where the input term expresses the same concept as the output but differs in minor respects. For example, the input term may be plural and the output term singular. This rule also accommodates other grammatical variations such as suffix variations (e.g. 'ing' or 'tion') and synonyms. 'List' applies when a single term is switched to multiple terms. In such instances the multiple terms express the same concept (e.g. Machmeters/Mach number, speed indicators). Finally, 'table' indicates the occurrence of 'tables'. That is, the input term is context sensitive and requires additional terms to clarify the concept.

Iyer and Giguere [27] investigated a prototype expert system interface enabling mathematicians to discover library resources classified according to DDC, but using the American Mathematic Society (AMS) Mathematics Subject Classification (MSC). Their expert system mapped the MSC scheme to 510 (Mathematics) in DDC. Although similarities were noted in the divisions of mathematics covered by both schemes, both differed significantly in their level of specificity and emphasis. Iyer and Giguere reconciled these differences by analysing both schemes to identify a series of 'mapping strategies': exact matches, specific to general, general to specific, many to one, cyclic mapping, no matches, and specific and broad class mapping. 'Exact matches' were considered instances whereby the MSC scheme has a corresponding DDC number at a similar level of specificity. 'Specific to general' accommodated those MSC concepts that are overly specific for DDC and were mapped to the nearest broader DDC class. 'General to specific' is the reverse of 'specific to general' and characterized instances where DDC exemplifies greater specificity than MSC. 'Many to one' matches were those where acknowledged subclasses of a particular MSC concept were distributed across several locations of the semantic hierarchy, whilst in DDC they are located within one class, thus necessitating several mappings to the same DDC number. 'Cyclic mapping' accommodated instances where a DDC number has a broader scope than a similar number in MSC. For example, 'Vector and tensor analysis' (53A45 MSC) was mapped to a DDC number representing vector, tensor *and* spinor analysis (515.63). 'Spinor analysis' (53A50 MSC) was therefore also mapped to 515.63. 'No matches' constituted instances whereby those areas of the mathematics discipline comprehensively treated in MSC have absolutely no equivalence in DDC. Finally, Iyer and Giguere considered 'specific and broad class mapping' to refer to instances in which primary divisions in MSC may map satisfactorily, but where further specificity can not be accommodated in the target scheme (DDC), thus forcing a 'broad class' mapping.

While researching the possibilities for a multilingual thesaurus, Riesthuis [28] investigated the use of six forms of equivalence to facilitate a model designed to aid cross-lingual retrieval. Riesthuis' model used 'partial equivalence', 'loan terms', 'inexact equivalence', 'single-to-multiple' and 'non-equivalence'. Owing to the fact that Riesthuis was exploring equivalence in a multilingual environment, several of the match types accommodate relationships that would not necessarily occur in a monolingual environment. 'Partial equivalence' denotes instances whereby a term in languages A and B of the same thesaurus are not satisfactorily equivalent; however, a degree of overlap is found. By way of example, Riesthuis notes that the Dutch word 'record' has a narrower sense than the English word; Dutch records are a *type* of the English term records. 'Loan terms' are terms from a source language that are adopted in a target language when they refer to concepts that are unfamiliar to the users of that language. This is particularly true of names (e.g. 'lei' will be adopted for the Romanian currency in the Hungarian version and 'forint' for the Hungarian currency in the Romanian version of the same thesaurus). Riesthuis concedes that such a strategy is less appropriate for abstract concepts and introduces 'inexact equivalence' to accommodate terms that in translation dictionaries are often cited as equivalent and are therefore used for indexing, but are actually not truly equivalent, e.g. 'Wissenschaft' (German) and 'Science' (English). 'Single-to-multiple' and 'non-equivalence' represent matches similar to (the converse of) 'many to one' and 'no match' respectively, as proposed by Iyer and Giguere [27].

3.1. Conceptual approaches: Renardus, SKOS Core, the semantic web

Koch et al. [3] developed a web-based service (Renardus) to facilitate searching and browsing across a variety of distributed European information services and subject gateways. DDC was used as a common switching terminology and browse structure. Koch et al. acknowledged the need to specify the degree of mapping equivalence and used the principle of set theory to create five separate mapping match types: fully equivalent, narrower, broader, major overlap and minor overlap, when compared with a DDC class. It is worth noting that the Renardus match types are less concerned with expressing the specific nature of matches (or otherwise) and instead seek to characterize relationships of a *conceptual* nature. For example, 'fully equivalent' denotes that there is good equivalence between the terminologies, irrespective of how that concept may be represented. Such a match type essentially subsumes those matches generally described as exactly or conceptually equivalent [8, 10]. It also subsumes those matches that might normally be differentiated on the grounds of grammatical or lexical variations (e.g. plural/singular, abbreviations/acronyms, etc.). Whilst the approach proposed by Koch et al. [3] jettisons the emphasis placed on terminological incongruities, it is consistent with traditional classification and indexing theory which attempts to reconcile concepts rather than the terms used to represent those concepts.

A similar approach has been adopted by semantic web approaches, such as the proposed W3C Simple Knowledge Organization System (SKOS) Core [29]. SKOS Core is based on the representation of concepts and is an application of the Resource Description Framework (RDF). It provides a model for expressing the structure and content of various KOS to enable easy machine processing. Miles and Brickley [30] have proposed the SKOS Core Mapping Vocabulary Specification (MVS) to support the mapping of concepts between different schemes using the SKOS Core framework. This emerged from similar work [31] undertaken by the SWAD-Europe project [32]. The properties proposed by SKOS are: exactMatch, broadMatch, narrowMatch, majorMatch and minorMatch. The SKOS Core MVS also supplements the match types with a series of classes (AND, OR, NOT) for combining or excluding concepts. For example, the class AND is used to denote the intersection of two or more concepts. The term 'health services administration' in terminology *A* may therefore map to 'health services' AND 'administration' in terminology *B*.

The definitions of the SKOS MVS match types are similar to those used by Koch et al. [3] and are based on the assumption that the number of resources assigned to a particular concept is known. For example, majorMatch is where a 'set of resources properly indexed against concept A shares more than 50% of its members with the set of resources properly indexed against concept B' [30]. It therefore remains unclear how appropriate the SKOS Core MVS currently is for terminology mapping services. Such match type definitions are conducive to static terminology mappings, but less suited to dynamic mappings (invoked via a terminology server) where little is known about the resources or the indexes held in the repositories with which a client will interact. Although some of the match types could theoretically be used, their application would probably be inconsistent with the conceptual underpinnings and assumptions inherent in the Specification (unless appropriate extensions or modifications are made). The SKOS Core MVS has yet to experience wide deployment or testing; however, Liang et al. [33] report difficulties while mapping from AGROVOC Thesaurus to the Chinese Agricultural Thesaurus. They cite ill-defined mapping properties and find the assumptions inherent in the Specification to limit particular applications. Liang et al. consequently propose some redefinitions.

3.2. Matches derived via co-occurrence mapping

OCLC have experimented significantly with co-occurrence mapping, e.g. [34, 35], involving statistical routines which extract 'loosely-mapped' terms from metadata records containing terms from more than one terminology [2]. For example, it is possible to implement a co-occurrence process using MARC21 Authority Format [36] metadata records that employ tag 082 (DDC number) and tags 600–651 (subject added entry) that use second indicator 0, denoting LCSH (Library of Congress Subject Headings) in order to derive a loose set of mappings between the two terminologies. Such techniques have been used to great effect by OCLC to provide popular LCSH with mapped DDC numbers for practitioners [37] and within the WebDewey service [38].

Vizine-Goetz et al. [7] recently conducted research to further develop such inter-vocabulary association techniques and mapped terms from the Educational Resources Information Center (ERIC)

Thesaurus to LCSH. Their methodology entailed encoding both terminologies according to the MARC21 Authority Format and implementing a series of algorithms to ascertain matches. To aid match analysis and to express inter-term relationships, Vizine-Goetz et al. categorized matches according to four separate match types, PT/PT, PT/NPT, NPT/NPT and NPT/PT, all of which signify exact matches between preferred and non preferred terms in source/target terminologies. While spacing, capitalization and punctuation are ignored during their matching process, Vizine-Goetz et al. acknowledge that they focus on exact matches and that various other potential matches (e.g. plural/singular, further specification, etc.) are not accommodated within their categorization. The match types cannot therefore be said to be exhaustive and are optimized for investigation of the relational vocabularies at hand (i.e. ERIC and LCSH). Their use as generic mapping types (i.e. applicable to all kinds of KOS) consequently remains unclear.

3.3. Chaplan match types

Arguably the most significant contribution to mapping match types has been proposed by Chaplan [8] whose investigation focussed on identifying the nature of term matches that could potentially be used to enhance the performance of switching systems. Chaplan's methodology entailed the intellectual mapping of terms from the Laborline Thesaurus to LCSH, resulting in the subsequent identification of 19 separate match types (Table 1). Chaplan concluded that the relationships between terminologies were 'vastly more complex than supposed' and stated that relatively straightforward matches (i.e. exact match, partial match, no match) were inadequate to accurately characterize the full range of relationship types evident between terms in different schemes. It is noteworthy that several of Chaplan's more complex, semantic matches confirm those identified during thesauri reconciliation experiments by Neville [9] (e.g. Chaplan: opposite or negative; Neville: antonymous terms), whilst others are of a morphological nature, e.g. singular/plural match. Chaplan notes that further research is required to ascertain whether these results are applicable across a variety of different terminologies. The work documented in this paper goes some way to testing the applicability of Chaplan's match types across a variety of KOS, details of which are provided in the methodology section.

4. Rationale and objectives

The work documented here attempts to examine terminology mapping match types in relation to a DDC-based terminology server. In particular, we assess the suitability and validity of Chaplan's 19 match types [8] as the basis of a *generic suite* of equivalence matches to be used by services employing terminology mapping. Match types will facilitate the expression of the nature of equivalence between terms from different schemes, thus improving the ability to search disparate collections employing different terminologies. Such an assessment requires consideration of whether mappings between disparate KOS terminologies can be adequately represented by Chaplan's set of match types or whether alternative and/or additional match types are required. This is particularly important since the majority of match type research focusses on mapping between similarly structured relational vocabularies.

An earlier instantiation of the DDC spine-based terminology server used match types based on the work of Chaplan. These match types were used primarily to aid users during the disambiguation process whereby they select their preferred term in context from the DDC hierarchies presented. The growth of semantic web applications [39] and the associated need to deconstruct and link lexically disparate search terms or phrases [40] suggests that the broad range of match types proposed by Chaplan may not always be required. However, the conceptual approach based on set theory (i.e. SKOS Core MVS), as noted in Section 3.1, can be limited for terminology services and for some of the terminological functions such services may wish to offer. We envisage instances within our framework where finer granularity may be required (e.g. during particular phases of user disambiguation). Similarly to Liang and Sini [41] we consider the conceptual approach to be somewhat abstract for the practical application of mappings in this context. Extensions to the SKOS Core MVS are outside the scope of

Table 1
Chaplan's terminology match types

Match type code	Definition	Chaplan's examples Laborline Thesaurus	Chaplan's examples LCSH
1	Exact match	Industrial relations	Industrial relations
2	Exact cross-reference match	Child labor	USE Children – employment
3	Exact match, but with intervening characters	Research management	Research – management
4	Plurals	Displaced worker	Displaced workers
5	Subordination, in the form of a species–genus relationship	Industrywide bargaining	Collective bargaining
6	Superordination, in the form of genus–species relationship	Motor vehicle industry	Automobile industry and trade
7	Part-of-speech difference	Employment interview	Employment interviewing
8	Word-order variation	Illegal alien	Aliens, illegal
9	Further specification	Absenteeism	Absenteeism (labor)
10	Spelling variation	No strike clause	No-strike clause
11	Suffix variation	Quality of working life	Quality of work life
12	Abbreviation or acronym	Alta.	Alberta
13	Subdivision (represents term that was used only as a subdivision in LCSH)	Measurement	Measurement
14	Concept match	Performance appraisal	Employees – Rating of
15	Homograph	Millinery [referring to hat industry]	Millinery [referring to costume hats]
16	Translation	Precedent	Stare decisis
17	Date or numerical variation	1935	Nineteen thirty-five
18	No match	Boulwarism	Deskilling
19	Opposite or negative	Desegregation	Segregation

this paper; we are interested in the extent to which Chaplan's match types could form the basis of a generic suite of match types to be used by terminology services. It is hypothesized that such a large number of match types – across a variety of terminologies and using the specified rules – is unnecessary and could easily be collapsed into a smaller number, possibly reflecting alternative approaches [3, 30]. It is also thought that the scope of some match types (specifically part-of-speech difference and suffix variation) is ill-defined, which may lead to misapplication.

5. Methodology

5.1. Selection of schemes

To test the validity of Chaplan's match types, four terminologies were selected for mapping to DDC [42]: LCSH [43], MeSH (Medical Subject Headings) [44], UNESCO Thesaurus [45] and AAT (Art and Architecture Thesaurus) [46]. The selection of these particular terminologies was purposive. Each of the selected terminologies experiences wide international use and two (MeSH and AAT) are discipline-specific, thus exemplifying significant subject detail and higher levels of granularity than the two general schemes (LCSH and UNESCO Thesaurus). A categorization of the terminologies used according to Zeng and Chan's [2] KOS typology is provided in Table 2. Relational vocabularies are those that emphasize the use of cross-references and associative relationships between

Table 2
Terminologies categorized according to KOS typology

Terminology	KOS type
AAT	Relational vocabulary (thesaurus)
DDC	Classification and categorization scheme (bibliographic classification scheme)
LCSH	Relational vocabulary (subject heading list)
MeSH	Relational vocabulary (thesaurus)
UNESCO Thesaurus	Relational vocabulary (thesaurus)

concepts and the terms used to represent those concepts (e.g. thesauri and subject headings). Classification schemes are those schemes that seek to order and group like concepts, thus establishing distinct subject groups (e.g. LIS classification schemes and taxonomies). The selection of schemes pertaining to each type of categorization enables the examination of different structures since, in reality, different collections and services use differently structured terminologies; a problem at the root of vocabulary interoperability. An example of term lists was not included in the investigation; however, our assumption is that term lists exemplify simpler structures than relational vocabularies. Any match types capable of accommodating the latter form of KOS should theoretically be more than capable of accommodating the former.

5.2. Selection of terms

Machine readable copies (in XML) of the terminologies were obtained and loaded into an appropriately structured database. A simple Java program was written to randomly select 50 terms from each terminology. The extracted terms were then mapped to DDC notation by both authors (A and B). To assist in the mapping process, terms from the selected terminologies and notation/captions from the terminological spine (DDC) were considered in context. That is, the nomenclature surrounding terms (in both extracted terms and DDC notation/captions), any broader and narrower relationships, related terms, and scope notes associated with terms, were all studied to ensure accuracy of mappings between terminologies. Notation of the nearest broader or narrower concept was considered if no suitable exact or concept match could be found in the target terminology [4]. WebDewey [38] was used to search and browse DDC schedules for appropriate mappings. These tasks were undertaken *independently* by each author in order to increase the validity of identified mappings, and results were recorded in an appropriately structured matrix (Table 3).

Further consistency was ensured by observing strict DDC application rules with respect to the ‘class here’ and ‘including’ notes, which were treated distinctly [47]. For example, DDC caption scope notes employing the use of ‘class here’ are considered to approximate the whole class and therefore are unlikely to receive separate numbers. When instructed to ‘class here’, a concept match was assumed. Similarly, ‘include’ notes were considered to constitute a narrower term match. As such, where ‘class here’ and ‘including’ notes were evident, between-term relationships were coded as match types 14 and 5 respectively. Authors A and B re-grouped following the mapping process to compare results. Contentious mappings were examined closely and resolved through a process of re-analysis of DDC schedules and any available instructions relating to the mapped schemes.

It should be noted that DDC notation is supplemented with captions used to inform practitioners during concept translation. Although these are often used for creating labelled hierarchical browsing structures in various systems, e.g. [48, 49], and are occasionally used by services for indexing, the primary indicator remains the *notation*. However, since the display of DDC notation in isolation is of limited value to users, our terminology server provides captions (alongside notation) to make mappings meaningful to users (as do other systems, see for example [49]) and to aid the usability of the alternative terminological functions offered. Consequently, mappings from satellite terminologies to the DDC spine take account of captions as a principal source of mapping information.

Table 3
Portion of example UNESCO to DDC mapping matrix

UNESCO term	DDC no.	DDC caption	Auxiliary notation used	Optional notes
Vocational schools	373.246	Secondary education > Secondary schools and programs of specific kinds, levels, curricula, focus > Academic, military, vocational schools > Vocational schools	N/A	
Fuel technology	662.6	Chemical engineering > Technology of explosives, fuels, related products > Fuels	N/A	
Aquaculture	639.8	Agriculture > Hunting, fishing, conservation, related technologies > Aquaculture	N/A	
Library technicians	023.3	Library & information sciences > Personnel management (Human resource management) > Technician positions	N/A	UNESCO term within DDC scope notes
Paramedical personnel	610.690233	Medicine and health > Organizations, management, professions > Medical personnel and relationships > Allied health personnel	Notation added from elsewhere in schedules	

5.3. *Categorization of mappings*

Authors A and B then categorized the mappings in accordance with Chaplan's 19 match types. These categorizations were undertaken independently and were encoded by adding 1–19 to an additional column of the matrix. The authors then reconvened to determine the level of agreement of codes assigned across all 200 mapped terms. Individual matrices were merged to ascertain where authors' match types agreed, or otherwise. Where concordance on match types did not occur, the relevant terminologies were revisited to clarify terms in context together with relevant nomenclature. Where necessary, additional research work was undertaken, including consulting reference works and domain-specific resources to elucidate term definitions and scope. In all instances, the authors were able to reach agreement on match codes assigned.

5.4. *Comment on linguistic analysis*

The area of linguistic analysis is extremely complex although for vocabulary mapping between disparate terminologies it is considered unnecessary and indeed unrealistic to attempt to take cognizance of all levels of linguistic analysis within an online search environment. The START Information Server [50], built at the MIT Artificial Intelligence Laboratory, concerned with sentence level processing, was created under the premise that morphological, syntactic and semantic information is necessary and sufficient for the system to understand natural language in the form of sentences. In the present context of a terminology server, users will typically search for a single term or phrase [51] and are likely to assess search results based on the incidence of these terms/phrases within the documents returned. It is unlikely that complete sentences would be entered as search queries. Users are likely to consider the form of words, their meaning and context when judging the relevance of results returned. As such, few levels of linguistic analysis are considered within the current study when establishing equivalence via terminology mapping, where single or compound terms from one scheme are being directly mapped to equivalent single or compound terms in another scheme. In this context phonetics and phonology appear peripheral since both concern the sound of language, rather than the meaning, context or sense of a term; elements typically at the centre of effective information retrieval. Although syntax is critical for effective computational linguistic analysis, primarily for sentence processing [52], the nature of user searching makes it less pertinent than other analytical levels. It follows that phonetics, phonology and syntax are largely overlooked in the present paper, with morphological, semantic and pragmatic differences between terms being considered of primary importance due to the online environment and the manual nature of the mappings implemented in this study.

This does not mean other elements of linguistic analysis are completely ignored however. Some aspects of analysis such as syntax and to some extent pragmatics will be facilitated by a process of disambiguation within a terminology server. The user will select the intended meaning of a search term or phrase, choosing from a range of contexts presented. For example, they will be able to specify whether a search for 'windows' refers to the Microsoft Operating System or the glass covered openings typically found within buildings and vehicles. It should be noted that for the purposes of automated mapping within the English language, however, far wider linguistic analysis and subsequent mapping algorithms would require to be undertaken and formulated.

5.5. Caveats

Chaplan's study assigned multiple match codes to mappings. This practice was not followed in the present study since the authors neither fully understood nor agreed with Chaplan's documented work. For example, the relationship between the terms 'watch making' and 'clock and watch making' could – according to Chaplan's definitions – be simultaneously considered as 2 (exact cross-reference match), 10 (spelling variation), 15 (homograph) and 18 (no match). In the authors' opinion a subordinate/superordinate relationship is also valid in such an example, but it is unclear why Chaplan has not encoded it accordingly. In addition, examples given to illustrate some match codes are ambiguous. For example, match code 8 (word order variation) is characterized as follows: A: Illegal alien; B: Aliens, illegal. Since this mapping also constitutes a singular/plural relationship it is not considered exclusive and is therefore a poor example with which to define word order variation. As a result of such uncertainty, and with a view to providing clarity to the user, the methodology asserted that only one match code could be assigned to any given mapping.

6. Findings

6.1. Match codes: level of agreement

The two sets of emergent data were combined to determine any areas of disagreement regarding the match codes assigned. The mean level of agreement between authors across all schemes was 164 (82%) with a standard deviation of 13.54. It was found (Table 4) that the level of agreement between authors was higher for discipline-specific schemes such as AAT and MeSH and somewhat lower for more generic schemes like LCSH and UNESCO. Taken together, the mean level of agreement for discipline-specific schemes was 93%, compared with 71% for the two general schemes investigated.

Authors A and B did not agree on the match type relationships between mapped terms on 36 of 200 occasions (18%) (Table 5). The highest proportion of disagreement was found between match codes assigned for LCSH and UNESCO, when compared to AAT and MeSH. The former two schemes elicited 80.56% of all disagreements between assigned match codes.

A total of 33 of the 36 (87.88%) between-author disagreements involved match code 14. Such disagreement constituted 91.67% of all disagreements across the 200 mappings implemented. That is, on 33 occasions one author categorized a mapping as a concept match while the other considered it to demonstrate an alternative type of equivalence. In 29 of the 33 cases, conflict arose between a concept match (14) and narrower (5) or broader (6) term matches. Disagreements involving match code 14 were the only type encountered when mapping AAT and MeSH to DDC.

Table 4
Level of agreement between match codes assigned by authors A and B

	AAT-DDC	LCSH-DDC	MeSH-DDC	UNESCO-DDC
Level of agreement	88%	74%	98%	68%

Table 5
Instances of conflict between match codes assigned by authors A and B

Match codes assigned by authors A/B or B/A	AAT-DDC	LCSH-DDC	MeSH-DDC	UNESCO-DDC	TOTAL
18/5	–	1	–	–	1
14/1	1	–	–	–	1
14/5	5	7	1	12	25
14/6	–	2	–	2	4
14/9	–	3	–	–	3
11/7	–	–	–	1	1
1/10	–	–	–	1	1
TOTAL	6	13	1	16	36

A further three distinct mismatches were evident from the LCSH and MeSH data (see Table 5). The first arose between code 18 (no match) and 5 (species-genus subordination). The second between code 11 (suffix variation) and 7 (part-of-speech difference), and the third between code 1 (exact match) and 10 (spelling variation). The latter was the only instance where authors A and B encoded a mapping differently and subsequently agreed on a third code when conflating their data. In all other cases, agreed match codes were consistent with at least one of the author's original categorizations.

6.2. Agreed match codes: frequencies

A frequency count of each of the agreed match codes was conducted (Table 6 and Figure 2). When mapping terms from AAT, LCSH, MeSH and UNESCO to DDC, match codes 1, 5 and 14 proved valid across all schemes. That is, terms from all four schemes elicited relationships categorized as exact match, narrower term and concept match when mapped to DDC (see Figure 2). The most commonly assigned match code was 5 (narrower term) constituting 113 (56.5%) of the 200 codes assigned. Beyond the 89% of mappings categorized as narrower, concept or exact matches, the remaining 11% were collectively indicative of match codes 3 (exact match with intervening characters), 4 (plural form), 6 (genus-species superordination (or broader)), 7 (part-of-speech difference), 9 (further specification) and 10 (spelling variation). Match code 6 (broader term) was only assigned on three occasions and was used to characterize the relationship between terms from the more general LCSH and UNESCO. Match code 6 was not applied when mapping from subject-specific schemes.

6.3. Ranking of agreed match codes, by scheme

Assigned match codes were ranked according to frequency for each of the four schemes involved (see Table 7). Match code 5 (species-genus subordination) ranked the most frequently assigned code across three of the four schemes used – AAT, LCSH and MeSH – and ranked second in the case of UNESCO, with only a single occurrence (or 2%) separating the two top ranked match types in this case. UNESCO, when mapped to DDC, elicited one more concept match than narrower term match, making match code 14 the most highly ranked for this scheme.

Table 6
Frequency count (and percentage) of assigned match codes, for individual schemes and totals

Match code	AAT		LCSH		MeSH		UNESCO		Total match types assigned across all terminologies	
number	#	%	#	%	#	%	#	%	#	%
1	4	8	6	12	5	10	11	22	26	13
2	–	–	–	–	–	–	–	–	0	0
3	–	–	–	–	1	2	1	2	2	1
4	–	–	1	2	–	–	–	–	1	0.5
5	39	78	22	44	38	76	14	28	113	56.5
6	–	–	1	2	–	–	2	4	3	1.5
7	–	–	–	–	–	–	1	2	1	0.5
8	–	–	–	–	–	–	–	–	0	0
9	–	-	8	16	1	2	5	10	14	7
10	–	–	–	–	–	–	1	2	1	0.5
11	–	–	–	–	–	–	–	–	0	0
12	–	–	–	–	–	–	–	–	0	0
13	–	–	–	–	–	–	–	–	0	0
14	7	14	12	24	5	10	15	30	39	19.5
15	–	–	–	–	–	–	–	–	0	0
16	–	–	–	–	–	–	–	–	0	0
17	–	–	–	–	–	–	–	–	0	0
18	–	–	–	–	–	–	–	–	0	0
19	–	–	–	–	–	–	–	–	0	0
Total Terms	50	100	50	100	50	100	50	100	200	100

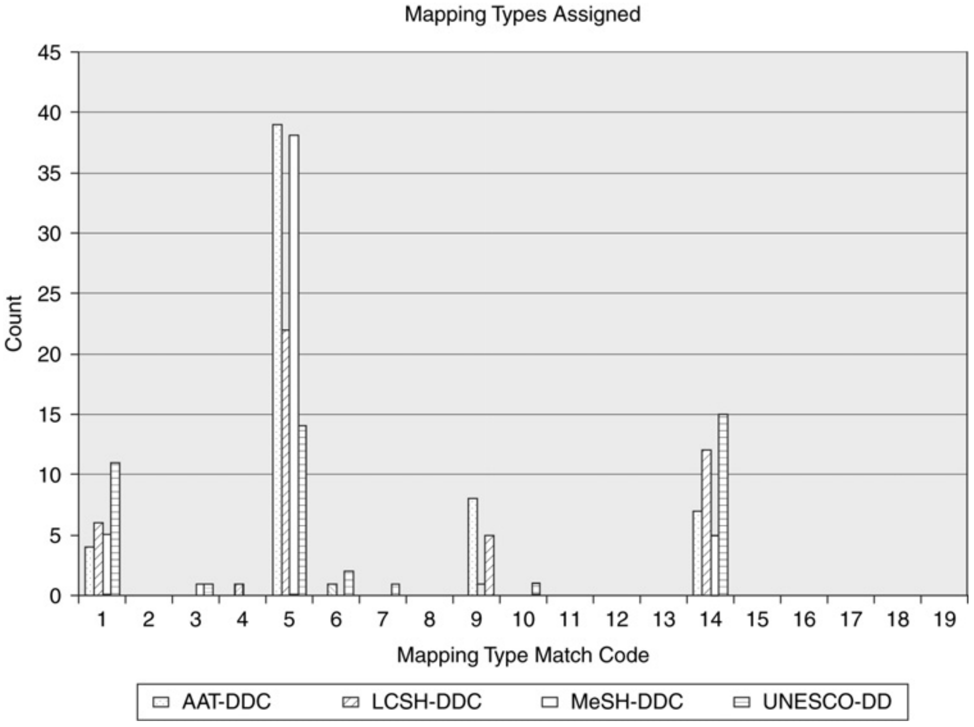


Fig. 2. Frequency count of assigned match codes, by scheme.

Table 7
Ranking of assigned match codes, by scheme

Match code number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
AAT	3	–	–	–	1	–	–	–	–	–	–	–	–	2	–	–	–	–	–
LCSH	4	–	–	5.5	1	5.5	–	–	3	–	–	–	–	2	–	–	–	–	–
MeSH	2.5	–	4.5	–	1	–	–	–	4.5	–	–	–	–	2.5	–	–	–	–	–
UNESCO	3	–	7	–	2	5	7	–	4	7	–	–	–	1	–	–	–	–	–
Mean Ranking (to 2 d.p.)	3.13	0	5.75	5.5	1.25	5.25	7	0	3.83	7	0	0	0	1.88	0	0	0	0	0

Table 8
Assigned match codes by ranking and scheme

Ranking	AAT	LCSH	MeSH	UNESCO
1	5	5	5	14
2	14	14	1/14	5
3	1	9	–	1
4	–	1	3/9	9
5	–	4/6	–	6
6	–	–	–	3
7	–	–	–	7
8	–	–	–	10

The second most frequently assigned match code was 14: concept match. This was the case for AAT and LCSH. It also ranked joint second for terms mapped from MeSH to DDC along with code 1: exact match.

Exact matches (code 1) were the third most frequently assigned mapping type, constituting 13% of match types across all schemes. Exact matches were the third most frequently encountered relationship between terms mapped from AAT and UNESCO. MeSH mappings elicited an equal number of exact matches and concept matches (5 occurrences or 10%). For LCSH, code 9 (further specification) ranked in third place.

Table 8 summarizes these findings by presenting assigned match codes by ranking, illustrating that fewer mapping types were required to characterize relationships between subject-specific schemes and DDC. A greater range of relationships was evident when considering equivalence relationships between terms in LCSH and UNESCO; that is to say, a more varied set of match codes was applied.

6.4. *Match types validated*

Tables 7 and 8 indicate that a total of nine Chaplan match types were deemed valid for the purpose of expressing equivalence relationships from terms in AAT, LCSH, MeSH and UNESCO, to DDC. Illustrative examples are provided in Table 9. Of the remaining 10 relationship types identified by Chaplan only two were assigned throughout the study, and were subsequently ruled out following the merging of authors' encoded match types. On one occasion author A categorized a match as 11 (suffix variation); and on a second occasion as 18 (no match). In the former case, the assignation of code 11 was replaced with 7 (part-of-speech difference). In the latter, 18 was replaced with 5 (narrower term).

Table 9
Examples of the nine match types verified

Match code number	Match type	Scheme term	DDC no.	DDC caption/ hierarchy
1	Exact match	AAT: Strasbourg The AAT scope note states: 'Refers to the style of faience produced at the Strasbourg pottery and porcelain factory in the 18th century. Widely imitated throughout Europe, the style features naturalistic floral decoration rendered in brightly colored enamel.'	738.309443954	DDC: Ceramic arts > Earthenware and stoneware > Historical and geographic treatment > Europe Western Europe > France and Monaco > Champagne-Ardenne, Ile-de-France, Lorraine, Alsace > Alsace > Bas-Rhin department > Strasbourg
3	Exact match with intervening characters	UNESCO: Viet Nam	959.7	Southeast Asia > Vietnam
4	Plural form	LCSH: Eye	573.88	Specific physiological systems in animals, regional histology and physiology in animals > Nervous and sensory systems > Eyes
5	Species–genus subordination	MeSH: Chromosomes, Human, Pair 5	572.87	Life sciences; biology > Internal biological processes and structures > General internal processes common to all organisms > Biochemistry > Specific biochemicals and biochemical genetics > Biochemical genetics > Chromosomes
6	Genus–species superordination	LCSH: Cultural industries	338.470705	Production > Secondary industries and services > Services and specific products > Documentary media, educational media, news media; journalism; publishing > Publishing
7	Part-of-speech difference	UNESCO: Heating	536	Physics > (Specific forms of energy) > Heat
9	Further specification	LCSH: Managed care plans (Medical care)	362.104258	Social welfare problems and services > Physical illness > Special topics of physical illness > Social aspects > Forms of assistance > Managed care plans

(Continued)

Table 9
(Continued)

Match code number	Match type	Scheme term	DDC no.	DDC caption/ hierarchy
10	Spelling variation	UNESCO: Educational programmes	344.0769	Labor, social service, education, cultural law > Education > Finance > Educational programs
14	Concept match	AAT: Scottish	306.09411	Culture and institutions > Historical, geographic, persons treatment > Treatment by specific continents, countries, localities; extraterrestrial worlds > Europe Western Europe > British Isles > Scotland

6.5. *Key findings: summary*

- Match codes were assigned more consistently for subject specific schemes than for more general schemes.
- 91.67% of between-author disagreements (as shown in Table 5) involved match code 14.
- A total of nine of Chaplan's original 19 match types were verified.
- Exact matches, concept matches and narrower term matches were the three most frequently assigned match codes, and were the only three to prove valid across all four schemes investigated. Between them, they accounted for 178 of 200 (89%) codes assigned.
- A narrower range of match codes was required to categorize relationship types when mapping terms from subject-specific schemes to DDC, compared to that of the general schemes LCSH and UNESCO.

7. Discussion

7.1. *Match codes: level of agreement*

It was noted in Section 6 that between-author variation arose in relation to particular assigned match codes; 91.67% of the said variation involved match code 14 (concept match), suggesting that the nature of this equivalence relationship is poorly defined, resulting in blurring of boundaries with other match codes. The fact that 80.56% of variations involving code 14 also involved codes 5/6 suggests that there is general confusion over what may constitute a narrower/broader term match and a concept match. It seems that concepts are often considered equivalent when one is actually a superset/subset of the other. The blurring of concept matches and narrower/broader term matches could result from an inability to distinguish sufficiently between an equivalent concept and super/ sub concepts (i.e. X is part of Y) or it may be a symptom of limited subject knowledge on the part of the authors. Besides disagreement involving code 14, a total of three additional disagreements were encountered.

Although evident on a single occasion only, the conflict between codes 7 and 11 suggests that such linguistic distinctions are not required for terminology mapping within this context. In this instance the UNESCO term 'Heating' was mapped to DDC 'Physics > (Specific forms of energy) > Heat' (536), in accordance with Chaplan's definitions. Author B considered this to be an example of 'part-of-speech difference' in line with Chaplan's example:

A: Employment interview

B: Employment interviewing

The authors recognize that, depending on the context, 'Heating' – 'Heat' might convey quite different concepts and therefore constitutes a weak example of the 'part-of-speech difference' match type proposed by Chaplan. However, it was the only example potentially capable of signifying this match type that emerged from the study. It is likely that with an increased set of mapped terms, a more sound example might have been uncovered. The need for a degree of morphological processing may be evident here, since the roots of the two words are identical although the meanings are quite different.

In contrast to author B's categorization as 'part-of-speech difference', author A categorized this mapping as a 'suffix variation', conforming to Chaplan's example:

A: Quality of working life

B: Quality of work life

On revisiting Chaplan's definitions there is no significant difference between the two examples quoted above, suggesting that they are sufficiently equivalent and could be merged. Indeed, Chaplan herself implied that these two measures of equivalence may not be sufficiently distinct from one another following a 50% overlap in terms categorized with both codes during her study.

Authors A and B assigned codes 18 and 5 respectively to the following example:

LCSH: Don Juan (Legendary character)

DDC: (808.80351) Rhetoric and collections of literary texts from more than two literatures > Collections of literary texts from more than two literatures > Arts and literature dealing with specific themes and subjects > Humanity > Specific persons

When revisited by the authors it became evident that a scope note under DDC Table 3-C [36] provides instruction to 'Include Don Juan'. It therefore follows that code 5 was agreed upon.

Where disagreement arose from one author assigning code 1 (exact match) and the other code 10 (spelling variation), the overall outcome was inconsistent with both authors. This was the only single occurrence of neither authors' codes being assigned following re-analysis of the terms in context. The authors agreed that example:

A: Viet Nam

B: (959.7) Southeast Asia > Vietnam

should be assigned match code 3 (exact match with intervening characters). Since each character is exactly matched it was agreed that capitalization did not constitute a spelling variation as such, but that the space in case A constituted – in a machine readable sense – an 'intervening character'. It could be argued that the above example also provides 'further specification'; however, the decision was taken early on that DDC hierarchies should be taken into consideration since contextual detail was required to ascertain, for example, whether a DDC caption was broader or narrower than its equivalent in an alternative scheme.

7.2. Agreed match codes: frequencies

It is likely that the reason match code 5 was the most frequently encountered relationship characterizing mappings from AAT, LCSH and MeSH, and the second most frequently occurring in the case of UNESCO, was due to the use of a universal classification scheme as the target terminology. DDC attempts to provide an epistemological interpretation of knowledge and the treatment of concepts is therefore often more broad, even when analytico-synthetic features are employed. The mapped schemes' terms tend to be more granular as indicated by the proportion of species–genus relationships. The infrequent assignation of reciprocal match code 6 (genus–species or broader term match) in our study appears to support this assertion. Code 5 was the second most frequently assigned in the case of UNESCO, although the first and second place rankings only differed by a single mapping. The inability of DDC to match the granularity of the mapped schemes is telling and suggests that in many cases the target will actually degrade the signal for the user [15, 18], as discussed in Section 2. Furthermore, the nature of DDC as a bibliographic classification scheme dictates that it is complex when compared to relational vocabularies or term lists, and is often not conducive to term-to-term mappings. Analytico-synthetic features have to be regularly employed to express particular concepts and therefore concepts do not exist in a formal sense. This renders the identification of direct equivalence problematic. It is therefore important to note that although the use of target schemes (e.g. DDC) often proves advantageous [3, 21, 22] and theoretically sound [19, 20, 53], their use may actually compromise retrieval performance for users.

Concept match (code 14) proved the second most frequently assigned match type in characterizing relationships between terms mapped from AAT and LCSH, and joint second (with exact match) for MeSH. For UNESCO, code 14 was the most frequently applied. This indicates that concept matches are evident across all schemes and are a necessary means of identifying like terms. This assertion holds when considering both general and discipline-specific schemes, indicating a good degree of conceptual equivalence across all schemes and accounting for 19.5% of total equivalence. Closer examination reveals a higher proportion of conceptual equivalence between universal schemes and DDC – i.e. LCSH (24%); UNESCO (30%) – than is evident in the case of more granular terminologies such as AAT (14%) and MeSH (10%).

Code 9 (further specification) ranked the third most frequently assigned match type when characterizing relationships between concepts from LCSH and DDC. We consider this to be a consequence of the structural nature of LCSH. For example, LCSH is a relational vocabulary (i.e. subject heading list) employing the use of subdivisions. Where the DDC hierarchy reads ‘Computer programming, programs, data > Programming > Programming languages’ (005.13) the equivalent LCSH heading would read ‘Programming languages (Electronic computers)’. While the DDC hierarchy provides contextual information clarifying that the programming languages being referred to directly relate to computers, the lack of sufficient hierarchical semantic structure in LCSH necessitates the use of qualifiers (i.e. ‘Electronic computers’), thus providing ‘further specification’.

Aside from the three most frequently assigned match types (narrower term, concept match and exact match), which the data highlight as characteristic of frequently occurring relationships between terms in disparate schemes, and with the exception of code 6 (broader) as a reciprocal entity of narrower, it appears that the remaining codes assigned (3, 4, 7, 9 and 10) (Table 6) can each be considered as a form of exact or concept match. This would suggest that the 9.5% of match types defined by codes 3 (exact match with intervening characters), 4 (plural form), 7 (part-of-speech difference), 9 (further specification) and 10 (spelling variation) could be combined and considered more generally as exact or concept matches. It is considered unlikely that the user of a terminology server would benefit from the knowledge that, for example, ‘absenteeism (labor)’ has further specification than ‘absenteeism’. Users simply want to know that the terms show some level of equivalence in respect of the concepts they represent [54] unless the further specification alters the context of the term.

7.3. Match types validated

Recall that the primary aim of this study was to test the hypothesis that such a large number of match types – across a variety of terminologies and using Chaplan’s rules of application – is unnecessary and could be collapsed into a smaller (perhaps more manageable) number. The present study validated the application of nine of Chaplan’s 19 match types as detailed in Table 10.

All four mapped schemes demonstrated incidences of exact matches, species–genus subordination and concept matches. This suggests that these three forms of equivalence should be retained in any future set of mapping types proposed. Exact match, narrower (and broader) terms and concept match all constitute benefits for the retrieval of information since they provide the user with further information on a subject area and potentially relevant terms with which to search. In addition, code 3 (exact

Table 10
Match types verified

Match Type Code	Mapping Type
1	Exact match
3	Exact match with intervening characters
4	Plural form
5	Species–genus subordination
6	Genus–species superordination
7	Part-of-speech difference
9	Further specification
10	Spelling variation
14	Concept match

match with intervening characters) was assigned to mapped terms originating from MeSH and UNESCO; code 4 (plural form) was applied to one term mapped from LCSH to DDC; code 6 (genus–species subordination) was verified by terms from LCSH and UNESCO; code 7 (part-of-speech difference) was assigned to a mapping from UNESCO; code 9 (further specification) applied to relationships between terms from LCSH, MeSH and UNESCO; code 10 (spelling variation) proved valid in characterizing the association between one UNESCO term and a DDC equivalent. A closer examination of these match types suggests that they are not sufficiently distinct to warrant their inclusion in a reduced set of mapping types, with the possible exception of code 6. Code 6 is likely to be more frequently assigned should a scheme with extremely broad subject groupings and a low level of specificity be mapped to DDC. In other words, code 6 could prove valid when a scheme contains top terms exemplifying a broader subject scope than those contained within the target terminology.

It is proposed that where frequency counts are low and/or scheme-specific they could be combined, thus reducing the overall range of mapping types required within a terminology service. Code 3 is essentially an exact match and it is proposed that such cases be characterized accordingly. The addition of e.g. a space, a hyphen or a colon does not sufficiently change the meaning of a term to warrant the need for an additional match type. It is proposed that codes 4, 7, 9 and 10 constitute a form of concept match and, as such, should be assigned code 14. In each of these cases mapped terms convey equivalent concepts.

7.4. Extraneous match types

Based on the results shown in Table 6, codes 2, 8, 11, 12, 13, 15, 16, 17, 18 and 19 appear redundant, given the current schemes and term sets extracted. This is not safe to assume at this stage, however, since the methodology led authors to actively seek exact matches, or as near to an exact match as possible. It is therefore probable that additional match codes are required depending on user circumstances. For example, it is possible that in a search for ‘employment’, the term ‘unemployment’ may be more useful than ‘work’. It follows that more of Chaplan’s match types than indicated by the data above may be relevant and that the set of match types presented in Section 6.4 should be supplemented accordingly.

Further research is required in this area to determine whether or not selected match types that appear extraneous in the present study – but were proven necessary within Chaplan’s study – may in fact prove valid to the user in specific scenarios. The current authors would argue that codes 2, 8, 11, 12, 13, 16 and 17 constitute forms of concept match. In the case of code 15 (homograph), it is questionable that any such form of relationship would be imposed between terms during intellectual mapping. Although homographs appear as exact matches on a presentational basis, their meanings are not equivalent. It follows that no level of exact or partial match is relevant and that such terms essentially constitute a ‘no match’ relationship. In the current study, conducted within the context of an M2M terminology server employing the use of a DDC spine, such terms would be presented to the user within their DDC hierarchies enabling their sense of meaning to become apparent. The process of user disambiguation should be sufficient to handle any potential confusion over the sense of homographs.

Chaplan did not find any incidence of match code 17 (date or numerical variation), making justification for the inclusion of this type of equivalence in her set of 19 unclear. This leaves match code 18 (no match), the relevance of which within an intellectual mapping scenario is questionable.

8. Conclusion and further work

The present study has confirmed that Chaplan's set of 19 match types is unnecessary for the purpose of characterizing equivalence relationships between terms in disparate schemes within the context of a DDC-spine based terminology server. The study examined the equivalence relationships necessary to map a subset of terms extracted from AAT, LCSH, MeSH and UNESCO to DDC. A total of nine of the 19 equivalence relationships were verified, with exact, concept and species–genus subordination proving the most frequently encountered types across all four schemes. This supports our stated hypothesis and provides us with a generic suite of match types.

Results of the present study indicate that general lessons for the field of information retrieval, such as improvements to relevance ranking algorithms, improved information for users relating to the relationship between their query and returned results and so on, may be learned. It would, however, in the authors' opinion be wrong to speculate on the significance of such indications within a real-world information retrieval context at this stage, in the absence of any concrete evidence for doing so. Further work will involve the implementation of our results (the reduced set); the effects on subsequent information retrieval via a terminology server will then be investigated.

It is considered likely that the nine match types verified from Chaplan's set could be further reduced, provided they are sufficiently well defined, to form a set closer to that proposed by the set theory-based SKOS Core MVS model. The present study indicates that the developing SKOS Core MVS is probably insufficient as it stands and requires modification, since currently only three match types (exactMatch, broaderMatch and minorMatch) appear applicable in the context of a distributed terminology server. This is consistent with the work of Liang et al. [33] and Liang and Sini [41] who found that the MVS required supplementing and redefining to express match types identified between AGROVOC and the Chinese Agricultural Thesaurus sufficiently. Nevertheless, the value of either approach for users engaging in a process of disambiguation remains unclear. We propose to conduct an appropriate user study to verify that a conceptual basis for match types is sufficient for the purposes of retrieval via a terminology server and that lexical differences do not compromise user success in this context.

The authors consider a principal focus of further work to be reconciling the differences between these disparate equivalence approaches. The set theory approach is considered advantageous because it provides a layer of abstraction, thus allowing equivalence types to be easily derived between extremely dissimilar KOS (e.g. between a classification and a thesaurus). Such approaches [3, 30] achieve this by maintaining abstraction, thus lowering equivalence specificity which, as we have seen, may be required for some applications. Conversely, approaches that provide specificity [8, 9, 27] can be difficult to apply across disparate KOS and are susceptible to misapplication, as demonstrated in this study. Within our model we see benefits to both approaches and instances where these approaches could be combined, depending on client service requirements. However, the principal focus should seek to reconcile their respective differences in order to identify a suite of equivalence matches that can balance the advantageous aspects of both approaches:

- To be sufficiently generic to enable deployment by communities of practice and in a wide variety of contexts or applications, including the semantic web.

- To be easily applied across a variety of KOS by practitioners or service administrators with only a working knowledge of terminological issues.
- To offer a degree of specificity, but without compromising easy application.
- To provide sufficiently robust equivalence definitions so as to minimize misapplication and to promote the consistent characterization of equivalence relationships.

A possible limitation of the present study has been noted in Section 5.5; however, areas where potential limitations might be identified relate to the choice of schemes and the design of the mapping analysis respectively. Firstly, it is feasible that the choice of schemes investigated together with the extraction of terms to be mapped (although random) may have affected the outcome. In order to eliminate any such bias, the study should be extended to look at a wider range of schemes and a greater selection of terms from each. Secondly, the establishment of mappings and the analysis of equivalence types were conducted by two individuals. This was found to be necessary since the nature of the analysis required a high degree of proficiency in terminology mapping and KOS generally, therefore prohibiting the use of multiple researchers. However, we consider the controls used to stem bias and increase validity with respect to deriving mappings (see independent mapping in Section 5.2) and assigning match codes (see independent match type categorization in Section 5.3) to ameliorate any significant bias. It would nevertheless be valuable to observe the effect a larger number of suitably qualified researchers would have on match code assignation. Such limitations will be considered in future work, as described above.

References

[1] L.M. Chan and M.L. Zeng, Ensuring interoperability among subject vocabularies and Knowledge Organization Schemes: a methodological analysis, *IFLA Journal* 28(5/6) (2002) 323–7.

[2] M.L. Zeng and L.M. Chan, Trends and issues in establishing interoperability among Knowledge Organization Systems, *Journal of the American Society for Information Sciences and Technology* 55(5) (2004) 377–95.

[3] T. Koch, H. Neuroth and M. Day, Renardus: Cross-browsing European subject gateway via a common classification system (DDC). In: I.C. McIlwaine (ed.), *Proceedings of the IFLA Satellite Meeting Held in Dublin, Ohio, 14–16 August 2001 and Sponsored by the IFLA Classification and Indexing Section, the IFLA Information Technology Section and OCLC* (K.G. Saur, München, 2003) 25–33.

[4] M. Doerr, Semantic problems of thesaurus mapping, *Journal of Digital Information* 1(8) (2001). Available at: http://jodi.tamu.edu/Articles/v01/i08/Doerr/ (accessed 10 July 2006).

[5] C. Binding and D. Tudhope, KOS at your Service: Programmatic Access to Knowledge Organization Systems, *Journal of Digital Information* 4(4) (2004). Available at: http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Binding/ (accessed 10 July 2006).

[6] J.-E. Mai, The future of general classification, *Cataloging and Classification Quarterly* 37(1/2) (2003) 3–12.

[7] D. Vizine-Goetz, C. Hickey, A. Houghton and R. Thompson, Vocabulary mapping for terminology services, *Journal of Digital Information* 4(4) (2004). Available at: http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/ (accessed 10 July 2006).

[8] M.A. Chaplan, Mapping Laborline Thesaurus terms to Library of Congress Subject Headings: implications for vocabulary switching, *Library Quarterly* 56(1) (1995) 39–61.

[9] H.H. Neville, Feasibility study of a scheme for reconciling thesauri covering a common subject, *Journal of Documentation* 26(4) (1970) 313–36.

[10] J.P. Silvester and P.H. Klingbiel, An operational system for subject switching between controlled vocabularies, *Information Processing and Management* 29(1) (1993) 47–59.

[11] E. Freyre and M. Naudi, MACS: subject access across languages and networks. In: I.C. McIlwaine (ed.), *Proceedings of the IFLA Satellite Meeting Held in Dublin, Ohio, 14–16 August 2001 and Sponsored by the IFLA Classification and Indexing Section, the IFLA Information Technology Section and OCLC* (K.G. Saur, München, 2003) 3–10.

[12] G. Clavel-Merrin, MACS (Multilingual access to subjects): a virtual authority file across languages, *Cataloging and Classification Quarterly* 39(1/2) (2004) 323–330.

[13] E. McCulloch, A. Shiri and D. Nicholson, Challenges and issues in terminology mapping: a digital library perspective, *The Electronic Library* 23(6) (2005) 671–7.

[14] V. Horsnell, *Intermediate Lexicon for Information Science: a Feasibility Study* (Polytechnic of North London, London, 1974).

[15] V. Horsnell, The Intermediate Lexicon: an aid to international co-operation, *Aslib Proceedings* 27(2) (1975) 57–66.

[16] R.T. Niehoff, Development of an integrated energy vocabulary and the possibilities for online subject switching, *Journal of the American Society for Information Science* 27(1) (1976) 3–17.

[17] P.S. Kuhr, Putting the world back together: mapping multiple vocabularies into a single thesaurus. In: I.C. McIlwaine (ed.), *Proceedings of the IFLA Satellite Meeting Held in Dublin, Ohio, 14–16 August 2001 and Sponsored by the IFLA Classification and Indexing Section, the IFLA Information Technology Section and OCLC* (K.G. Saur, München, 2003) 33–42.

[18] E.J. Coates, Switching languages for indexing, *Journal of Documentation* 26(2) (1970) 102–10.

[19] G.A. Lloyd, The Universal Decimal Classification as an international switching language. In: H. Wellisch and T.D. Wilson (eds), *Subject Retrieval in the Seventies: Proceedings of an International Symposium Held at the University of Maryland, Maryland, 14–15 May 1971* (Greenwood, Connecticut, 1972).

[20] E. Svenonius, Use of classification in online retrieval, *Library Resources and Technical Services* 27(1) (1983) 76–80.

[21] D. Nicholson, Subject-based interoperability: issues from the High Level Thesaurus (HILT) project. In: *Proceedings of the 68th IFLA General Council and Conference – Classification and Indexing, Glasgow, UK, August 18–24 2002* (IFLA, The Hague, 2002). Available at: www.ifla.org/IV/ifla68/papers/006–122e.pdf (accessed 10 July 2006).

[22] M. Balikova, Multilingual subject access to catalogues of National Libraries (MSAC): Czech Republic's collaborations with Slovakia, Slovenia, Croatia, Macedonia, Lithuania and Latvia. In: *Proceedings of the World Library and Information Congress: 71st IFLA General Conference and Council – Classification and Indexing with Cataloguing, Oslo, Norway, August 14–18 2005* (IFLA, The Hague, 2005). Available at: www.ifla.org/IV/ifla71/papers/044e-Balikova.pdf (accessed 10 July 2006).

[23] I. Dahlberg, Towards establishment of compatibility between indexing languages, *International Classification* 8(2) (1981) 86–91.

[24] D. Nicholson, A. Dawson and A. Shiri, HILT: a pilot terminology mapping service with a DDC spine, *Cataloging and Classification Quarterly* 42(3/4) (2006) 187–200.

[25] A. Shiri, D. Nicholson and E. McCulloch, User evaluation of a pilot terminologies server for a distributed multi-scheme environment, *Online Information Review* 28(4) (2004) 273–83.

[26] D. Nicholson, *HILT: High-Level Thesaurus Project M2M Feasibility Study: Final Report to JISC* (CDLR, Glasgow, 2005). Available at: http://hilt.cdlr.strath.ac.uk/hiltm2mfs/0HILTM2MFinalReportRepV3.1.pdf (accessed 7 February 2007).

[27] H. Iyer and M. Giguere, Towards designing an expert system to map mathematics classificatory structure, *Knowledge Organization* 22(3/4) (1995) 141–7.

[28] G.J.A. Riesthuis, Information languages and multilingual subject access. In: I.C. McIiwaine (ed.), *Proceedings of the IFLA Satellite Meeting Held in Dublin, Ohio, 14–16 August 2001 and Sponsored by the IFLA Classification and Indexing Section, the IFLA Information Technology Section and OCLC* (K.G. Saur, München, 2003) 11–17.

[29] A. Miles and D. Brickley (eds), *SKOS Core Guide: W3C Working Draft 2 November, World Wide Web Consortium (W3C)* (2005). Available at: www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/ (accessed 10 July 2006).

[30] A. Miles and D. Brickley (eds), *SKOS Mapping Vocabulary Specification, World Wide Web Consortium (W3C)* (2004). Available at: www.w3.org/2004/02/skos/mapping/spec/ (accessed 10 July 2006).

[31] SWAD-Europe, *Inter-Thesaurus Mapping: a Guide to the SKOS-Mapping RDF Schema for Inter-Thesaurus Mapping, World Wide Web Consortium (W3C)* (2003). Available at: www.w3.org/2001/sw/Europe/reports/thes/8.4/ (accessed 10 July 2006).

[32] *SWAD-Europe*. Available at: www.w3.org/2001/sw/Europe/ (accessed 10 July 2006).

[33] A. Liang, M. Sini, C. Chun, S.J. Li, W.L. Lu, C.P. He and J. Keizer, The mapping schema from Chinese Agricultural Thesaurus to AGROVOC, *6th Agricultural Ontology Service (AOS) Workshop on Ontologies: the more practical issues and experiences, July 25–28, Vila Real, Portugal, 2005* (Food and Agriculture Organization, Rome, 2005). Available at: ftp://ftp.fao.org/docrep/fao/008/af241e/af241e00.pdf (accessed 10 July).

[34] L.M. Chan and D. Vizine-Goetz, Feasibility of a computer-generated subject validation file based on frequency of occurrence of assigned LC Subject Headings. Phase II, nature and patterns of invalid headings, *Annual Review of OCLC Research* 1995 (1996). Available at: http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?objid=0000003370 (accessed 10 July 2006).

[35] D. Vizine-Goetz, Popular LCSH with Dewey numbers, *OCLC Newsletter* (233) (1998). Available at: http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?objid=0000003449 (accessed 10 July 2006).

[36] Library of Congress, *MARC21 Concise Format for Authority Data* (2005). Available at: www.loc.gov/marc/authority/ (accessed 10 July 2006).

[37] J.S. Mitchell (ed.), *People, Places and Things: a List of Popular Library of Congress Subject Headings with Dewey Numbers* (Forest Press, Ohio, 2001).

[38] OCLC, *WebDewey* (2006). Available at: www.oclc.org/dewey/versions/webdewey/ (accessed 10 July 2006).

[39] M. Lytras, M.-A. Sicilia, J. Davies and V. Kashyap, Digital libraries in the knowledge era: knowledge management and Semantic Web technologies, *Library Management* 26(4/5) (2005) 170–75.

[40] L. Cantara, Encoding controlled vocabularies for the Semantic Web using SKOS Core, *OCLC Systems andServices* 22(2) (2006) 111–14.

[41] A.C. Liang and M. Sini, Mapping AGROVOC and the Chinese Agricultural Thesaurus: definitions, tools, procedures, *New Review of Hypermedia and Multimedia* 12(1) (2006) 51–62.

[42] OCLC, *Dewey Services: Dewey Decimal Classification* (2006). Available at: www.oclc.org/dewey/ (accessed 10 July 2006).

[43] Library of Congress, *Library of Congress Authorities* (2006). Available at: http://authorities.loc.gov/ (accessed 10 July 2006).

[44] United States National Library of Health, *Medical Subject Headings* (2006). Available at: www.nlm.nih.gov/mesh/ (accessed 10 July 2006).

[45] UNESCO and the University of London Computing Centre, *UNESCO Thesaurus* (2002). Available at: www2.ulcc.ac.uk/unesco/ (accessed 10 July 2006).

[46] J. Paul Getty Trust, *Art and Architecture Thesaurus Online* (2000). Available at: www.getty.edu/research/conducting_research/vocabularies/aat/ (accessed 10 July 2006).

[47] OCLC, *Introduction to Dewey Decimal Classification* (OCLC, Ohio, 2006). Available at: www.oclc.org/dewey/versions/ddc22print/intro.pdf (accessed 10 July 2006).

[48] *BUBL Information Service* (2006). Available at: http://bubl.ac.uk/ (accessed 7 February 2007).

[49] Koninklijke Bibliotheek, *Renardus* (2002). Available at: www.renardus.org/ (accessed 10 July 2006).

[50] B. Katz, *From Sentence Processing to Information Access on the World Wide Web* (Massachusetts Institute of Technology, Cambridge, 1999). Available at: http://people.csail.mit.edu/boris/webaccess/ (accessed 7 February 2007).

[51] D. Nicholson, A. Shiri and E. McCulloch, *HILT: High-Level Thesaurus Project Phase II: Final Report to JISC* (Centre for Digital Library Research, Glasgow, 2003). Available at: http://hilt.cdlr.strath.ac.uk/hilt2web/finalreport.htm (accessed 7 February 2007).

[52] M.W. Eysenck and M.T. Keane, *Cognitive Psychology* (3rd edition) (Psychology Press, East Sussex, 1995).

[53] S.Q. Liu, Decomposing DDC synthesized numbers. In: *Proceedings of the 62nd IFLA General Conference, August 25–31, Beijing, 1996* (IFLA, The Hague, 1996). Available at: www.ifla.org/IV/ifla62/62-sonl.htm (accessed 10 July 2006).

[54] R. Krovetz and W.B. Croft, Lexical ambiguity and information retrieval, *ACM Transactions on Information Systems* 10(2) (1992) 115–41.

Principles in Patterns (PiP): Evaluation

WP7:37 Evaluation of systems pilot

Phase 2: User acceptance testing of Course and Class Approval Online Pilot (C-CAP)

February 2012

University of Strathclyde

Contents

Figures

Tables

1. Introduction

The PiP Evaluation Plan [1] documents four distinct evaluative strands, the first of which entails an evaluation of the PiP system pilot (WP7:37 – Systems & tool evaluation) (Figure 1). Phase 1 of this evaluative strand focused on the heuristic evaluation of the PiP *Course and Class Approval Online Pilot* system (*C-CAP*) and was completed in December 2011. A report documenting the principal findings is available from the PiP project website [2]. Phase 2 is the final phase of the system and tool evaluation (WP7:37) and forms the basis of this report.

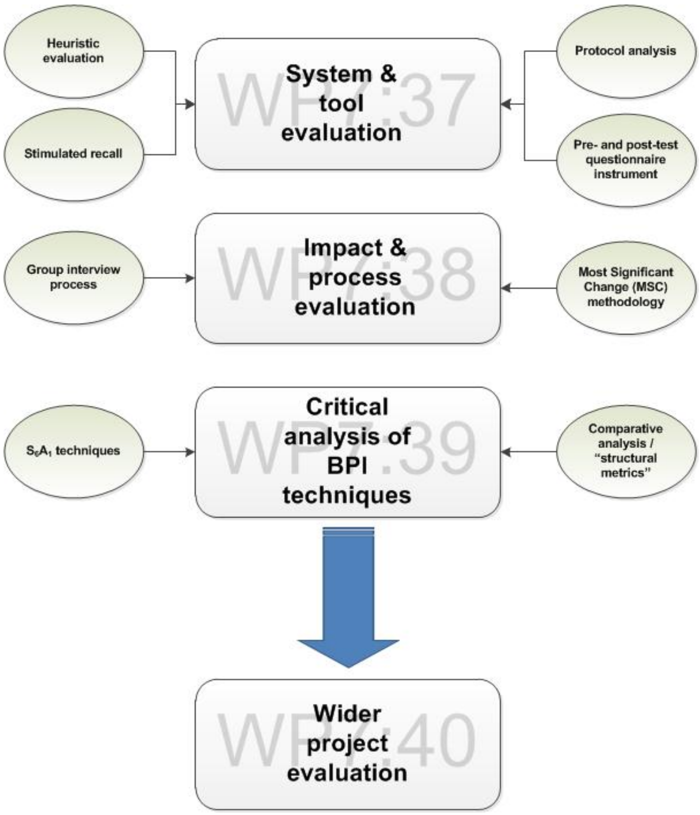

Heuristic evaluation
Stimulated recall
WP7:37
System & tool evaluation
Protocol analysis
Pre- and post-test questionnaire instrument
Group interview process
WP7:38
Impact & process evaluation
Most Significant Change (MSC) methodology
S_eA_1 techniques
WP7:39
Critical analysis of BPI techniques
Comparative analysis / "structural metrics"
WP7:40
Wider project evaluation

Figure 1: Overview of evaluative strands and evaluative sub-phases of PiP.

Smith and Brown [3] and Lai [4] discuss the importance of technology facilitated approaches to design and approval for the purposes of improving pedagogy and, in Lai's case, in increasing the portability and sharing of curricula within specific educational contexts. With the exception of PiP [5] and T-SPARC [6] - both funded under the JISC Institutional Approaches to Curriculum Design Programme [7] - very little is available in the literature to influence the development and evaluation of technology supported approaches to curriculum design and approval. Smith and Brown [3] and Lai [4] merely discuss the theoretical opportunities of technology supported curriculum design. PiP therefore represents a unique testbed with little academic research upon which to guide the evaluative approach adopted for such a project.

Phase 2 of the evaluation is broadly concerned with "user acceptance testing". This entails exploring the extent to which C-CAP functionality meets users' expectations within specific curriculum design tasks, as well as eliciting data on C-CAP's overall usability and its ability to support academics in improving the quality of curricula. The general evaluative approach adopted therefore employs a combination of standard Human-Computer Interaction (HCI) approaches and specially designed data collection instruments, including protocol analysis, stimulated recall and pre- and post-test questionnaire instruments. This brief report summarises the methodology deployed, presents the results of the evaluation and discusses their implications for the further development of C-CAP. It is anticipated that some solutions will be implemented within the lifetime of the project. This is consistent with the incremental systems design methodology that PiP has adopted. However, it should be recognised that the implementation of some solutions may not be feasible, either because there are insufficient project resources to implement them or because they lie outside the project scope.

2. Methodology

2.1 Aims

The PiP Evaluation Plan details the wider objectives of the project evaluation [1]. The aim of this phase of the evaluation was to expose C-CAP to facets of HCI testing in order to validate aspects of phase 1 and evaluate C-CAP within in a real user context, including C-CAP's ability to support academic participants in the design of curricula. The following broad research questions influenced the evaluative design:

1. The extent to which C-CAP functionality meets users' expectations within specific curriculum design tasks
2. Assessing the performance of C-CAP in supporting the participants in curriculum design task and approval process and its potential for improving pedagogy
3. Eliciting data on current approval process and how C-CAP could contribute to improvements in the process (i.e. its fitness for purpose).
4. Measuring the overall usability of C-CAP (e.g. interface design and functionality instinctive, navigable, etc.) and capture data on users' preferred system design/features

Details of the study participants are provided in section 2.3 and an overview of the procedure adopted in section 2.4.

Phase 1 of the evaluation formed an important basis for preparing the C-CAP system for phase 2. The following section (2.2) summarises the role of the heuristic evaluation in preparing for the user acceptance testing.

2.2 Phase 1: C-CAP interface improvements for optimising data collection

The use of heuristic evaluation in phase 1 was an integral part of ensuring C-CAP demonstrated a high degree of heuristic compliance prior to commencing phase 2. Heuristic compliance was considered imperative for two related reasons: minimising users' extraneous cognitive load during user acceptance testing, and; optimising user acceptance testing data.

"Intrinsic cognitive load" pertains to the inherent difficulty of a task while "extraneous cognitive load" relates to the task presentation, which is normally controlled by the task designer [8]. If the intrinsic cognitive load of a task is high, and extraneous cognitive load is also high, then problem solving or task completion may fail to occur. Adjusting the presentation of the task to lower extraneous cognitive load can facilitate task completion or problem solving if such adjustments mean that the resulting total cognitive load falls within the mental resources of the user [9]. A prominent theme in recent HCI research therefore pertains to how best to minimise the extraneous cognitive load users often experience as a result of interface or system design. Poor system usability and design has been shown to increase users' disorientation and cognitive load during system use [10–12]. As extraneous cognitive load increases so the cognitive resources available to the user to complete their primary task (e.g. locating information, interacting with a system to complete a work task, booking flights, etc.) decreases.

Systems that expose users to high levels of extraneous cognitive load as a result of poor system design and usability have been shown to erode human cognitive processing. This generally manifests itself in a measurable decline in task performance, inefficiency in task completion, increased error rates and user frustration [11–15]. In some user task settings a decline in higher-level metacognitive skills can also be observed [12]. Any system engaging users in high levels of intrinsic cognitive load (i.e. a system engaging academics in curriculum design) must therefore strive to minimise extraneous cognitive load if the system is support them in task completion. Given the frequent complexities and intellectual demands associated with the curriculum design process [16], any system has to ensure a high level of usability if it is to truly support and inspire academics in the curriculum design process. Failure to address the threat of extraneous cognitive load in this instance could potentially have resulted in poor task performance.

The above noted threat of extraneous cognitive load also has implications for the quality of data gathered during user acceptance testing. A system demonstrating high levels of extraneous cognitive load generally fails to engage the user with the primary task sufficiently [12]. The consequences for typical HCI testing is that user participants are therefore more likely to comment on trivial or superficial interface issues, or system errors that could easily be debugged prior to user exposure, rather than deeper system issues, or aspects of how the system supports them in the primary task (which, in this context, would be the curriculum design and approval process). A valid data collection environment is consequently not achieved and data can become skewed towards superficial system problems which are often not indicative of a system's wider raison d'être.

Phase 1 (heuristic evaluation) was therefore used to optimise C-CAP and ergo the data collection environment, thus minimising the potential for extraneous cognitive load during user acceptance testing. Phase 1 detected 27 heuristic violations in the C-CAP system [2]. Of these violations, 67% (n = 18) were classified at a mean severity rating of ≤ 2.67, and of these 11% (n = 3) were classified at severity rating 1 (Cosmetic problem only). Only 33% (n = 9) were classified at a mean severity rating ≥ 3. Over 93% of all detected heuristic violations were resolved prior to commencing user acceptance testing, leading to numerous system and interface improvements. Unresolved violations were attributable to factors outside the control of the PiP team, e.g. University process issues or the limitations of InfoPath. Appendix E provides indicative screen dumps of the C-CAP system as deployed for this phase of the evaluation.

2.3 Participants

The evaluation participants were drawn from the academic departments of the University of Strathclyde. Early outreach and stakeholder activity meant that many participants were already familiar with PiP and its work; however, participants for this evaluative phase were recruited via faculty list emails (circulated on behalf of the evaluator by faculty managers) and an all-staff announcement via the Weekly Digest[†] [*]. To be eligible participating academics were required to have experience of the curriculum design and approval process and to have been involved in the creation of new classes and/or courses in within last 2 years. In reality, almost all participants had been involved in either class or course design within the past 6 months. It was originally the intention of phase 2 to include faculty managers in the user acceptance testing; but since faculty managers only become involved with C-CAP to administer the approval process *after* curricula have been designed their involvement would amount to using a single interface screen. Faculty manager involvement was therefore considered unproductive at this stage and was deferred until WP7:38 when faculty piloting is scheduled to take place.

Ten academic participants agreed to participate in the study. Table 1 sets out participants' faculty, departmental and discipline affiliations. Despite the small sample numbers, the group originated from a broad range of academic backgrounds, including physics, economics, mathematics and statistics

[†] http://www.strath.ac.uk/weeklydigest/

[*] Phase 2 of the evaluation plan was required to be considered by the University Ethics Committee (UEC). The UEC mandated adjustments to the methodology to further protect the anonymity of academic participants. This included no direct recruitment of participants.

and biomedical sciences. Unfortunately no Humanities & Social Sciences (HaSS) faculty were recruited[‡].

Table 1: Faculty and departmental affiliations of study participants.

Participant No.	Faculty	Department / subject
1	Strathclyde Business School	Management Science
2	Faculty of Science	Department of Physics
3	Strathclyde Business School	Economics
4	Faculty of Science	Strathclyde Institute of Pharmacy and Biomedical Sciences
5	Strathclyde Business School	Management Science
6	Faculty of Engineering	Department of Mechanical and Aerospace Engineering
7	Faculty of Science	Department of Computer and Information Sciences
8	Strathclyde Business School	Economics
9	Strathclyde Business School	Management Science
10	Faculty of Science	Department of Mathematics and Statistics

2.4 Procedure

The user acceptance testing sessions were designed to include four distinct sections: Pre-session questionnaire instrument, protocol analysis, stimulated recall, and a post-session questionnaire. Each session was circa 60 mins in duration, including ethical conditions (e.g. signing of consent form, explanation of research scope, etc). Data collection was conducted throughout January 2012 in a controlled IT lab setting.

The following sections detail the methods used and describes the overall procedure.

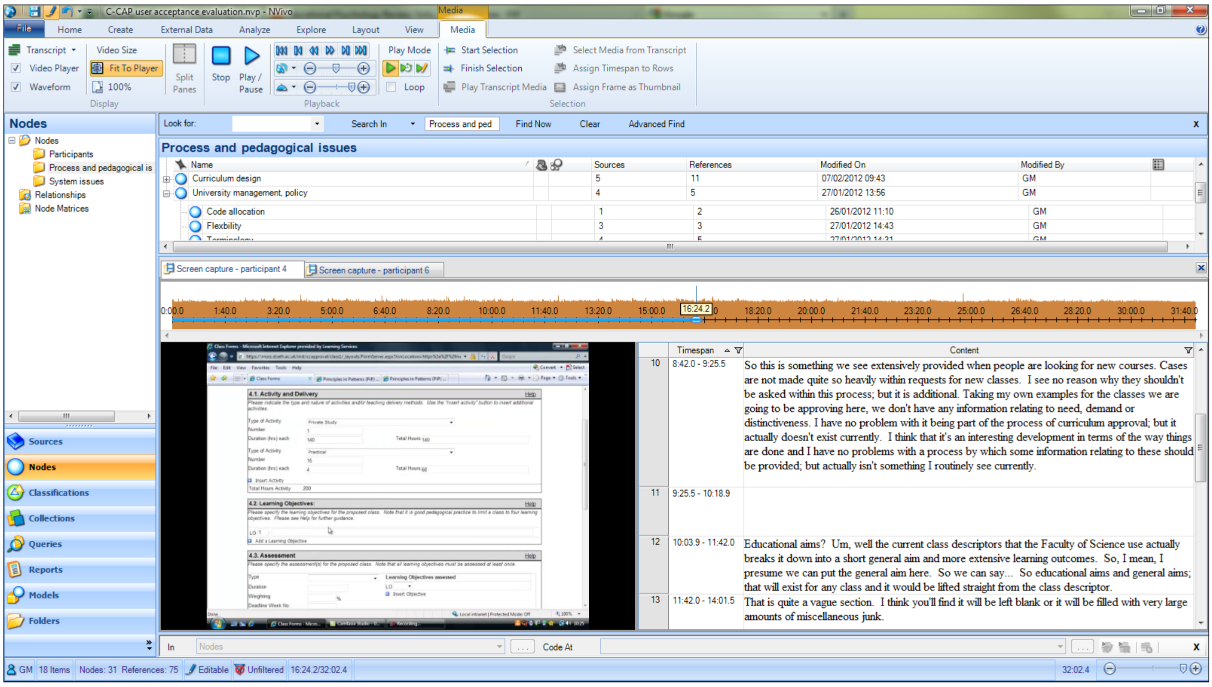
Protocol analysis

Protocol analysis (also known informally as the "think aloud protocol") is a frequently deployed user testing methodology for software, interfaces, systems, etc. in which participants are asked to complete a series of tasks with the test/pilot system while simultaneously verbalising their thoughts. Verbalisations (or protocols) are sound recorded and transcribed for analysis. Additional data may also be gathered (e.g. screen captures, evaluator logs, etc). The methodology is considered to have a high level of face validity as the data captured tends to focus on the *actual use* of a system rather than on user judgements concerning its *perceived* usability or efficacy. Protocol analyses are based on direct participant observation and attempt to model users' real world interaction with a system. As such, evaluators gain an insight into users' cognitive processes as the methodology tends to expose a wide variety of user problems, assumptions or misconceptions, many of which would otherwise go undetected. Protocol analysis was originally formalised by Ericsson and Simon [17] and later van Someren et al. [18] and has since become a widely used technique in user testing studies in a wide variety of system contexts [19–27].

To best model a genuine curriculum design process and test the C-CAP system in supporting curriculum design and approval, participants were asked to bring a recently drafted curriculum design form with them to the session. Participants were then instructed to replicate their form using the C-CAP system while thinking aloud, recognising that the form structure in C-CAP was different and often more detailed than existing curriculum design forms. For example, C-CAP offers a more structured approach by using efficiency tools [28] to accelerate form completion (e.g. drop down lists, auto-calculation of teaching hours / assessment weightings, etc.) and imposes some basic principles of curriculum design theory (e.g. adherence to constructive alignment [29], greater consideration of learning activities, etc.). Participants were briefed on the process of thinking aloud, which was in line with established protocol analysis procedures [18], [24]. Screen capture software was used to record both participants' C-CAP interface interaction (visual data) and to sound record their "think aloud"

[‡] Two HaSS participants were originally recruited but for external reasons were unable to participate.

protocols (audio data). Screen capture and associated audio data from the protocol analysis were uploaded into QSR NVivo 9 for content analysis, coding and further analysis (Figure 2). Data analysis was conducted according Holsti's [17] methodologies for content analysis and van Someren et al.'s techniques for category creation [18]. NVivo 9 was also used for audio transcription.



Process and pedagogical issues

Name	Sources	References	Modified On	Modified By
Curriculum design	5	11	07/02/2012 09:43	GM
University management, policy	4	5	27/01/2012 13:56	GM
Code allocation	1	2	26/01/2012 11:10	GM
Flexbility	3	3	27/01/2012 14:43	GM

	Timespan	Content
10	8:42.0 - 9:25.5	So this is something we see extensively provided when people are looking for new courses. Cases are not made quite so heavily within requests for new classes. I see no reason why they shouldn't be asked within this process; but it is additional. Taking my own examples for the classes we are going to be approving here, we don't have any information relating to need, demand or distinctiveness. I have no problem with it being part of the process of curriculum approval; but it actually doesn't exist currently. I think that it's an interesting development in terms of the way things are done and I have no problems with a process by which some information relating to these should be provided; but actually isn't something I routinely see currently.
11	9:25.5 - 10:18.9	
12	10:03.9 - 11:42.0	Educational aims? Um, well the current class descriptors that the Faculty of Science use actually breaks it down into a short general aim and more extensive learning outcomes. So, I mean, I presume we can put the general aim here. So we can say... So educational aims and general aims; that will exist for any class and it would be lifted straight from the class descriptor.
13	11:42.0 - 14:01.5	That is quite a vague section. I think you'll find it will be left blank or it will be filled with very large amounts of miscellaneous junk.

Figure 2: Screen capture data and transcribed audio as prepared for analysis in NVivo 9.

Throughout the protocol analysis session evaluator logs were used to record "significant events" that occurred during participants' interaction with the C-CAP system. "Significant events" can be defined as those moments where C-CAP was especially difficult for the participant to use or where C-CAP did not function as they expected (e.g. navigation was not located where the participant anticipated, C-CAP experienced a system error, participant experienced difficulty using the drop down menus for aligning assessment with learning objectives, etc.). The logs were created and maintained in MS Excel and included a time stamp and a brief description of the significant event (see example log in Appendix D). The overall purpose of the log was to record any events which might otherwise go unnoticed through the protocol analysis or to mark significant events worthy of further exploration via stimulated recall.

Stimulated recall

The stimulated recall technique (or "retrospective think aloud") is similar to protocol analysis but differs in that data are not collected until after the participant has completed their primary task [20], [24]. Often researchers use one or the other, normally owing to cost considerations; but research studies report on the benefits of both in identifying different HCI issues [28]. In stimulated recall a recorded screen capture of the participant's system interactions is played back to the participant who is then asked to articulate their cognitive processes and actions at specific points of the recording. Stimulated recall is generally considered favourable because although the participant is asked to verbalise after they have completed the task, they are often able to provide more detailed verbalisations owing to reduced cognitive load.

Stimulated recall was used immediately after participants had completed their "think aloud" curriculum design task using C-CAP (i.e. after the protocol analysis). A common drawback of protocol analysis is that some verbalisations can be inadequate. This is often the case when the user is engaged in cognitively onerous tasks, e.g. when the user is asked to verbalise while using a complex system interface [24]. Since participants in the user acceptance testing were engaging in the fictional but nonetheless cognitively onerous process of curriculum design with C-CAP, it was important that a brief stimulated recall phase be included in the testing session. Participants were only asked to engage in stimulated recall if significant events were logged during the "think aloud" curriculum design task. Stimulated recall would therefore focus the nature of those significant events and seek to tease out participants' thinking at the relevant stage of the screen capture video.

Stimulated recall was conducted immediately after the collection of protocol analysis data in order to review participants' system behaviour, thus teasing out potentially important data which may have been missed during protocol analysis. A total of six participants provided stimulated recall data. Stimulated recall data were sound recorded and uploaded to NVivo 9 for transcription and analysis alongside protocol analysis data.

Pre- and post-session questionnaire instruments

A pre-session questionnaire was administered prior to the commencement of the protocol analysis session in order to collect basic demographic information* and capture participants' IT efficacy. IT efficacy was measured using an adapted version of Murphy et al.'s [30] original Computer Self-Efficacy (CSE) scale, modified by Torkzadeh et al [31]. The instrument was also designed to elicit from participants their opinions and perceptions of the current curriculum approval process and its current issues.

The post-session questionnaire was administered after the completion of stimulated recall (if applicable). The post-session instrument was designed to capture data on users' success with the system and gather definitive data on the aspects of the system that participants perceived most favourably and those they did not. This was based on a customised version of the standard System Usability Scale (SUS) post-test instrument, first proposed by Brooke [32] and subsequently developed, deployed and validated by other usability researchers (e.g.[33], [34], [35], [36]). Brooke's instrument comprises a 10 item questionnaire using 5 point Likert scale response options. The post-session questionnaire also sought to capture perceptions of how C-CAP supported them in the curriculum design process and its potential for improving approval processes at the University of Strathclyde.

Both questionnaire instruments were administered using Bristol Online Surveys (BOS), an online survey tool [37]. Data from BOS was exported to a .csv file for analysis in MS Excel and in SPSS. The post-session instrument was also imported to NVivo 9 for analysis of open-ended question responses (i.e. Q.3).

Screen dumps of the questionnaire instruments as displayed in BOS are available in Appendices F and G.

Procedure summary

To summarise, the following data collection methods were used in the following order:

1. Pre-session questionnaire
2. Protocol analysis using C-CAP ("think aloud" curriculum design task)
3. Stimulated recall (based on recording playback of "think aloud" curriculum design task using C-CAP).
4. Post-session questionnaire

* The demographic information requirements of the questionnaire instruments were reduced in line with UEC requirements.

2.5 Methodological restrictions and limitations

The methodological approach adopted for this phase of the evaluation was subject to a variety of restrictions which, in turn, constitute limitations to the present design. This phase of the evaluation was ideally suited to a repeated measure approach in which participants would be exposed to alternative versions of C-CAP, thus permitting statistical inferences to be made between treatments. Unfortunately the timetable for the PiP project precluded the use of an additional development phase between treatments. It is also worth noting that the participant recruitment restrictions would have rendered such an approach untenable even if the timetable for evaluation was favourable. The current approach is therefore a compromise, with a suite of data collection techniques administered instead in order to gather rich data about participant interactions with C-CAP.

An additional limitation relates to the artificial nature of the curriculum design task that participants were asked to engage in during the testing session. To best model a genuine curriculum design process and the extent to which the C-CAP system can support academics in curriculum design and approval, participants were asked to replicate an existing curriculum design form within C-CAP. The new form structure and the peculiarities of C-CAP meant that this task was more than simply cutting and pasting, or re-typing from a hard copy. However, this nevertheless represents a compromise on requiring participants to draft curricula from scratch, which was deemed unfeasible as it would require excessively long protocols and would not necessarily capture the genuine drafting process, which is often incremental and protracted. It is anticipated that the piloting of C-CAP within faculties as part of the next evaluative strand (WP7:38 - Impact & process evaluation – see Figure 1) will better expose C-CAP to the verities of curriculum design and approval. Rich qualitative data is expected to be gathered for this strand, via group interviews and Most Significant Change (MSC) stories [1].

3. Results and discussion[*]

3.1 Questionnaire instrument data

Owing to the detail of the qualitative data gathered during the user acceptance testing it is necessary to first summarise the findings from both the pre- and post-session questionnaire instruments.

Pre-session questionnaire data

Table 2: Computer Self Efficacy (CSE) results.

Computer Self Efficacy (CSE) scale - statements†	**Participant results**		
	M	***Mdn***	***SD***
a. I feel confident calling up a data file to view on the monitor screen	4.9	5	0.32
b. I feel confident working on a personal computer or laptop	4.7	5	0.48
c. I feel confident getting software up and running	4.4	5	0.84
d. I feel confident using the user's guide when help is needed	4.9	5	0.32
e. I feel confident entering and saving data (numbers or words) into a file	4.9	5	0.32
f. I feel confident escaping / exiting from a program or software	4.9	5	0.32
g. I feel confident calling up a data file to view on the monitor screen	4.6	5	0.52
h. I feel confident understanding terms/words relating to computer hardware	4.6	5	0.52
i. I feel confident understanding terms/words relating to computer software	4.6	5	0.70
j. I feel confident handling a CD-R/DVD correctly	4.7	5	0.48
k. I feel confident learning to use a variety of software applications	4.8	5	0.42
l. I feel confident making selections from an on-screen menu	4.9	5	0.32
m. I feel confident copying an individual file	4.8	5	0.42
n. I feel confident adding and deleting information from a data file	4.9	5	0.32
o. I feel confident moving the cursor around the monitor screen	4.9	5	0.32
p. I feel confident using the computer to write a letter or essay	4.6	5	0.52
q. I feel confident seeking help for problems with my computer	4.8	5	0.42
r. I feel confident using the computer to organise information	4.6	5	0.70
s. I feel confident getting rid of files when they are no longer needed	4.9	5	0.32
t. I feel confident organising and managing files	4.4	5	0.84
u. I feel confident troubleshooting computer problems	4.8	5	0.42
v. I feel confident browsing the World Wide Web (WWW)	4.8	5	0.42
w. I feel confident surfing the World Wide Web (WWW)	4.7	5	0.48
x. I feel confident finding information on the World Wide Web (WWW)	4.9	5	0.32
Results across participant group	***4.74***	***5***	***0.34***

† CSE uses a 5-point Likert scale where 1 = *I have very little confidence* and 5 = *I have a lot of confidence*. Adapted version of Murphy et al.'s [30] original Computer Self-Efficacy (CSE) scale, modified by Torkzadeh et al [31].

Recall that the purpose of the pre-session questionnaire was to collect basic demographic information and was designed to capture data on participants' IT efficacy and their perceptions of the current curriculum approval process. IT efficacy was measured using an adapted and modified version [31] of the CSE scale [30]. Internal consistency of the scales was tested using Cronbach's alpha and demonstrated "excellent" reliability (α= 0.952) [38]. Table 2 sets out the results of the CSE instrument used within the pre-session questionnaire. CSE results across the group revealed a high level of efficacy (*M* = 4.74; *Mdn* = 5). The ICT efficacy of participants was found to be very high across all CSE scale items, with little variation across the participant group (*SD* = 0.34). Such a high CSE score was anticipated given the academic composition of the participants.

Participants' perceptions of the existing curriculum approval process is summarised in Table 3. With such ordinal data it is conventional to consider the median values, which were largely neutral in nature (M_{ub} = 2.88; Mdn_{ub} = 3; SD_{ub} = 0.31). It should be noted that an unbalanced ($_{ub}$) Likert scale was used for this section owing to difficulties in positively wording those statements pertaining specific aspects of the curriculum approval process (i.e. *I, j, k, l*). Table 3 therefore separates positively and

[*] The extended nature of the results is such that their presentation has been combined with their discussion.

reverse coded results. Balanced ($_b$) results for the reverse coded results and balanced ($_b$) results across the entire participant group are also presented.

Scale reliability using Cronbach's alpha was found to be high (α= 0.862) and well above recognised reliability thresholds [38]. Balanced results across the participant group for all statements suggests a negative profile with general dissatisfaction with the current process (M_b = 2.68; Mdn_b = 2.5; SD_b = 0.55).

Examining the results for the positively coded statements separately reveals a negative profile for statements *a* – *h* with limited dispersion *(M* = 2.66; *Mdn* = 2.5; *SD* = 0.50). The profile of the reverse coded statements (*i* – *l*) almost mirrors the positively coded (*M* = 3.3; *Mdn* = 3.5; *SD* = 0.39). This can be verified by the balanced reverse coded results (M_b = 2.7; Mdn_b = 2.5; SD_b = 0.39). With the exception of statement *b* - which only demonstrated moderate approval (*M* = 3.3; *Mdn* = 4; *SD* = 0.95) - it is interesting to note that no single mean response suggested outright satisfaction with the current curriculum approval process, with participants inclined to view the current process as onerous and stifling class/course design (*k*) (Mdn_b = 4), or in needing improvements to render it more efficient (*l*) (Mdn_b = 4). This appears to be corroborated by statements *c* (*Mdn* = 2) and *g* (*Mdn* = 2).

Table 3: Results for the participant perception statements on the current curriculum approval process.

Current curriculum approval process: participant perception statements*	**Participant results**		
	M	***MDN***	***SD***
a. The curriculum approval process at the University of Strathclyde is an efficient process	2.6	2.5	0.97
b. The curriculum approval process at the University of Strathclyde is simple to understand	3.3	4	0.95
c. The curriculum approval process at the University of Strathclyde is a trivial process	1.8	2	0.79
d. The curriculum approval process at the University of Strathclyde is a process that demonstrates a quick turnaround time (i.e. time from submission to final approval)	2.3	2.5	0.82
e. The curriculum approval process at the University of Strathclyde is an effective process	3.1	3	0.74
f. The curriculum approval process at the University of Strathclyde is a process that is easy to manage	3.1	3	0.88
g. The curriculum approval process at the University of Strathclyde is a process that is well placed to respond to the demands from industry and the employment market	2.4	2	0.84
h. The curriculum approval process at the University of Strathclyde is a process that ensures quality teaching is delivered	2.7	2.5	1.06
Positively coded results	***2.66***	***2.5***	***0.50***
i. The curriculum approval process at the University of Strathclyde is a process requiring too many decisions by other people	2.9 (3.1)	3	0.88
j. The curriculum approval process at the University of Strathclyde is a convoluted process	3.1 (2.9)	3	0.74
k. The curriculum approval process at the University of Strathclyde is onerous and stifles innovation in course/module design	3.4 (2.6)	4 (2)	1.07
l. The curriculum approval process at the University of Strathclyde is a process requiring improvements to enhance efficiency	3.8 (2.2)	4 (2)	0.63
Reverse coded results	***3.3***	***3.5***	***0.39***
Reverse coded results ($_b$ = balanced)†	***2.7***	***2.5***	***0.39***
Results across participant group ($_{ub}$ = unbalanced)	***2.88***	***3***	***0.31***
Results across participant group ($_b$ = balanced)†	***2.68***	***2.5***	***0.55***

*Curriculum approval process perception statements use a 5-point Likert scale where 1 = *Strongly disagree* and 5 = *Strongly agree*. **Note** that statements *l, j, k,* and *l* were reverse coded.

† Reverse coded results balanced: *reverse score(x) = max(x) + 1 - x*

Post-session questionnaire data

Brooke's [32] System Usability Scale (SUS) formed the focus for the post-session questionnaire. The SUS instrument experiences wide use and has been subsequently developed, deployed and validated by other usability researchers (e.g. [33], [34]). The version of SUS used in this study included an adjustment to item 8, supplanting the word "cumbersome" for "awkward", as per the findings of Finstad [35] and research of Bangor et al. [33].

The results from the SUS are presented in Table 4 as are the individual SUS scores for each participant. SUS scores are calculated as follows: odd numbered items in the SUS are scored as the item score minus 1 and even items are scored as 5 minus the item score. This balances all scores and permits zeroes at the bottom of the range. The sum of the scores is then multiplied by 2.5. The resulting SUS score has a range of 0 to 100. The higher the SUS score, the easier a user feels it is to operate a system (i.e. C-CAP). SUS scores for individual items are included in Table 4 but are not in themselves meaningful; SUS produces a single value representing a combined measure of the overall usability of the system being studied.

Table 4: SUS scores per participant and group SUS results.

Brooke's System Usability Scale (SUS)[32]	**Individual participant SUS scores**			**Bangor et al's Adjective Rating Statement (ARS)**[33]
	#	***Faculty affiliation***	***SUS score***	***ARS score***
1. I think that I would like to use this system frequently	1	Strathclyde Business School	85	6
2. I found the system unnecessarily complex	2	Faculty of Science	67.5	5
3. I thought the system was easy to use	3	Strathclyde Business School	42.5	1
4. I think that I would need the support of a technical person to be able to use this system	4	Faculty of Science	80	6
5. I found the various functions in this system were well integrated	5	Strathclyde Business School	55	4
6. I thought there was too much inconsistency in this system	6	Faculty of Engineering	97.5	5
7. I would imagine that most people would learn to use this system very quickly	7	Faculty of Science	67.5	5
8. I found the system very awkward to use	8	Strathclyde Business School	77.5	5
9. I felt very confident using the system	9	Strathclyde Business School	75	5
10. I needed to learn a lot of things before I could get going with this system	10	Faculty of Science	87.5	5
	Group score (M)		***73.5***	***4.7***
	SD		***16.12***	***1.42***
	IQR		***16.25***	

The post-session questionnaire yielded an overall mean SUS score of 73.5 (*SD* = 16.12; *IQR* = 16.25). Researchers note [33] that "promising" SUS scores are generally > 70. A SUS score of 73.5 therefore places participants' perceptions of C-CAP at a favourable level. This SUS score increases to 77 when the outlying score for participant #3 is removed. It is also interesting to note that 40% of participants yielded SUS scores ≥ 80. Lowering the threshold further we note that 70% of participants generated SUS scores > 60. To supplement the SUS instrument and triangulate its findings, Bangor et al.'s [33] Adjective Rating Statement (ARS) was used (see Appendix G). The ARS is administered after the SUS questionnaire items and uses a 7-point scale from "Worst imaginable" to "Best imaginable", with the numeric values of 1 to 7 assigned respectively. This provides a qualitative response that can be used in combination with the SUS score to better interpret participants' overall experience with C-CAP.

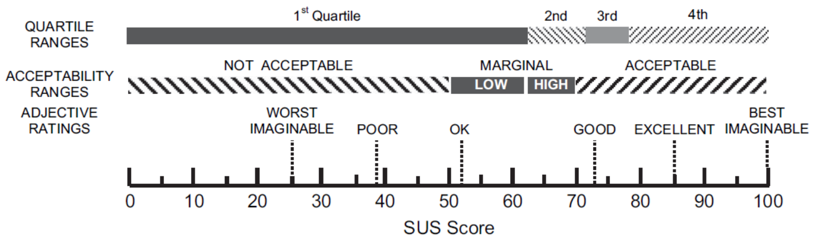


Figure 3: Comparative figure of SUS scores (by quartile), ARS and Bangor et al.'s [33] acceptability.

The post-session questionnaire yielded a mean ARS rating of 4.7 (*M* = 4.7; *SD* = 1.42), placing C-CAP within the “Good” ARS user-friendless category. Again, the ARS score increases and demonstrates less dispersion when outlying data are removed (*M* = 5.1; *SD* = 0.6). The mean ARS rating is consistent with Bangor et al.’s [33] validation of ARS with SUS and maps perfectly to Bangor et al.’s [33] SUS score guide and acceptability ranges (see Figure 3). Regression analysis appears to support the overall assertion that SUS scores predict ARS ratings in this instance (R^2 = 0.61, $F_{1,8}$ = 12.419, p < 0.01). It is nevertheless interesting to note that the SUS scores for participants #6 (*SUS* = 97.5; *ARS* = 5) and #10 (*SUS* = 87.5; *ARS* = 5) do not map comfortably to these acceptability ranges. This is borne out by the associated chart (Figure 4). For example, the SUS score for participant #6 was exceptionally high (*SUS* = 97.5) inferring an associated ARS score of 7 (“Best imaginable”; predicted *ARS* = 6.34); yet this participant represented a statistical anomaly by assigning an ARS score of 5 (“Good”). The lack of synergy between the SUS and ARS scores of participant #10 is less severe (*SUS* = 87.5; *ARS* = 5). Bangor et al.’s data is based on a far larger participant group (*n* = 212) which reveals levels of data variability not dissimilar to those presented in Table 4. It could be suggested that within a larger group the individual results of participants #6 and #10 would appear less anomalous. Such an anomaly in this case could therefore be attributable to the small participant numbers and the consequent lack of predictive power [39]. It should nevertheless be remembered that the overall SUS score for the participant group maps comfortably to Bangor et al.’s anticipated ARS rating and acceptability range. This places C-CAP within the 3rd quartile. It is possible that the perceived “goodness” of C-CAP is partly attributable to the high computer efficacy of the participant group, as demonstrated by a group CSE score of > 4.7.

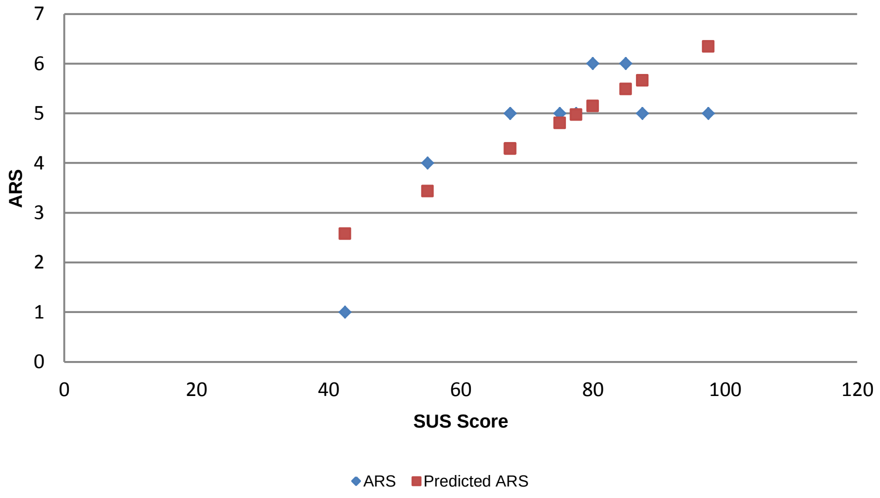


Figure 4: Predicted and actual ARS rating based on SUS score.

Recall that the post-session questionnaire also sought to capture perceptions of how C-CAP supported them in the curriculum design process and its potential for improving approval processes at the University of Strathclyde. Table 5 sets out the results for this section of the questionnaire instrument. Although positive values can be observed for statement *a* (*M* = 3.5*; Mdn* = 4*; SD* = 0.97), the overall results for this section were neutral (*M* = 3.12; *Mdn* = 3.2; *SD* = 0.91). The relatively high standard deviation reveals a high level of variation between participant responses, three of which were > 1. Such variability in the perceived potential of C-CAP to support participants in curriculum design and improve the approval process was a general theme that emerged from the protocol analysis and stimulated recall data, and appears to reinforce a dichotomy that emerged between participants' acceptance of the system and their understanding of the approval process.

Table 5: Post-questionnaire instrument: C-CAP participant statements.

C-CAP participant perception statements[†]	Participant results		
	M	*Mdn*	*SD*
a. The PiP system supports the curriculum design and approval process	3.5	4	0.97
b. The PiP system could greatly improve the curriculum design and approval process at the University of Strathclyde	2.9	3	1.10
c. The PiP system could support me in improving the pedagogical quality of curricula I design	2.9	3	0.88
d. The PiP system could support me in making curriculum design more efficient	3.3	3.5	1.16
e. The PiP system is sympathetic to the needs of my discipline	3	3	1.15
Results across participant group	***3.12***	***3.2***	***0.91***

[†] C-CAP participant perception statements use a 5-point Likert scale where 1 = *Strongly disagree* and 5 = *Strongly agree*.

3.2 Protocol analysis and stimulated recall data

Analysis of the qualitative data captured by the "think aloud" protocols, stimulated recall and open-ended questionnaire item (Q.3 of the post-session questionnaire) generated a detailed hierarchical coding framework (see Appendices A and B). This framework directed further querying of the data. Two super-nodes emerged from the data: *system issues*, and; *process and pedagogical issues*. These super-nodes contained 32 and 18 sub-nodes respectively and reflected the nature of the user acceptance evaluation, which was deliberately designed to elicit data on the extent to which C-CAP could support participants in the curriculum design and approval process. It was also designed to expose system and usability issues which were not identified during the heuristic evaluation (Phase 1). Interestingly, the qualitative data exposes among participants a dichotomy between the system and the curriculum design and approval process. This dichotomy will be explored in more detail later in this report.

Table 6: General word frequency query, including synonyms. (Top ten only.)

Word	Length	Count	Weighted Percentage (%)	Similar Words
class	5	246	2.43	*categories, category, class, classes, courses, forms, sorts, years*
think	5	159	1.26	*believe, consider, considered, guess, guessed, guessing, imagine, intended, means, reason, reasonably, recall, remember, remembering, suppose, supposed, think, thinking, thought*
assessment	10	112	1.18	*assess, assessed, assessment, assessments, evaluated, evaluation, value, values*
learning	8	144	1.15	*checking, determine, knowledge, knows, learn, learning, reading, readings, scholarships, seeing, study, teach, teaching*
students	8	63	0.69	*student, students*
hours	5	59	0.59	*hours, minutes*
objectives	10	54	0.59	*objective, objectives*
should	6	51	0.56	*should*
number	6	82	0.52	*amount, amounts, comes, coming, counts, figure, figures, issue, issued, issues, listing, lists, number, numbers, numerical, routinely, total*
activity	8	49	0.51	*activities, activity, dynamic, dynamics, participants*

Appendices A and B present the coding frameworks for the super-nodes. These frameworks detail all sub-nodes, node codes (to indicate hierarchical level), node definitions and indicative supporting quote(s). Columns for data references are also provided using the following definitions:

- **Sources:** Sources refers to the number of individual data sources (e.g. protocol analysis data, stimulated recall data, open-ended questionnaire responses) within which data has been coded at the associated node.
- **References:** References is a count of the number of selections within the source(s) that have been coded at a particular node.
- **Unique sources:** A unique source refers to the number of unique participants whose data has been associated with a particular node. Since most participants are associated with two or more data sources (e.g. protocol analysis data, stimulated recall data, open-ended questionnaire responses) and since multiple references to the same node may exist within any given source, a unique source count provides a means of determining how many participants have referred to particular node in their data.

For example, *Class rationale* (PPI:2.1 – Appendix A) has 9 sources, 14 references and 8 unique sources. This means that there exists 9 sources (likely a mixture of protocol analysis and stimulated recall data) within which 14 references to the node PPI:2.1 have been made. However, a unique source figure of 8 indicates that one participant has in fact referred to this node twice: once during protocol analysis and once during stimulated recall.

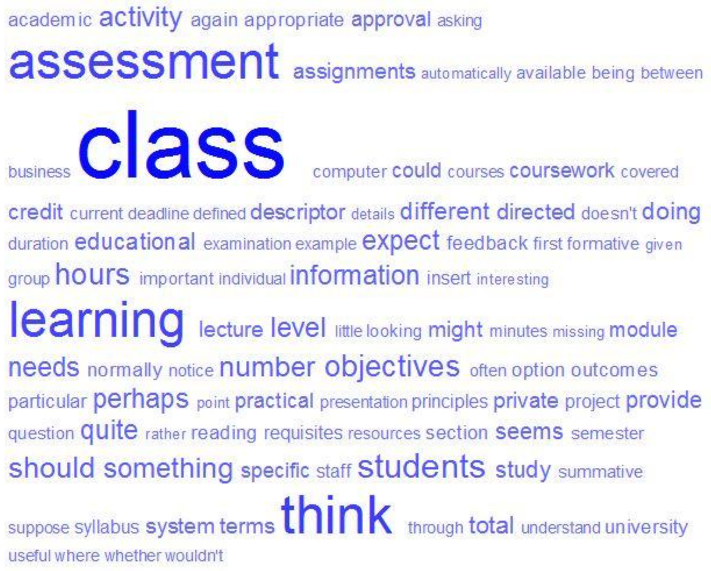


Figure 5: General word frequency query, including synonyms, diagrammed as a cloud.

A tree map diagramming the hierarchical nodes within the coding framework is provided in Appendix C. The result of a general word frequency query (with synonyms) is provided in Table 6 and is diagrammed as a cloud in Figure 5. These tend to reflect those aspects of the curriculum design process that participants found most difficult during the sessions (e.g. the design assessments and aligning them to stated learning objectives and/or outcomes, participant uncertainty over the credit-to-hours mappings used, etc.). Some of these issues will be revisited when the *process and pedagogical issues* super-node is discussed later in this section.

The following additional super-nodes were also created: *participant*; *participant attitudes* (i.e. mixed, negative, neutral, positive), and; *interesting quotes*. These additional super-nodes were used to facilitate data querying and did not to reflect the intellectual content of the data. They have therefore been omitted from the framework.

Although comprising 32 sub-nodes, the *system issues* framework primarily captures those C-CAP system issues that evaded exposure via the heuristic evaluation. Many of the nodes therefore address specific C-CAP functionality or system issues (e.g. *System navigation* [SI:2.9] or *Form submission errors* [SI:5.2]) or capture user requirement issues necessitating further investigation (e.g. *Dummy codes* [SI:2.3]). The *process and pedagogical issues* super-node comprises fewer sub-nodes, although some capture broader issues which are less conducive to enumeration. The nodes are too numerous and many are too trivial to discuss in detail here; for example, to facilitate the resolution of many interface or systems focused issues a table was derived from the protocol analysis data to assist in their prioritisation (see example in Appendix H). This table followed a format similar to the heuristic evaluation in phase 1 [2] and adopted a severity ratings system [28]. Suffice to state that the coding framework and its nodes will direct future C-CAP development work (to be completed prior to departmental / faculty piloting). We therefore restrict ourselves to further discussion of those nodes of substantive value.

Analysis of the data exposed participants' overall perception of the C-CAP system (*C-CAP perceptions* [SI:2]). C-CAP perceptions were generally positive, triangulating the positive SUS score from the post-session questionnaire instrument. Some participants frequently made positive comments throughout their interaction with the C-CAP system, with participants #9, #6 and #10 providing indicative comments:

> *It's actually very easy to use, in terms of development. It's quite intuitive. Ahhhh, much better... […] Generally the system is quite intuitive to use, so it's easy, it's straightforward.* (Participant #9)
>
> *So... read the information at the start is the first thing to do! It seems you can edit, which is quite useful. And there's help information as we go along. Good.* (Participant #6)
>
> *Lectures. Okay, so, this is lectures in hours, of which there are 48. But I guess we're going to have 24 lectures at 2 hours. Oh, it even does the maths for me! Splendid!* (Participant #10)

Some participants also commented in more detail on why their perceptions of C-CAP were generally positive. These more detailed comments often emerged from stimulated recall when the participant had an opportunity to reflect on their interactions with C-CAP. These comments were often more holistic insofar as they also considered the potential of C-CAP to improve the curriculum approval process. Said participant #4, for example:

> *It [C-CAP] has the potential to become a very efficient system in terms of both creating the approval system and going all the way to having a formalised descriptor document that one can present to staff and to students, saying "This is the class, this is what the class is about...". So in approving a class one has done the next step. Which, in a sense, we are already doing but in a paper based system. This is a draft class descriptor which is going to an academic committee tomorrow, and we will look at it and we will say "yes, that sounds like a very sensible class to be running". You can now apply for a class code, you now put it in the calendar. It now exists! Then they'll take that away, they'll update it and shove it all on the VLE. This can completely automate that process!*

However, the data also exposed participant hostility to the use of any system to aid the curriculum design and approval process. Participant #3 was perhaps most vocal in their disdain for the C-CAP system; and it should be noted that such fierce critiques were confined to this participant. The following illustrative quote from participant #3 was motivated by a C-CAP form submission error:

> *You see, this bothers me... This always bothers me about these things where you have these pre-set forms and you're entering information. I mean, it's easy for me to just use a form*

> *because when I'm sticking to a pre-set piece of software, y'know, I can't really see very well what I've written. And I hate that. If you can imagine, I did this under great pressure of time, um, and so that last thing I want to do is spend my time trying to figure out what it is I've just written. And then if I accidentally erase it...*

The aspect of C-CAP that perhaps inspired most comment from participants related to their experiences while using C-CAP to complete learning activity and assessment details. Sections 4.1 (Activity and delivery) and 4.3 (Assessment) require users to indicate the nature of the intended learning and assessment activities for the proposed class. Both sections were driven by drop down menus to promote efficiency in use and to minimise user error [28]. A notes box was also provided in section 4.3 to allow users to insert additional comments about their intended assessment activities. Although the values for these drop down menus mapped to the QAA's indicative learning and teaching methods list [40], almost all participants commented on the appropriateness of these values for their particular discipline and suggested alternatives (coded at *Option values* [SI:1.3] and *Learning activity options* [SI:4.2]). For example:

> *So these are very generic categories. So, "individual assignment", "group assignment", "group work", "group presentations"; all these things are all missing.* (Participant #5)
>
> *I was looking for a debate or presentation... It's quite narrow in terms of your descriptions of assessment. I would expect to see a break down between… A case study and a project are relatively similar, in a business context perhaps. Essay, report, presentation. Other formats we may use are debate, as I say; but we also... If you have an attendance requirement, in terms of they have to come to compulsory tutorials then that needs to be in as an assessment weighting as well because it tends to have marks attached to it.* (Participant #9)

In total 21 different learning activity types[*] and 16 different assessment activity types[†] were proposed by participants during the sessions. Data querying suggests that those participants proposing alternative learning or assessment activities were from outside the Faculty of Science and – although their proposed learning and assessment activities could be captured by the list and notes field – there was a perception that the values failed to reflect the "non-standard" teaching delivery methods or assessment techniques used by these faculties. Think aloud protocols from the following Strathclyde Business School participant were typical in this respect:

> *We've got labs, we've got tutorials, we've group activities, activity sessions, there's... It is, in essence... Everyone does lectures. We don't really have placements. Practicals? We don't do practicals - that's an Engineering view of the world. Fieldwork? Some courses do in the Business School, but not that many. That's more for HASS faculty staff. So, this should be a lot more extensive.* (Participant #5)

In other instances data suggest that the issue was primarily terminological. For example, some participants would not make the conceptual link between specific learning activities, such as a lab, and its practical nature ("Practical" – list value):

> *Right, okay, for the activity, actually, we've got a lecture, and also we have, from, erm, tutorial, which incorporate a lab as well. But, actually, but I cannot find this [lab] option for me; it doesn't provide other types of class session.* (Participant #1)

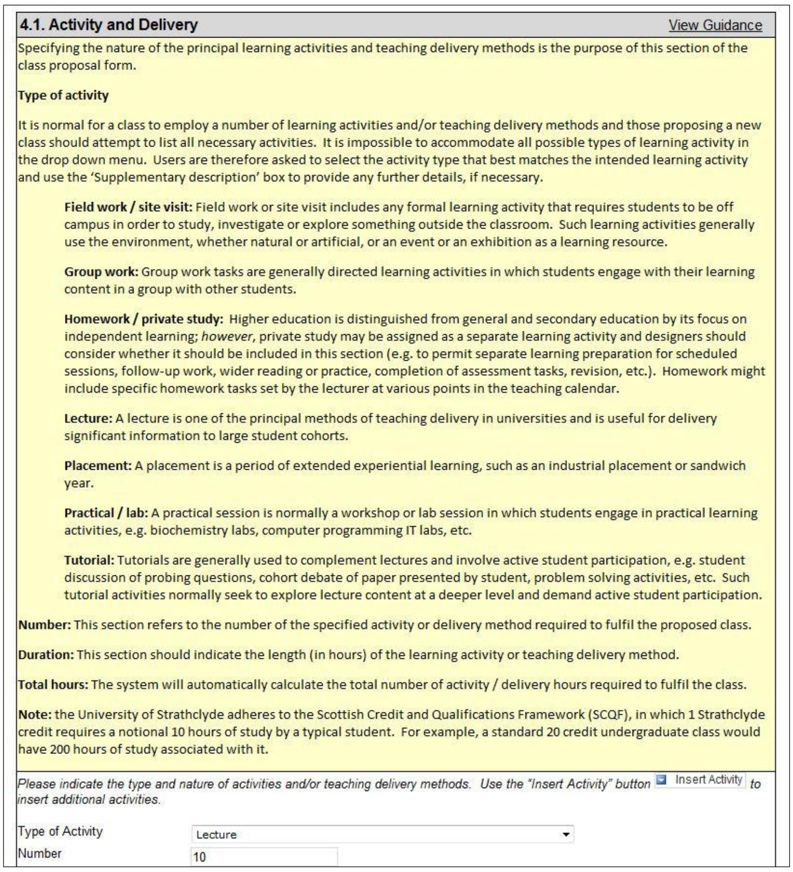
[*] Lecture, Tutorial, Seminar, Computer lab, Group work, Activity session, Group work, Group activities, Assignment, Individual reading, Interactive discussion, Class test, Site visit, Laboratory, Project work, Crit, Private study, Field work, Placement, Workshop, Presentation, Self-study

[†] Examination, Coursework, Class test, Lab books, Individual assignment, Group assignment, Group work, Group presentations, Debate, Presentation, Essay, Report, NCQ exam, Short answer exam, Attendance, Project

> *I would call them "computer labs". It doesn't really fit anywhere under those topics there. I would like to have "computer lab" added to the list of activities. Can I add it in manually? In that case, I will call my computer lab a "practical".* (Participant #6)

Others were also influenced in their suggestions by the way in which they perceived their teaching practice to differ from prevailing practice. In some instances this even called into question the legitimacy of the term “lecture” to describe a delivery method where an academic introduces ideas or delivers facts to a large group of students:

> *Probably I would put in there "Interactive discussion"; because when I lecture it's more a seminar than a lecture. Students come back and the pre-set lecture format often disappears. I am often sure I impart the analytical material I need to but students will ask questions... There's leeway. I would maybe put in a "Seminar", or something like that too.* (Participant #3)



4.1. Activity and Delivery

Specifying the nature of the principal learning activities and teaching delivery methods is the purpose of this section of the class proposal form.

Type of activity

It is normal for a class to employ a number of learning activities and/or teaching delivery methods and those proposing a new class should attempt to list all necessary activities. It is impossible to accommodate all possible types of learning activity in the drop down menu. Users are therefore asked to select the activity type that best matches the intended learning activity and use the ‘Supplementary description’ box to provide any further details, if necessary.

Field work / site visit: Field work or site visit includes any formal learning activity that requires students to be off campus in order to study, investigate or explore something outside the classroom. Such learning activities generally use the environment, whether natural or artificial, or an event or an exhibition as a learning resource.

Group work: Group work tasks are generally directed learning activities in which students engage with their learning content in a group with other students.

Homework / private study: Higher education is distinguished from general and secondary education by its focus on independent learning; *however*, private study may be assigned as a separate learning activity and designers should consider whether it should be included in this section (e.g. to permit separate learning preparation for scheduled sessions, follow-up work, wider reading or practice, completion of assessment tasks, revision, etc.). Homework might include specific homework tasks set by the lecturer at various points in the teaching calendar.

Lecture: A lecture is one of the principal methods of teaching delivery in universities and is useful for delivery significant information to large student cohorts.

Placement: A placement is a period of extended experiential learning, such as an industrial placement or sandwich year.

Practical / lab: A practical session is normally a workshop or lab session in which students engage in practical learning activities, e.g. biochemistry labs, computer programming IT labs, etc.

Tutorial: Tutorials are generally used to complement lectures and involve active student participation, e.g. student discussion of probing questions, cohort debate of paper presented by student, problem solving activities, etc. Such tutorial activities normally seek to explore lecture content at a deeper level and demand active student participation.

Number: This section refers to the number of the specified activity or delivery method required to fulfil the proposed class.

Duration: This section should indicate the length (in hours) of the learning activity or teaching delivery method.

Total hours: The system will automatically calculate the total number of activity / delivery hours required to fulfil the class.

Note: the University of Strathclyde adheres to the Scottish Credit and Qualifications Framework (SCQF), in which 1 Strathclyde credit requires a notional 10 hours of study by a typical student. For example, a standard 20 credit undergraduate class would have 200 hours of study associated with it.

Please indicate the type and nature of activities and/or teaching delivery methods. Use the “Insert Activity” button *to insert additional activities.*

Type of Activity: Lecture
Number: 10

Figure 6: Example of contextual help / guidance provided in section 4.1 (Activity and delivery) of C-CAP.

Kolås and Staupe [41] note the difficulties in attempting to systematise pedagogical design patterns in online contexts and it is therefore conceivable that similar issues were encountered when attempting to do the same with more traditional forms of pedagogy in C-CAP. One possible explanation could be participants’ resistance to using the context sensitive help, available in the top right hand corner of every section of C-CAP (Figure 5). Only one participant used the context sensitive help (participant #6), which included detailed guidance on the learning activity values available and their scope. Had participants been more inclined to view this help then they may have been more likely to perceive their peculiar teaching delivery methods to fall within the scope of C-CAP’s values. It may be that future C-CAP development work should better expose context sensitive help, either by pre-expanding the help sections so that users have to collapse them thus revealing its content, or by improving the visibility of the help features in a collapsible state. It is clear, however, that the large number of disparate list values (as proposed by a small number of participants) precludes inclusion as it would render sections of C-CAP unusable. Data derived from the list would also lack the specificity required for institutional reporting and wider curriculum management.

Aspects of section 4.3 (Assessment) that caused further confusion for many participants (*n* = 6) pertained to assessment deadline. The collection of such data is intended to encourage curriculum designers and course leaders to consider cohort assessment load during semesters. Many participants considered the collection of such information to be undesirable:

> *Again, the coursework would be issued across the entire duration of the semester, so there would be no specific deadline week. Y'know, it could be weeks three, five, seven, nine - so specifying the deadline week number doesn't help.* (Participant #2)

Or they considered it be unfeasible, because assessment activities and their deadlines are often only decided immediately prior to class delivery:

> *Deadline week numbers may vary, again, depending on how the coursework is split up. We don't know precisely how many pieces of coursework there might be. But the expectation is that there would be a minimum of two but probably a maximum of three. That's something that we might decide early on once we saw the number of people attending the course.* (Participant #7)

Others were more circumspect for reasons of teaching flexibility:

> *I'm fairly flexible with some the deadlines, actually. I wouldn't like to be prescriptive about it because I think it would vary a little bit according to the progress you make in terms of the lectures and labs. And that depends on the cohort of students and how quickly they learn. I do adapt it a bit in practice. I don't like these being too prescriptive. So I'd rather not have to have fixed deadlines.* (Participant #6)

Negative comments about these information requirements in C-CAP were a component of broader data themes pertaining to flexibility in teaching practice (coded at *Flexibility* [PPI:3.2]) and the perceived pointlessness of some curriculum design requirements in C-CAP (coded at *Form requirements* [SI:2.4] and [SI:2.10] *Unnecessary information*). Many participants reported their unease with drafting overly prescriptive curricula which might in future restrict their teaching practice and lead to further bureaucracy, whilst others felt it was disingenuous to provide prescription so far in advance of teaching delivery. The following protocol excerpts illustrate these varying participant viewpoints:

> *I want to just say there are four classes that take place this week, this week, that week. You know? It's almost as if there's too much information being asked in this. Some of this information should be given to the students by the department when they are delivering the class, rather than going in... making up the approval form.* (Participant #2)

> *Assessment description… Hmmmm… A general point here... The more detail we have to put in here in terms of the assessment, the more it becomes necessary to update this every year. Because, typically, you'd have maybe different assessments. That means modifying all these forms. So I'm not convinced a highly specified description of the assessment and when it's due is a good idea. It means more work and having to update it more regularly.* (Participant #6)

> *Typically you would want to be able to say what the assessment is, how long it lasts, if it's an exam, although that can't be a mandatory field. It's weighting. Timetabling information I wouldn't think is part of the approval process. Really the only timetabling information one needs at the approval process is whether it's an end of class examination or piece of in-class*

> *coursework, which is defined by the type of assessment. I'm not sure at the stage people are planning classes they would know enough about the structure of the class to be able to say, "Oh we're going to have a deadline in week 6 or 7". That, to me, is not relevant.* (Participant #4)

Finding a balance between the needs of the University (and ergo C-CAP) to improve pedagogy (e.g. promote more 'high impact' learning activities, greater alignment of assessment with stated learning objectives, etc.) and the information requirements of the centre (e.g. timetabling, estates management, library, procurement, etc.) on the one hand, and what academics are prepared to tolerate during curriculum design on the other, is clearly an area that requires further investigation by PiP. The curriculum descriptor structure and information requirements within C-CAP were derived from a number of extant forms used within the University and modelled the stated information requirements of key stakeholders (e.g. Educational Strategy Committee [42], Student Experience & Enhancement Services Directorate (SEES) [43], etc.). Restructuring of the forms in C-CAP and Phase 1 of the evaluation helped to rationalise the information demanded from users. Usability engineering techniques (such as heuristic evaluation) promote the use of efficiency tools to accelerate the speed with which users can complete tasks [28]; and it is possible that C-CAP requires further refinement in this respect in order to make the collection of such information less onerous for users. The role curriculum information can perform in improving the operational efficacy of the University was not fully recognised by several of the participants. Only those participants with administrative experience at higher academic levels (e.g. HoD) appreciated the significance of such information gathering by C-CAP. It is therefore possible that groups such as the Educational Strategy Committee need to better communicate the importance of such information for institutional monitoring, portfolio management and resource planning.

The *process and pedagogical issues* super-node contains 18 sub-nodes. The PiP project focuses on the potential of C-CAP to improve curriculum approval processes; but it is also within the remit of the project to explore the role C-CAP can perform in delivering new paths through which the University's range of policies and best practice guidelines on curriculum design can be brought to the fore in the minds of designers. Curriculum design represents a key "teachable moment" that is rarely exploited [44]. Indeed, it is often one of the few opportunities to influence the quality of the curricula that will eventually be delivered. One aspect of curriculum design that dominates educational literature is the idea of constructive alignment [29], [45], [46]; optimising assessments to best measure student learning against the stated learning objectives. The version of C-CAP used for the user acceptance evaluation therefore required participants to engage in constructive alignment (i.e. explicitly stating which assessments will assess which learning objectives); however, few participants viewed this requirement favourably. Data coded at *Aligning learning outcomes* [PPI:2.6.1] indicated that the majority of academics either considered their learning objectives to be assessed by all stated assessments, or felt it was irrelevant to include such detail as it can be highly ephemeral. For example:

> *So what do we mean by learning objectives assessment? It's actually all of them! Yeah, because I think it needs to reflect all objectives not just some.* (Participant #1)

> *For most of our classes the examination and coursework are essentially going to assess all of these things. So do I have to click four times to put them all in? It would be nice to have them altogether, I think. Because the exam is essentially going to assess the whole course...* (Participant #10)

> *It's not possible to pre-determine which learning objectives would be assessed by coursework. Because this may change from year to year… We don't pre-determine that. It's unlikely it would be all the learning objectives but I couldn't say in advance which it would be.* (Participant #7)

The process of aligning assessments with learning objectives in C-CAP was driven by inserting a new objective and then selecting from a drop down menu the objective which was to be aligned (Figure 6).

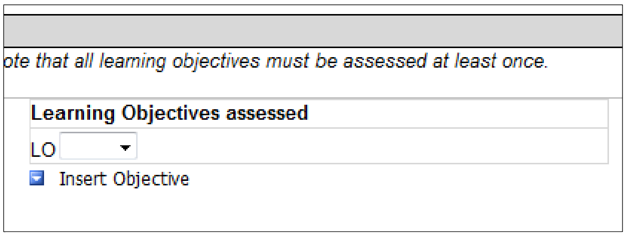

ote that all learning objectives must be assessed at least once.

Figure 7: Inserting learning objectives in C-CAP.

There were indications from the protocol data and the screen capture videos that the hostility towards aligning learning objectives was occasionally motivated by the awkwardness and tediousness of the alignment process in C-CAP:

> *And again, the examination is designed to assess all the learning outcomes, so I don't think that it's a helpful... well, from my point of view, it's a not a helpful thing. There should be a box that says "All". And that way you don't have to enter all five.* (Participant #2)

> *There is unnecessary repetition of clicking to add, e.g. learning outcomes to assessment…* (Participant #9)

It is possible that this aspect of C-CAP exerted higher levels of extraneous cognitive load on the participant, which in turn forced many to abandon the process of alignment altogether to seek interface options that would facilitate an “all objectives” solution. It is also possible that the artificial nature of the curriculum design task limited participants’ potential for creativity in this instance. Participants were replicating existing designs in C-CAP and although many had not explicitly aligned assessments with learning objectives in their original designs, many attended the testing session with the majority of their creative work essentially completed. These participants may therefore have felt disinclined to use C-CAP’s functionality in this respect. General participant antipathy towards rigorous adherence to standard curriculum design principles cannot be discounted either.

Neither did mandating constructive alignment appear to support C-CAP’s ability to promote greater reflection of assessment strategy [*Inspiring reflection* [PPI:2.5] AND *Aligning learning outcomes* [PPI:2.6.1]). Querying of the data indicates that only one participant considered C-CAP to inspire reflection during constructive alignment. This participant had experience of HoD responsibilities and was appreciative of C-CAP’s ambitions in this respect; but even this participant recognised the difficulties in implementing such a system more widely:

> *Learning objectives... assessment. I think... Interesting that one. It is clearly something which is beneficial to understand how the class works, and the students would better understand the linkage between what the class is meant to achieve and the assessment, but it's not something we routinely list. It is an additional and new idea. It [C-CAP] would force people to think a bit harder about their assessments and their learning objectives. I can see it being met with some... Hmmmm... worry, shall we say! Or people will simply say "all learning outcomes" and it will degenerate into an uninformative piece of information.* (Participant #4)

The data presented in Table 5 suggested that participants were generally positive about the potential of C-CAP to support them in curriculum design (*M* = 3.5; *Mdn* = 4; *SD* = 0.97) but were generally indifferent about the potential of C-CAP to improve their pedagogy or the quality of the curricula they design. Whilst some (like participant #4 above) could appreciate the potential of C-CAP in improving aspects of curriculum design or its potential to improve the departmental efficiency, data querying (*Curriculum approval* [PPI:1] AND *C-CAP perceptions* [SI:2]) appears to corroborate participants’ indifference, with only two participants commenting, one positively and one negatively. Participant #9 was positive about a relatively superficial aspect of the C-CAP system (i.e. form design) rather than the system itself:

> *I like this one, "Justify the need for the new course...", which is good. That first box makes you go through... makes you think clearly, erm, why the class is there in the first place [..]*

> *because there are too many classes that are put on the books with very small numbers. So... it's good.* (Participant #9)

Participant #3 (captured during stimulated recall) was vehement in their view that such a system usurped the creativity inherent to the curriculum design process and restricted innovative practice:

> *I found that this was a hindrance to good course design, because it was first of all tedious and everything is pre-set. I mean, just the thing about not being able to cut and paste things easily. You've got to type them in. And it comes back with errors, which is irritating. So, I found it wasn't conducive to thinking in an innovative way about a course the way I could when I sat down and.... Because originally, what I did, was I sat down and I just wrote down a course proposal. And then I was given a template which I was able to cut and paste things into. But if I had to sit and do it... I would never sit and do it from here. So what this is going to do is.... I will do this first and then I'll just have to sit down and do even more work, cutting and pasting and putting this in. So... And it's just.... You just feel that everything is standardised. There's no leeway to add something that is distinctive about the course. So I found it kind of like a straitjacket.*
>
> *[…]*
>
> *If we're going to be forced to fill these things out… I will not work from this to design a course so, for me, it's useless. I would just do it this way [in MS Word] and then I would.... So it's really for the people who are approving the course, from my point of view. In my opinion I would not have come up with the courses I did if this was what I was working from, for sure - no way! And I think I've designed an excellent course, as external experts in the field have said; so I think it could suffer as a result.* (Participant #3)

Again, it is interesting to note that in many cases the depth of information requested via C-CAP – and the structure of the information requested - was consistent with several extant curriculum descriptors used at the University of Strathclyde or was rendered more efficiently for users (e.g. accelerators to speed up interaction with C-CAP). It is therefore apparent that negative comments such as those from participant #3 are more a consequence of the approval requirements mandated by the institution than C-CAP. Stimulated recall with participant #3, for example, sought further clarity on the participant's issues with the University's 12 Principles of Assessment and Feedback [47], which provoked the following response:

> *The idea, the innovation in the course; the thing that's going to make this course different from a course offered anywhere else is nothing to do with whether I'm able to think about the University's Principles of Assessment. It's completely convoluted. […] There's too much emphasis on this sort of stuff. I just think back to my own background, where I was taught at a university where professors had Nobel prizes. They were not sitting down designing the fantastic courses that I took with them with this sort of stuff. It's just... It's always this thing that "we're not doing enough"; this second guessing. This thing where you have to put everything in the form of language that really... You're often struggling to understand what they are getting at. Where the most important thing - the substantive content of the course - it comes secondary. I won't use it. Honestly. I wouldn't have designed the course, as I said, as I did.* (Participant #3)

It should be noted that the views of participant #3 were exceptional and no other participant commented quite so negatively during protocol analysis or stimulated recall. Nevertheless, this participant represents a particular academic viewpoint about which the PiP project needs to be cognisant. Communicating to similarly-minded academics of the benefits to curriculum design and approval, institutional monitoring of students' educational experience, portfolio management, resource planning and the operational efficiencies to be achieved with C-CAP will be essential to ensure successful advocacy were such a system to be implemented across the institution.

Participants often expressed uncertainty about aspects of the approval process and certain information requirements. An aspect of the design process which caused uncertainty among participants - and area in which C-CAP could incorporate additional user support – pertains to the relationship between credit weightings of the class being proposed and the required number of student study hours. Many participants (*n* = 6) discussed this aspect of design in their protocols extensively, such that it is reflected in Table 6 (which notes “hours” and “numbers” as two of the most mentioned words in the qualitative data). This issue was perhaps most acute in the number of student study hours associated with 20 credit classes. Although participants were replicating an existing curriculum approval form in C-CAP, many descriptors had originally been ambiguous about the number of student study hours associated with their class, perhaps because faculty administration or academic quality teams clarified the study hour expectations after the substantive content had been submitted. The uncertainty experienced by participants in some cases appears to be attributable to their reliance on faculty staff; but their uncertainty also appears to validate an original aim of PiP: to provide academics with a suite of discipline specific curriculum designs (i.e. patterns) that could be used as the basis for pedagogical innovation and the development of new curricula. Such designs would enable academics to focus on innovative curriculum design safe in the knowledge that the ‘foundations’ were sound.

The University of Strathclyde adheres to the Scottish Credit and Qualifications Framework (SCQF) [48] which, in turn, maps to the European Qualifications Framework (EQF) [49]. The SCQF promotes a notional 10 hours of study by a typical student per academic credit [50]. This means that a typical 20 credit class should have 200 hours of student study associated with it. Data querying extracted two passages that illustrate the uncertainty some academic staff have about University curriculum approval requirements:

> *I don't know if I've ever seen it written down, exactly how many hours there should be for 10 credits; but I've heard informally that it should be about 100 hours. And I assume that that includes students doing their assessments... assessment activity. I may be wrong, but that's what I've heard.* (Participant #6)

> *Perhaps if there's a standardised model in terms of the number credits that you put in? Perhaps there should be a total hours of activity that you've got to get to?* (Participant #9)

As might be expected, the protocols also revealed inconsistent practices between faculties and across a number of areas; however, this appeared to extend to what academics considered to constitute compulsory study activity when assigning class study hours. For example, some included hours towards summative assessment, while others expected the time spent on completing assessments to be in addition to the stated study hours. Some participants also acknowledged the disparate practice and its absurdity from an operational perspective:

> *We expect you to spend two hours on them, so there would be 24 hours load associated with that. It's not covered there, and if you look at the way our form is laid out. You've "Practical"... It's specific to Science, I suppose. If that wasn't running.... Erm, the devolved nature of the University allows different Faculties to do different things, which is stupid!* (Participant #2)

Improved guidance and support tools to flag when classes are under or over the credit-to-hours threshold would therefore be a useful addition to C-CAP, and would help to reduce the faculty burden associated with resolving trivial curriculum design errors. However, there is clearly a need to clarify curriculum design practice across the institution to, a) make the process and its requirements more transparent to academics, and b) to establish equitable learning pathways for students, particularly as radical differences in assessment practice and study hours allocation can be found within small investigations such as this. It is apposite to note that previous work conducted under the auspices of PiP [51] found that one of the principal obstructions to efficient curriculum approval was the failure of academics to meet the faculty paperwork requirements. This frequently creates additional work for faculty staff and often delays the approval process unnecessarily as staff are then required to pursue academics for clarification on the details of the proposed curricula, or to deliver feedback to the authors of rejected submissions. Supporting faculty in the approval process is an important aspect of C-CAP. C-CAP, for instance, uses techniques to reduce careless errors in forms and promotes "good" curriculum designs; but clearly there is a wider need to better communicate the expectations of the curriculum design and approval process, and to make the requirements of design more transparent to academics, many of whom are misinformed about the process [51]. C-CAP can be viewed as vital to achieving this since C-CAP exemplifies - and seeks to standardise - the curriculum approval process. This assumption will be tested during the next evaluative strand of PiP (WP7:38 - Impact & process evaluation).

4. Conclusion

This report has sought to summarise the methodological approach and principal findings of phase 2 of WP7:37. This phase was principally concerned with assessing the extent to which C-CAP functionality met users' expectations within specific curriculum design tasks and evaluating the performance of C-CAP in supporting curriculum design tasks and the approval process, as well as its potential for improving pedagogy. Measuring the overall usability of C-CAP (e.g. interface design and functionality instinctive, navigable, etc.), capturing data on users' preferred system design/features, and eliciting data on current approval processes and how C-CAP could contribute to improvements in the process, were also an additional aims of this evaluative phase. This phase of evaluation has therefore focussed on a small but nevertheless important aspect of the overall PiP evaluation plan [1]. Piloting of C-CAP within faculties will form the basis for the next evaluative strand (WP7:38 - Impact & process evaluation) in which rich qualitative data is expected to be gathered (via group interviews and MSC stories).

In this phase of evaluation C-CAP, as a system, was positively received, achieving a positive SUS score and ARS rating. Whilst this could be partially attributable to the high computer efficacy of the participants, protocol and stimulated recall data did reveal that participants were, in general, favourably disposed to the C-CAP system. Numerous problems with the usability of C-CAP were nevertheless identified and it is the intention of PiP to implement appropriate modifications to enhance user acceptance. Users' preferences will also be incorporated where possible.

It is clear, however, that a dichotomy exists between the *system* (which received generally positive feedback) and the overall curriculum design *process*, which was less well received. Although no such data was collected from participants, anecdotal evidence indicated that those participants who had been exposed to the curriculum approval process from a managerial perspective (e.g. as a Head of Department or Vice Dean) were the most encouraged by the potential of C-CAP to assist in the approval process; their views clearly influenced by their professional practice and an holistic understanding of the approval process issues involved. Whilst other users lacked this insight, data from both quantitative and qualitative sources indicated that all participants were dissatisfied with the existing process, tacitly acknowledging that adjustments and improvements were justified. At many stages in their interactions with the C-CAP system, participants were not required to produce more information than they otherwise would; yet the demands of the University's policies and regulations on curriculum approval meant that many participants were unconvinced of the overall process, as facilitated by C-CAP. In this respect it could simply be that the forms served by C-CAP – although based on existing curriculum descriptors – were sufficiently different to give the impression that large amounts of additional data was being collected. It could also be surmised that the pressures of increased teaching loads and departmental research expectations have made academics increasingly sceptical of the merits of new IT systems; but, as we have also observed, hostility to improved specificity in curriculum design has links to strongly held views on academic freedom and attitudes that novel educational concepts are antithetical to HE teaching contexts. There is therefore a need to clarify curriculum design practice across the institution to render the process and its requirements more transparent to academics, and to establish equitable learning pathways for students, particularly as radical differences in assessment practice and study hours allocation were found to exist. From this perspective, C-CAP can, over the longer term, be viewed as integral to achieving this since it embodies and seeks to standardise the curriculum approval process.

Given the methodological restrictions imposed on the PiP project, the evaluative approach adopted was of value and exposed rich data on a multitude of systems focussed and process issues which can guide further development prior to departmental / faculty piloting (WP7:38). Data will also inform wider recommendations to key stakeholders, such as the SEES Directorate [43] and the Educational Strategy Committee [42], on how best to advocate C-CAP as a tool to improve operational efficiency and educational quality.

Future research attempting to test the efficacy of technology supported approaches to curriculum design should seek to model the 'real world' design process more accurately. Perhaps the most disappointing finding was C-CAP's failure to inspire reflection or creativity among the majority of participants during the curriculum design process (leading to improved designs). Whilst the results and discussion section of this report (section 3) identified areas of C-CAP that could be improved to inspire such creativity, it is probable that the artificial nature of the curriculum design task compromised our ability to engage participants in the task sufficiently, particularly as many would have already invested creativity in their original curriculum designs. It is nevertheless hoped that the next evaluative strand (WP7:38) will enable an improved understanding of C-CAP's potential in this respect. Future work should instead employ 'design diaries' in which participants would note or verbalise their experiences designing curricula with C-CAP. Verbalisations and reflections could be captured via video diary [52]. Such an approach would lack the control enjoyed by the current study but would, a) yield useful data on how C-CAP can stimulate new curricula, b) would allow time for users to improve their C-CAP efficacy, and c) would enable participants to reflect upon their designs and how C-CAP inspired the adoption of innovative designs. Participant numbers need not exceed ten, as patterns in participant responses quickly emerge; but recruiting participants with greater knowledge of the administrative bottlenecks involved in curriculum approval would also yield richer data on the merits of the system in expediting the approval process.

5. References

[1] G. Macgregor, 'PiP Evaluation Plan (Draft)', University of Strathclyde, Glasgow, Nov. 2011. [Online]. Available: http://www.principlesinpatterns.ac.uk/Portals/70/PiPEvaluationPlan.pdf. [Accessed: 26-Jul-2012]

[2] G. Macgregor, 'Heuristic Evaluation of Course and Class Approval Online Pilot (C-CAP)', University of Strathclyde, Glasgow, Dec. 2011. [Online]. Available: http://www.principlesinpatterns.ac.uk/Portals/70/pip%20document%20library/ProjectReports/WP7-37-1heuristicevaluation.pdf. [Accessed: 26-Jul-2012]

[3] J. E. Smith and A. M. Brown, 'Building a culture of learning design: Reconsidering the place of online learning in the tertiary curriculum', pp. 615 –623, 2005.

[4] F.-Q. Lai, 'Five Tens, Eighteen Circles, & Online Learning of Educational Technology in China', *International Journal of Technology in Teaching and Learning*, vol. 3, no. 2, pp. 69–84, 2007.

[5] 'Principles In Patterns', 2012. [Online]. Available: http://www.principlesinpatterns.ac.uk/. [Accessed: 29-Feb-2012].

[6] 'T-SPARC', 2012. [Online]. Available: http://blogs.test.bcu.ac.uk/tsparc/. [Accessed: 29-Feb-2012].

[7] S. Knight, 'Institutional approaches to curriculum design', 2012. [Online]. Available: http://www.jisc.ac.uk/curriculumdesign. [Accessed: 29-Feb-2012].

[8] J. Sweller, J. J. G. van Merrienboer, and F. G. W. C. Paas, 'Cognitive architecture and instructional design', *Educational Psychology Review*, vol. 10, no. 3, pp. 251–296, 1998.

[9] P. Chandler and J. Sweller, 'Cognitive Load Theory and the Format of Instruction', *Cognition and Instruction*, vol. 8, no. 4, pp. 293–332, 1991.

[10] J. S. Ahuja and J. Webster, 'Perceived disorientation: an examination of a new measure to assess web design effectiveness', *Interacting with Computers*, vol. 14, no. 1, pp. 15–29, Dec. 2001.

[11] J. P. Tracy and M. J. Albers, 'Measuring Cognitive Load to Test the Usability of Web Sites', in *Proceedings of the Annual Conference for Technical Communication 2006*, 2006, pp. 256–260.

[12] S. Oviatt, 'Human-centered design meets cognitive load theory: designing interfaces that help people think', in *Proceedings of the 14th annual ACM international conference on Multimedia*, New York, NY, USA, 2006, pp. 871–880.

[13] R. Ignacio Madrid, H. Van Oostendorp, and M. C. Puerta Melguizo, 'The effects of the number of links and navigation support on cognitive load and learning with hypertext: The mediating role of reading order', *Computers in Human Behavior*, vol. 25, no. 1, pp. 66–75, Jan. 2009.

[14] P. Schmutz, S. Heinz, Y. Métrailler, and K. Opwis, 'Cognitive load in ecommerce applications: measurement and effects on user satisfaction', *Adv. in Hum.-Comp. Int.*, vol. 2009, pp. 3:1–3:9, Jan. 2009.

[15] D. DeStefano and J.-A. LeFevre, 'Cognitive load in hypertext reading: A review', *Computers in Human Behavior*, vol. 23, no. 3, pp. 1616–1641, May 2007.

[16] G. Conole, 'Learning design - making practice explicit', presented at the ConnectEd Design Conference, 28 June - 2 July 2010, Sydney, Australia, 2010.

[17] K. Ericsson and H. Simon, *Protocol analysis*. Cambridge (Mass.) ;;London: The MIT Press, 1985.

[18] M. van Someren, Y. Barnard, and J. Sandberg, *The think aloud method : a practical guide to modelling cognitive processes*. London ;;San Diego: Academic Press, 1994.

[19] R. BenbunanFich, 'Using protocol analysis to evaluate the usability of a commercial web site', *Information & Management*, vol. 39, pp. 151–163, Dec. 2001.

[20] Z. Guan, S. Lee, E. Cuddihy, and J. Ramey, 'The validity of the stimulated retrospective think-aloud method as measured by eye tracking', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2006, p. 1253.

[21] M. van den Haak, M. De Jong, and P. Jan Schellens, 'Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue', *Behaviour & Information Technology*, vol. 22, pp. 339–351, Sep. 2003.

[22] C. Cool and I. Xie, 'Affective utterances as contextual feedback in interactive information retrieval', in *Proceedings of the third symposium on Information interaction in context (IIiX '10)*, 2010, p. 277.

[23] H. Terai, H. Saito, Y. Egusa, M. Takaku, M. Miwa, and N. Kando, 'Differences between informational and transactional tasks in information seeking on the web', in *Proceedings of the third symposium on Information interaction in context (IIiX '08)*, 2008, p. 152.

[24] D. Kelly, *Methods for evaluating interactive information retrieval systems with users*. Hanover MA: now Publishers, 2009.

[25] D. Lottridge, M. Chignell, and S. E. Straus, 'Requirements analysis for customization using subgroup differences and large sample user testing: A case study of information retrieval on handheld devices in healthcare', *International Journal of Industrial Ergonomics*, vol. 41, pp. 208–218, May 2011.

[26] M. Jaspers, T. Steen, C. Bos, and M. Geenen, 'The think aloud method: a guide to user interface design', *International Journal of Medical Informatics*, vol. 73, pp. 781–795, Nov. 2004.

[27] T. Boren and J. Ramey, 'Thinking aloud: reconciling theory and practice', *IEEE Transactions on Professional Communication*, vol. 43, pp. 261–278, Sep. 2000.

[28] J. Nielsen, *Usability inspection methods*. New York: Wiley, 1994.

[29] J. B. B. (John Biggs, *Teaching for quality learning at university.*, 3rd ed. / by John Biggs and Catherine Tang. Maidenhead: Open University Press, 2007.

[30] C. A. Murphy, D. Coover, and S. V. Owen, 'Development and Validation of the Computer Self-Efficacy Scale', *Educational and Psychological Measurement*, vol. 49, pp. 893–899, Dec. 1989.

[31] G. Torkzadeh, J. C.-J. Chang, and D. Demirhan, 'A contingency model of computer and Internet self-efficacy', *Information & Management*, vol. 43, pp. 541–550, Jun. 2006.

[32] J. Brooke, 'SUS - A quick and dirty usability scale', in *Usability evaluation in industry*, London: CRC Press, 1996, pp. 189–194.

[33] A. Bangor, P. T. Kortum, and J. T. Miller, 'An Empirical Evaluation of the System Usability Scale', *International Journal of Human-Computer Interaction*, vol. 24, no. 6, pp. 574–594, 2008.

[34] J. R. Lewis and J. Sauro, 'The Factor Structure of the System Usability Scale', in *Proceedings of the 1st International Conference on Human Centered Design: Held as Part of HCI International 2009*, Berlin, Heidelberg, 2009, pp. 94–103.

[35] K. Finstad, 'The System Usability Scale and Non-Native English Speakers', *Journal of Usability studies*, vol. 1, no. 4, pp. 185–188, 2006.

[36] J. Sauro and E. Kindlund, 'A method to standardize usability metrics into a single score', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, New York, NY, USA, 2005, pp. 401–409.

[37] 'Bristol Online Surveys (BOS)', 2012. [Online]. Available: http://www.survey.bris.ac.uk/. [Accessed: 29-Feb-2012].

[38] Darren George, *SPSS for Windows step by step : a simple guide and reference, 12.0 update*, 5th ed. Boston: Pearson Education, 2005.

[39] S. B. Green, 'How Many Subjects Does It Take To Do A Regression Analysis', *Multivariate Behavioral Research*, vol. 26, no. 3, pp. 499–510, 1991.

[40] HESA, 'HESA - Higher Education Statistics Agency', *Calculation of assessment methods and learning and teaching methods*, 2011. [Online]. Available: http://www.hesa.ac.uk/component/option,com_studrec/task,show_file/Itemid,233/mnl,12061/href,Calculations_methods.html/#LearningandTeaching. [Accessed: 26-Feb-2012].

[41] L. Kolås and A. Staupe, 'Implementing delivery methods by using pedagogical design patterns', *EDMEDIA 2004*, vol. 2004, no. 1, pp. 5304–5309.

[42] University of Strathclyde, 'Education Strategy Committee', 2012. [Online]. Available: http://www.strath.ac.uk/committees/strategiccommittees/educationstrategycommittee/. [Accessed: 08-Mar-2012].

[43] University of Strathclyde, 'SEES Directorate', 2012. [Online]. Available: http://www.strath.ac.uk/sees/seesdirectorate/. [Accessed: 08-Mar-2012].

[44] P. Bartholomew and J. Everett, 'Socio-technical ramifications of a new approach to course design and approval', presented at the JISC Innovating e-Learning Online Conference November 2011, 2011.

[45] A. Walsh, 'An exploration of Biggs' constructive alignment in the context of work-based learning', *Assessment & Evaluation in Higher Education*, vol. 32, no. 1, pp. 79–87, 2006.

[46] C. Rust, 'The Impact of Assessment on Student Learning', *Active Learning in Higher Education*, vol. 3, no. 2, pp. 145–158, Jul. 2002.

[47] University of Strathclyde, '12 Principles of Good Assessment & Feedback - University of Strathclyde', 2008. [Online]. Available: http://www.strath.ac.uk/learnteach/teaching/staff/assessfeedback/12principles/. [Accessed: 28-Feb-2012].

[48] University of Strathclyde, 'University procedure and guidelines on course and class approval', University of Strathclyde, Glasgow, 2009.

[49] 'European Commission - Education & Training - lifelong learning policy - The European Qualifications Framework (EQF)'. [Online]. Available: http://ec.europa.eu/education/lifelong-learning-policy/doc44_en.htm. [Accessed: 28-Feb-2012].

[50] 'Scottish Credit and Qualifications Framework - Home'. [Online]. Available: http://www.scqf.org.uk/. [Accessed: 13-Jan-2012].

[51] D. Cullen, J. Everett, and C. Owen, 'The curriculum design and approval process at the University of Strathclyde: baseline of process and curriculum design activities', University of Strathclyde, Glasgow, 2009. [Online]. Available: http://www.principlesinpatterns.ac.uk/Default.aspx?tabid=2923. [Accessed: 26-Feb-2012]

[52] S. Carter and J. Mankoff, 'When participants do the capturing: the role of media in diary studies', in *Proceedings of the SIGCHI conference on Human factors in computing systems*, New York, NY, USA, 2005, pp. 899–908.

6. Appendix A: Coding framework: Process and pedagogical issues (super-node)

Table 7: Coding framework for the super-node "Process and pedagogical issues" only.

Super-node: Process and pedagogical issues						
Node code	**Node**	**Node definition / scope note**	**Example quote(s)**	**Sources**	**References**	**Unique source**
PPI:1	Curriculum approval	Content coded at this node captures participant views on the current curriculum approval process, or the potential for C-CAP to impact upon a future approval process.	*“It [C-CAP] has the potential to become a very efficient system in terms of both creating the approval system and going all the way to having a formalised descriptor document that one can present to staff and to students, saying ‘This is the class, this is what the class is about...’. So in approving a class one has done the next step, which, in a sense, we are already doing but in a paper based system. This is a draft class descriptor which is going to an academic committee tomorrow, and we will look at it and we will say ‘yes, that sounds like a very sensible class to be running’. You can now apply for a class code, you now put it in the calendar. It now exists! Then they'll take that away, they'll update it and shove it all on the VLE. This can completely automate that process!”*	4	7	3
PPI:2	Curriculum design	Content coded at this node relates to participant experience or issues with the practical aspects of curriculum design or their knowledge of curriculum design theory and/or practice.	*“I find these kinds of questions - Educational Aim - um, and rationale... I just find... I get a little irritated by these sorts of things because I sort of feel because it could be answered in the one go. And then I have to sort of think, ‘What are they wanting me to answer here?’, as opposed to rationale.”*	5	11	4
PPI:2.1	Class rationale	Content coded at this node concerns participant views or uncertainty over providing a rationale for a class (esp. section 3.1 of C-CAP), e.g. general views of its applicability, unsure what information should be provided, unnecessary because they feel it has already been provided elsewhere (i.e. course specification).	*“Provide rationale...blah...blah. A lot of this information will already be there in the programme specification, so it seems, sort of, it is being included for no additional value. Providing evidence for the need for the new class; that's normally something we wouldn't have. We would have a rationale for the class, in terms of scope; but things like employers, etc. would be in the covering note. And now we're onto classes, not courses...”*	9	14	8
PPI:2.2	Course linkage	Data coded at this node pertains to the numerous links that can exist between the classes that comprise a course and any issues therein.	*“My goodness, I have to put in all the courses that this is part of, which is of the order of 10 different courses? Because... It could be optional.... Well, we have 10 different degrees: Maths, Stats and Accounts, Maths, Stats and Management Science, Maths and Physics, Maths and Computer Science... So it looks like I have to put everything in here for each one, which is not so good. I'll just enter one for now; but that's just an observation.”*	1	1	1
PPI:2.3	Credit weightings	Data coded at this node evidences wider pedagogical and curriculum design issues with respect to credit weightings and their association with activity hours, including participant uncertainty on the regulations.	*“I mean, there's 20 credits, but how those would be divided up, um, that would require more information, which is something I haven't really considered at this stage.”* *“Another thing is private study. There's obviously, um, there must be a template out there that says that if a course is worth 20 credits the student should be spending a certain amount of time in private study. I mean, I would hope they would go off and study privately but, y'know, I don't know why I always have to say that. Y'know, if I say private study "5 hours", that's going to look ridiculous. It would be good to have the mapping of what's expected. I know it's out there but it's not in my head. But then again, I wouldn't be sitting here doing this. I would probably go and find out and then enter it in.”*	5	9	4
PPI:2.4	Disparate practice	Content coded at this node denotes participants' perceptions of differing curriculum design practice within the University. This might across faculties or within departments.	*“Several of the questions did not correspond to SBS requirements, while several other questions used language that was appropriate for other faculties or did not include SBS relevant terms. The system needs to be appropriate for all faculties or customisable by relevant Academic Committees.”*	3	5	3
PPI:2.5	Inspiring reflection	Content at this node captures participant views on the potential for C-CAP to	*“Learning objectives... assessment. I think... Interesting that one. It is clearly something which is beneficial to understand how the class works, and the students would understand*	3	6	3

Super-node: Process and pedagogical issues						
Node code	**Node**	**Node definition / scope note**	**Example quote(s)**	**Sources**	**References**	**Unique source**
		inspiring reflection in the curriculum design process.	*the linkage between what the class is meant to achieve and the assessment, but it's not something we routinely list. It is an additional and new idea. It would force people to think a bit harder about their assessments and their learning outcomes."* *"I think that's really, really derogatory, to think that, y'know, people sit down and they're not... Because this takes away the thought. The idea, the innovation in the course; the thing that's going to make this course different from a course offered anywhere else is nothing to do with whether I'm able to think about the University's principles of assessments. It's completely convoluted."*			
PPI:2.6	Learning outcomes	Content coded at this node denotes data relating to participant comments about learning outcomes.	*"What's the difference between a learning outcome and a learning objective? Right, okay, we would... four... now we have a very bland learning outcomes statement here on this class; but many others we specify very tightly what we expect the students to demonstrate a knowledge of and an ability to use. And then saying... Limiting it to four is not necessarily valid. Unless, of course, you put learning outcome 1, "Students shall show a basic understanding of dynamics, which will include a knowledge of X, Y Z". But that's then... circumventing... cheating."*	10	32	8
PPI:2.6.1	Aligning learning outcomes	Data coded at this node pertains to participant difficulties in aligning learning outcomes/objectives, e.g. difficulty aligning with assessment, desire to assess all outcomes, etc.	*"For most of our classes, the examination and coursework are essentially going to assess all of these things. So do I have to click four times to put them all in? It would be nice to have them altogether, I think. Because the exam is essentially going to assess the whole course."* *"Instead of just matching learning objectives to assessment you need to map your learning outcomes to your assessment, which is equally as important as objectives. In my opinion they are different things".*	9	11	7
PPI:2.6.2	Cognitive outcomes	Data coded at this node explores the additional need for C-CAP accommodation of - or University wide adoption of - cognitive based outcomes. These are typically transferrable skills which students are likely to acquire or develop in addition to discipline specific learning outcomes.	*"The only piece of information that I'm aware of that this online system hasn't asked me for that I would normally provide, either on a class descriptor or through the class approval process, what's called "key skills linkages", which we often ask - certainly within my own department ask for. So we would ask, what generic skills, key skills are covered by this class. So... verbal skills, academic skills, analytical skills... and... they are the framework of key skills which were produced many, many years ago, which we follow. I don't know whether that's still current or not..."* *"What we do is we have learning objectives and we also have learning outcomes, in terms of subject specific knowledge and skills that the students are developing and the general cognitive and non-subject specific skills. I think you need an additional two sections in there to cover those things."*	4	5	3
PPI:2.6.3	Syllabus	Node denotes content at which syllabus is discussed.	*"In summarising the syllabus, one of the issues that came up when designing the course was that we noted that these items were not all of the same weighting. That there would be more attention given to item two. So simply listing them all as individual bullets tends to obscure that aspect, even in the paper version of the course description."* *"Why is the syllabus disconnected from the learning objectives? I would say that you define the learning objectives and then you put the syllabus in place to support those learning objectives. So I would have thought the natural flow of the document was learning objectives and then syllabus."*	3	3	3
PPI:2.7	Personal ownership	Content coded at this node evidences the need for academic staff to assume personal ownership in the curriculum design process and/or the need for this to be reflected in the C-CAP system.	*"One of things that, again, I can't remember what I put on the form... But I don't think it made specific reference to personal ownership. One of the things that often happens when classes are created is that there's already a member of staff - an academic member of staff - who is designated with building the class and creating the class, and I think that needs to be indicated. Because they then become the point of contact that other people can refer back to for concerns or queries. On a class descriptor form, even on a draft planned class,*	1	1	1

Super-node: Process and pedagogical issues						
Node code	Node	Node definition / scope note	Example quote(s)	Sources	References	Unique source
			you would still have identified the class coordinator, because they will be the person driving it forward."			
PPI:2.8	Principles of Assessment	Content coded here evidences participants' knowledge, experiences and views on the University's Principles of Assessment and Feedback.	*"As if anyone actually knows that the University's principles of assessment feedback actually are. It would be good to have a drop down menu so that you could randomly pick one to twelve, or is it one to four now? That's also not a requirement of the Faculty documentation, so... who knows?"* *"Principles of Assessment and Feedback. There are 12 principles of good assessment... yup. Right, and this, I think, is an area where some examples would be really quite useful. Again, I've seen a very high variation in what different lecturers put in here. Maybe, given the emphasis on feedback that the students are requesting and also in the student survey it seems to be quite important, it maybe better to have the feedback as a separate category here. So it's quite clear that the students can see exactly what the feedback is, what they can expect from the course... More guidance on that area would be useful, and perhaps the feedback as a separate issue."*	4	4	3
PPI:3	University management, policy	Content coded at this node pertains to participant feedback about University policies, procedures or management decisions that affect the curriculum design and approval process and/or teaching.	*"Class evaluation... That's interesting... I'm not quite sure what it means by self-evaluation. Who is the self - student or staff? Staff evaluation might be more appropriate. It's... Um... I wonder whether this is slightly redundant. I would hope the University is moving towards a specific... These should just be standard features of an academic activity which really don't need to be defined. They are there and they are used. All departments have staff-student committees. So all staff-student committees have the opportunity to comment on classes. All classes are required to go through an annual review process, so is it even necessary...? This is not something that features in the current process at all and I wonder whether it is even necessary. Not that class evaluation isn't necessary. Class evaluation is absolutely critically necessary, but it's there. There are University processes which are used and are known about. They don't need to be defined in the approval process."* *"I rather like... The form isn't asking me to confirm availability of a lecture room. Again, we would take that for granted. Why should a computer lab be any different?"* *"It's specific to Science, I suppose. If that wasn't running.... Erm, the devolved nature of the University allows different Faculties to do different things, which is stupid."*	4	5	4
PPI:3.1	Code allocation	Content coded at this node documents participants' understanding of the course/class code allocation process.	*"I'm still a bit worried about a request for a course code. If it really means a degree course code; most people involved in approval will have no idea that means, especially because the University currently runs duplicate systems of course coding. So, 2.1 is very confusing and unclear."*	1	2	1
PPI:3.2	Flexibility	Evidence of the need for academic flexibility in curriculum design and teaching delivery.	*"Deadline week number may vary, again, depending on how the coursework is split up. We don't know precisely how many pieces of coursework there might be. But the expectation is that there would be a minimum of two but probably a maximum of three. That's something that we might decide early on once we saw the number of people attending the course."*	3	3	3
PPI:3.3	Terminology	Content evidencing participant uncertainty, confusion or recognition of terminological problems in the class and/or course design and approval process.	*"What happens if "course" actually means "programme" name or "degree" name, and there are several? We have different terminologies in different faculties, you see. Nothing is standardised. So a class and programme and a course can be interchangeable depending on which faculty you're at."* *"It is a little bit unclear here, when I'm starting this. This is a class specification; really curriculum - or my understanding of curriculum - is the whole course rather than an individual class, so that's a little confusing, I think. And also "class"... Traditionally we'd call this a "module descriptor" form, rather than "class". A problem with definitions, I*	4	5	4

Super-node: Process and pedagogical issues						
Node code	**Node**	**Node definition / scope note**	**Example quote(s)**	**Sources**	**References**	**Unique source**
			guess."			

7. Appendix B: Coding framework: System issues (super-node)

Table 8: Coding framework for the super-node "System issues" only.

Super-node: System issues						
Node code	**Node**	**Node definition / scope note**	**Example quote(s)**	**Sources**	**References**	**Unique source**
SI:1	Assessment activity	Data coded at this node denotes a participant requirement for a greater number of assessment options.	*"Format, delivery and assessment? Okay, so you are able to insert.... Well, I mean, our descriptors have "lectures", "tutorials", "laboratories", "assignments", "self study". One of things you sometimes see in terms of activities is a distinction between private study and directed study, in that - and this is particularly important in terms of some of the accreditation activities; because private study would be time which you spent reading, revising, doing things that you wish to do in order to get you through the class. Directed study would be time spent your own in your own time doing specific tasks, such as writing up a lab report, producing an essay... So some.... Everybody recognises that within the hours of the class you don't... for a 20 credit class you don't teach 200 hours; but when you look at the bits you're not in contact with the student it is very differently divided into directed and private. It's important that one indicates that there is an element of directed study where a specific and intended task is being completed. This is particularly important in things like practical work where a very large amount of the class might be involved in directed study as opposed to private study."*	10	33	9
SI:1.1	Assessment deadline	Data coded at this node evidences a participant view that "assessment deadline" should not be associated with particular assessments, e.g. examinations, courseworks, etc.	*"I'm fairly flexible with some the deadlines, actually. I wouldn't like to be prescriptive about it because I think it would vary a little bit according to the progress you make in terms of the lectures and labs. And that depends on the cohort of students and how quickly they learn. I do adapt it a bit in practice. I don't like these being too prescriptive. So I'd rather not have to have fixed deadlines."* *"Deadline week number may vary, again, depending on how the coursework is split up. We don't know precisely how many pieces of coursework there might be. But the expectation is that there would be a minimum of two but probably a maximum of three. That's something that we might decide early on once we saw the number of people attending the course."*	6	9	6
SI:1.2	Assessment duration	Data coded at this node supports participant concerns over the validity of "assessment duration", as per section 4.1 of the C-CAP system.	*"Coursework, as an assessment... Duration may not make sense there. Some of the coursework might be done in labs, in which case the duration will be the duration of the labs. In other cases it may involve submitting an assignment. So the duration... does that mean the time between the coursework being issued and submitted. It might be several weeks. I'm not clear on how I would answer that."*	5	7	5
SI:1.3	Option values	Data coded at this value provides specific participant suggestions for additional assessment option values (for section 4.1 of C-CAP).	*"So these are very generic categories. So, "individual assignment", "group assignment", "group work", "group presentations"; all these things are all missing."* *"I was looking for a debate or presentation... It's quite narrow in terms of your descriptions of assessment. I would expect to see a break down between a case study and a project are relatively similar, in a business context perhaps. Essay, report, presentation. Other formats we may use are debate, as I say; but we also... If you have an attendance requirement, in terms of they have to come to compulsory tutorials then that needs to be in as an assessment weighting as well because it tends to have marks attached to it.* *This "coursework" is just a bit bland and a bit general for me. It doesn't give enough detail."*	8	10	7
SI:2	C-CAP perceptions	Data coded at this node evidences participants' general perceptions about the C-CAP system, e.g. its usability, its ability to support curriculum design, etc.	*"You see, this bothers me... This always bothers me about these things where you have these pre-set form and you're entering information. I mean it's easy for me to just use a form because when I'm sticking to a pre-set piece of software, y'know, I can't really see very well what I've written. And I hate that. If you can imagine, I did this under great pressure of time, um, and so that last thing I want to do is spend my time trying to figure out what it is I've just written. And then if I accidentally erase it.."*	7	14	7

Super-node: System issues						
Node code	**Node**	**Node definition / scope note**	**Example quote(s)**	**Sources**	**References**	**Unique source**
			"Learning objectives? They often are bulletted. It directly relates to the sort of information one would expect on a class descriptor. Interestingly, if this system performs well it could actually be the generator of a class descriptor. Ahhh, now I understand how this adding works. This is good." *"Generally the system is quite intuitive to use, so it's easy, it's straightforward."*			
SI:2.1	Class evaluation	Data coded at this node discusses class evaluation and related aspects in the C-CAP system.	*"There's no summative assessment in class evaluation. It's all formative. I think that's a... I don't think it's a relevant question, to be honest. What are the choices? "Self-evaluation" is hardly summative. Similarly with "Student feedback"... There is a wee bit of summative in that you give the students a list of one to five; but again, it's feedback that informs your teaching. There is no summative in there. Summative essentially has a final mark associated with it. That's my understanding of summative. There's a mark that counts towards something. Any form of feedback you can take on board or you can ignore. If you ignore it then, okay, you're making a rod for your own back."* *"These should just be standard features of an academic activity which really don't need to be defined. They are there and they are used. All departments have staff-student committees. So all staff-student committees have the opportunity to comment on classes. All classes are required to go through an annual review process, so is it even necessary...? This is not something that features in the current process at all and I wonder whether it is even necessary. Not that class evaluation isn't necessary. Class evaluation is absolutely critically necessary, but it's there. There are University processes which are used and are known about. They don't need to be defined in the approval process."*	7	7	7
SI:2.2	Course codes	Content coded at this node evidences participant concerns about identifying courses in C-CAP, e.g. need for drop down lists, potential for confusion of course codes with UCAS codes, etc.	*"Once you find the class you then enter... It automatically enters the course code because... The reason why I say that is: there are different codes depending on how you interact with the system. For example, BSc Physics is 0027/1 2 3 or 4, depending on which year it is, and that's the code that Registry use, I think, to identify a student with that. Whereas... With the UCAS application process there is a completely different set of codes associated with that. And the Admissions side of the degree has a different code from the actual Registry side of things. So, you can end up remembering too many codes. Maybe a simple drop down menu, or another box saying "This is a new course" would make more sense..."* *"Course code? Um, it's not clear what the course code refers to there at all. If it really means a degree course code, people won't understand that."*	7	7	7
SI:2.3	Dummy codes	Node denoting participant discussion of the perceived need for "dummy codes" to assist in the curriculum approval process.	*"These will often come to approval processes with dummy course codes anyway. Indeed, there is some confusion there in that sometimes you can't get a class code until your course is approved. And sometimes you can. This form here that working from as a draft is giving a dummy code, but others have already got their codes, so it's variable."*	2	2	2
SI:2.4	Form requirements	Participant comments concerning the detail or requirements of the form and the information required to be completed by participants.	*"I want to just say there are four classes that take place this week, this week, that week. You know? It's almost as if there's too much information being asked in this. Some of this information should be given to the students by the department when they are delivering the class, rather than going in... making up the approval form."*	8	13	8
SI:2.5	PoA menu	A node that evidences participants' C-CAP system perceptions or needs for section 4.5 (Principles of Assessment and Feedback).	*"Principles of Assessment and Feedback. There are 12 principles of good assessment... yup. Right, and this, I think, is an area where some examples would be really quite useful. Again, I've seen a very high variation in what different lecturers put in here. Maybe, given the emphasis on feedback that the students are requesting and also in the student survey it seems to be quite important, it maybe better to have the feedback as a separate category here. So it's quite clear that the students can see exactly what the feedback is, what they can expect from the course... More guidance on that area would be useful, and perhaps the feedback as a separate issue."*	2	2	2

Super-node: System issues						
Node code	**Node**	**Node definition / scope note**	**Example quote(s)**	**Sources**	**References**	**Unique source**
SI:2.6	Read only	Participant discussing the role of "read only" versions of the form.	*"How easy is it print this form out in its entirety? I can't work.... I can do this but I don't like. I just prefer writing on documents and it's faster to write than it is to call up a PDF sticky and type that in, or enter a comment."*	1	1	1
SI:2.7	Reading and resources	Data coded at this node evidences user confusion / issues with section 5.2 (Recommended reading and resources).	*"When the say "Availability", is it that something is available in the library? So, for example, because there are various journal, so journal they pick up from the library, some not. Can I just put "Not available"? Or, there are various journal [...] some we just provide for them. Also, an option.... "Available locally"?"* *"This 5.2 is incredibly tedious to do, to be honest. Resource? Does that mean an actual book, or does it mean books in the library? The forms that go to Academic Committee require a reading list, an indicative reading list, and that is different from the additional resources required for the class. So you'd have things like, "We need a room with flexible seating, AV - which is impossible in some cases - or you need white boards or this, that and the other". Books are separate. It's a reading list. This seems to be confusing two things together."*	9	11	9
SI:2.8	System consistency	Nodes denotes participant comments relating to C-CAP system (or lack of) consistency.	*"Now it's telling me, in red, that assessment weighting was sum to 100%. I believe they do. It would be nice if it didn't tell me that if it did. Otherwise I'm assuming there might be something wrong or incorrect. If it's going to add up figures earlier on but not add them up now, it seems inconsistent."*	1	1	1
SI:2.9	System navigation	Data coded at this node evidences participants' experiences with the C-CAP navigation.	*"It's not intuitive that you move along these top bars. That was a guess. I guessed. As you'll notice from the survey, I regard myself as reasonably IT literate. But I don't think it's intuitive that these five boxes are step boxes that you step along. Perhaps just a sentence..."*	3	4	3
SI:2.10	Unnecessary information	Code evidences examples of unnecessary information being provided in class forms, e.g. "not applicable", "none", etc.	*[Evidenced via screen capture video]*	3	3	3
SI:3	Class framework	General issues pertaining to class / module framework issues. Also acts as aggregate node for child nodes.	*"Many of the classes the Physics Department offers, and the Science Department offers, offer an exemption scheme whereby students will take a range of class tests. These will be done throughout the semester and then they will... and then if the student performs to a certain defined level, the student will be awarded the credits. Sorry, the student will not need to sit the January or June examination for that task because the Department has deemed that their performance is satisfactory such that the exam board can award the credits for the class. How can that be reflected under here? I know there's a notes field but, the way I look at it, the notes field relates to the examination and such like."*	10	25	10
SI:3.1	Academic level	Data coded at this node evidences participant issues with the assignation of UG or PG and a preference for "academic level".	*"Okay, level you should specify the academic level, not whether it's undergraduate or postgraduate. It should be level one, two, three, four, five - and then you can determine whether it's postgraduate from, erm, the level descriptor."* *"Credit value... Level... It's either undergraduate or postgraduate. Level, in my terminology, is 1, 2, 3, 4, or 5. MEng or MSc is level 5, for taught modules."*	5	5	5
SI:3.2	Credit values	Content at this node evidences participant issues with the credit values used in section 1.1 of C-CAP.	*"I don't know if I've ever seen it written down, exactly how many hours there should be for 10 credits; but I've heard informally that it should be about 100 hours. And I assume that that includes students doing their assessments... assessment activity. I may be wrong, but that's what I've heard."*	3	3	3
SI:3.2.1	Credit-to-hours mapping	General evidence of system need to assist participants in calculating the number of activity hours associated with the credit system.	*"One thing that we're advised is that for a 10 credit class there should be a total of 100 hours, so it would be useful to get some advice here, I guess, on the screen to make up their total to 100 hours; or, at least, to have some explanation why it's not 100 hours. So in this case I will insert an activity which is "private study", and make that 50 hours - and that gives me 100 hours, which is typical for a 10 credit class, I think."*	6	7	6

Super-node: System issues						
Node code	**Node**	**Node definition / scope note**	**Example quote(s)**	**Sources**	**References**	**Unique source**
			"Perhaps if there's a standardised model in terms of the number credits that you put in? Perhaps there should be a total hours of activity that you've got to get to?"			
SI:3.3	Mode of attendance	Data coded here evidence participant uncertainty relating to definitions of attendance modes, e.g. open, distance, etc.	*"The 'modes of attendance' is an interesting question. Many academics designing classes won't really necessarily be familiar with the distinction between "attending" or "open" class structures. So... I wonder whether that's something that's really relevant at the early stage of class approval."*	4	4	4
SI:3.4	NQ	Content that discusses the issues involved in "NQing" (Not Qualified to sit examination).	*"In that context I think, one of things class descriptors will often talk about - and it's an issue the University needs to consider more - there is a process called "NQ"; you deem a student "non qualified" to sit an assessment on the basis of some activity. Some failure to attend, some failure in another aspect of the course. And if one has an NQ procedure with their class it needs to be indicated; routes out of NQ procedure also need to be indicated. That's a tricky one because, to be quite honest, I don't like the whole concept of NQing anyway, so I'd rather not see it there at all. But I know it is quite heavily used by some classes and some departments."*	2	2	2
SI:3.5	Semester system	Content coded at this node captures participants' views on recording the teaching pattern of classes.	*"One other thing.... It doesn't apply to this particular class which I'm entering now, but some other classes that I have been involved with, is that the MSc - Power Plant Engineering - is taught throughout the year, so it's not tied to the semester system. So having semester one, semester two, wouldn't be applicable for some of the modules which we have on that course."*	2	2	2
SI:3.6	Taught hours	Node content pertains to participants' discussion of how hours for particular types of activity are allocated.	*"Format, delivery and assessment? Okay, so you are able to insert.... Well, I mean, our descriptors have "lectures", "tutorials", "laboratories", "assignments", "self-study". One of things you sometimes see in terms of activities is a distinction between private study and directed study, in that - and this is particularly important in terms of some of the accreditation activities; because private study would be time which you spent reading, revising, doing things that you wish to do in order to get you through the class. Directed study would be time spent your own in your own time doing specific tasks, such as writing up a lab report, producing an essay... So some.... Everybody recognises that within the hours of the class you don't... for a 20 credit class you don't teach 200 hours; but when you look at the bits you're not in contact with the student it is very differently divided into directed and private. It's important that one indicates that there is an element of directed study where a specific and intended task is being completed. This is particularly important in things like practical work where a very large amount of the class might be involved in directed study as opposed to private study."*	7	8	6
SI:4	Learning activity	Content at this node captures general issues pertaining to learning activities and their documentation in curriculum design approval forms.	*"For instance, as part of the class delivery hours we've got 76 hours allocated for assignments. Now that might partly be done in the labs but it may be submission of some kind of report."*	3	3	3
SI:4.1	Learning activity number	Data coded at this node evidences participants' views on the C-CAP requirement to specifiy the number of learning activities required in the class.	*"The number or duration... I don't think this detailed information is necessary. All that is mostly necessary is the number of hours within the class. So typically... A class like this might have, it's a 20 credit class - it's going to have round about a third of that; it might have 60 hours of practical. Again, my experience of most class descriptor processes; they don't bother to drill down to the number of sessions. So, okay, I'm going to say, for example, we might expect there to be round about 15 four hour sessions. Private study would be the rest."*	1	2	1
SI:4.2	Learning activity options	Data coded at this node supports the need for extra options in the drop down menu for "Type of activity" (section 4.1 in C-CAP).	*"Well, I mean, our descriptors have "lectures", "tutorials", "laboratories", "assignments", "self-study". One of things you sometimes see in terms of activities is a distinction between private study and directed study, in that - and this is particularly important in terms of some of the accreditation activities."* *"Now, we had "assignments" as a button on our list here, which is "Field work", "Lecture", "Placement", "Practical"... We had that separate from "Private study". That's just an*	8	10	8

Super-node: System issues						
Node code	**Node**	**Node definition / scope note**	**Example quote(s)**	**Sources**	**References**	**Unique source**
			observation. But we could just combine it. They do a lot of homework and that's how they get their feedback and so on. So we like to say, "Yes - you will be expected to spend time on this", rather than just this nebulous "private study" that, sometimes, I think they completely ignore that. Whereas if it says "You're expected to spend a certain amount of time on the assignments", it focusses them a bit more."			
SI:5	Technical impediment	Data at this - and sub-codes - pertain to specific technical issues or errors preventing meaningful use of the C-CAP system.	*[Facet node]*	0	0	0
SI:5.1	Delete button problems	User difficulties with the C-CAP delete button.	*[Evidenced via screen capture video]*	1	1	1
SI:5.2	Form submission errors	Content coded at this node evidences C-CAP form submission or form saving errors.	*[Evidenced via screen capture video]*	2	6	2
SI:5.3	Inputting class codes	Data coded here evidences participant concerns about entering or remembering class codes, e.g. re-ordering of form fields, requirement for look-up, etc.	*"I can't remember the correct class code. Yeah, yeah. This is just me; but I always think of the class code, not the class name. So the first thing I enter is the class code and not the class name. I always find it really disconcerting when you search the class catalogue and the first field is the class name rather than the class name, because it is more efficient to enter the class code than the class name. But, yeah, that's just me."*	1	1	1
SI:5.4	Insert button problems	Data coded at this node documents participant usability issues with the "insert item" buttons in C-CAP, e.g. insert button not visible to participant, insert button unresponsive, etc.	*"My impression is that I need to click on "Add a learning objective" twice each time, in order to get it to respond. I think that's happened... I'll double check next time. Yeah - that's confirmed."*	4	4	2
SI:5.5	Obscuration of text	Content coded at this node evidences instances in which C-CAP obscures inputted content thereby limiting usability, e.g. failure for text box to expand, important text above or below page fold, etc.	*"The later coursework is designed to assess all the learning objectives, so it was relatively easy; but it would be tedious if you were focussing on just one or two of these things to remember which learning objective is it, and having scroll back up, and then..."* *"Not being able to see learning objectives when using the drop down lists."*	4	4	3

8. Appendix C: Node tree map

A tree map is a representation of coded data, displaying items as nested rectangular boxes. These boxes diagram hierarchical data as nested boxes of varying sizes. The size of the box represents how many of source items are coded by the nodes displayed. The colour of each box also represents the number of coding references.

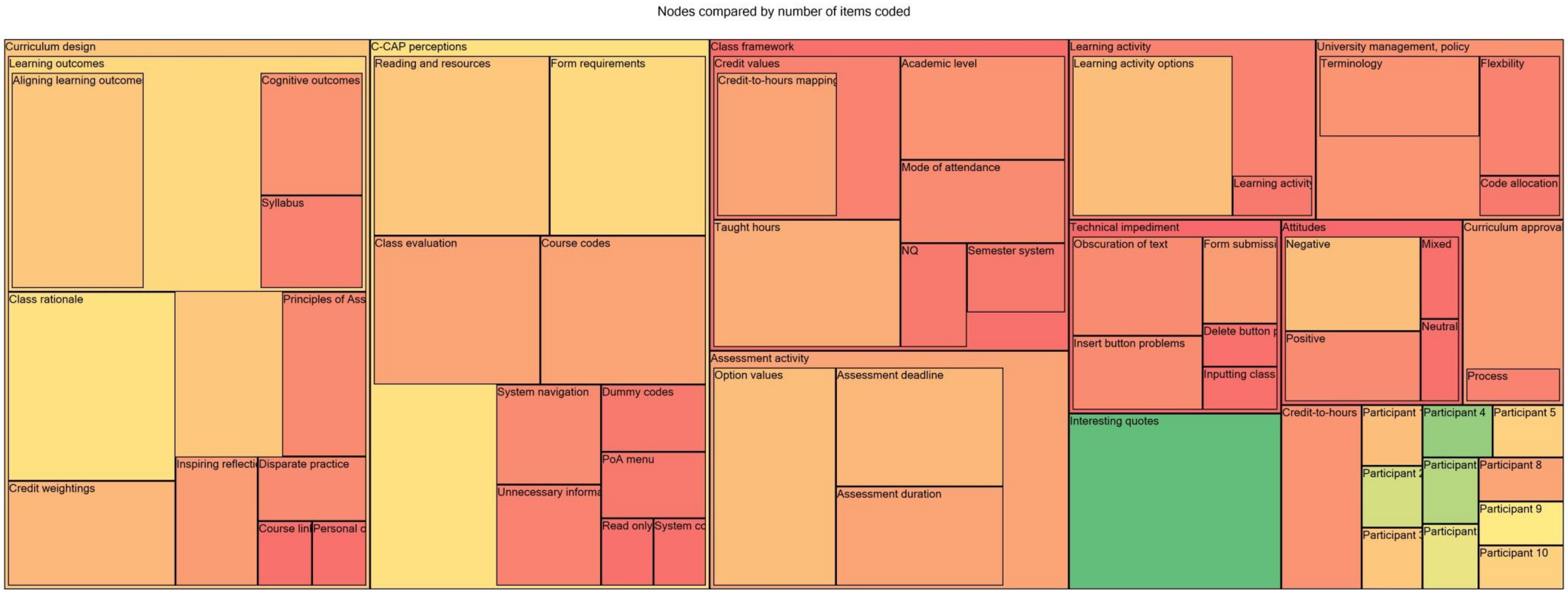
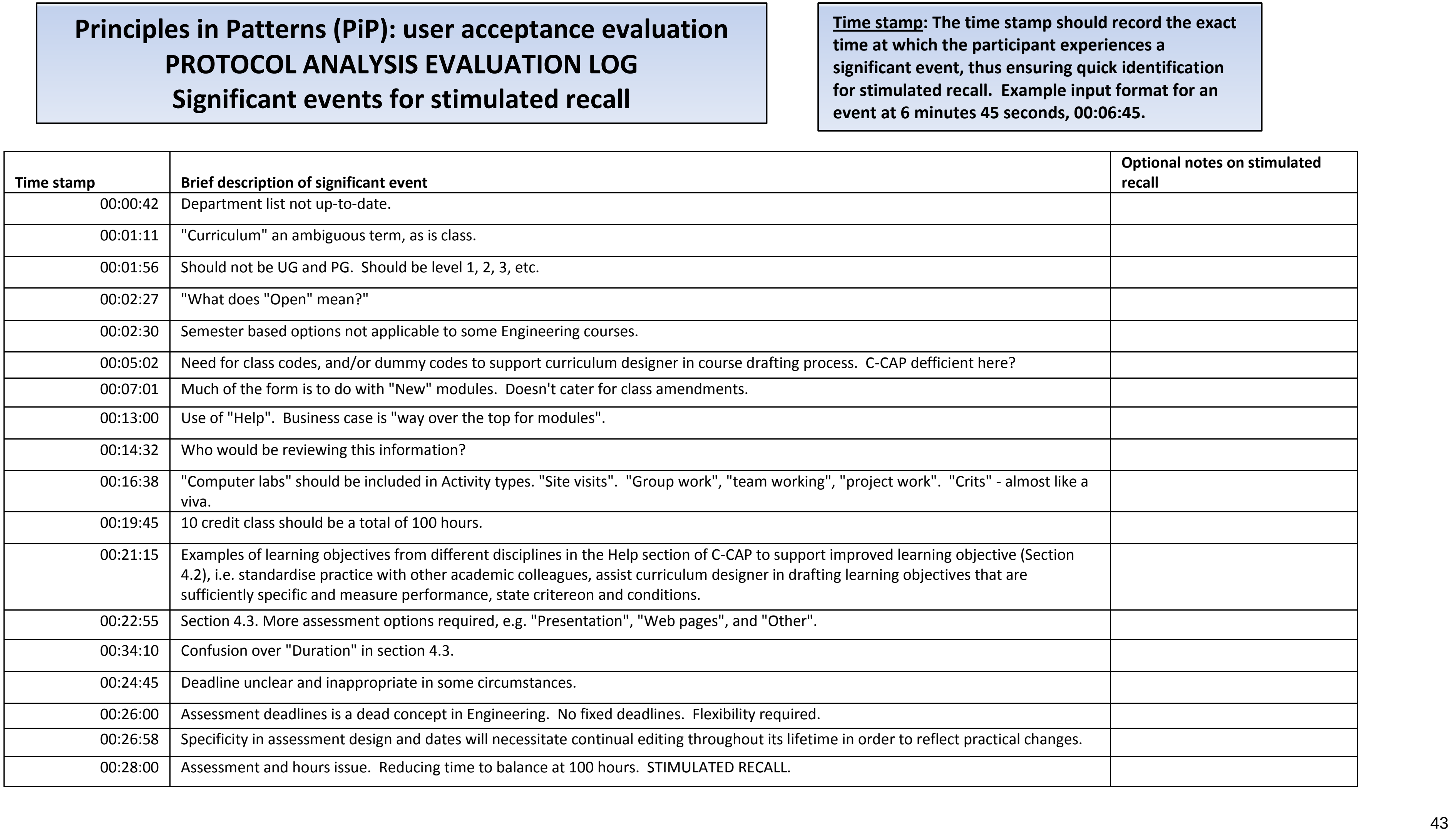


Figure 8: Node tree map representing nodes from the coding framework.

9. Appendix D: Evaluator log example



Principles in Patterns (PiP): user acceptance evaluation
PROTOCOL ANALYSIS EVALUATION LOG
Significant events for stimulated recall

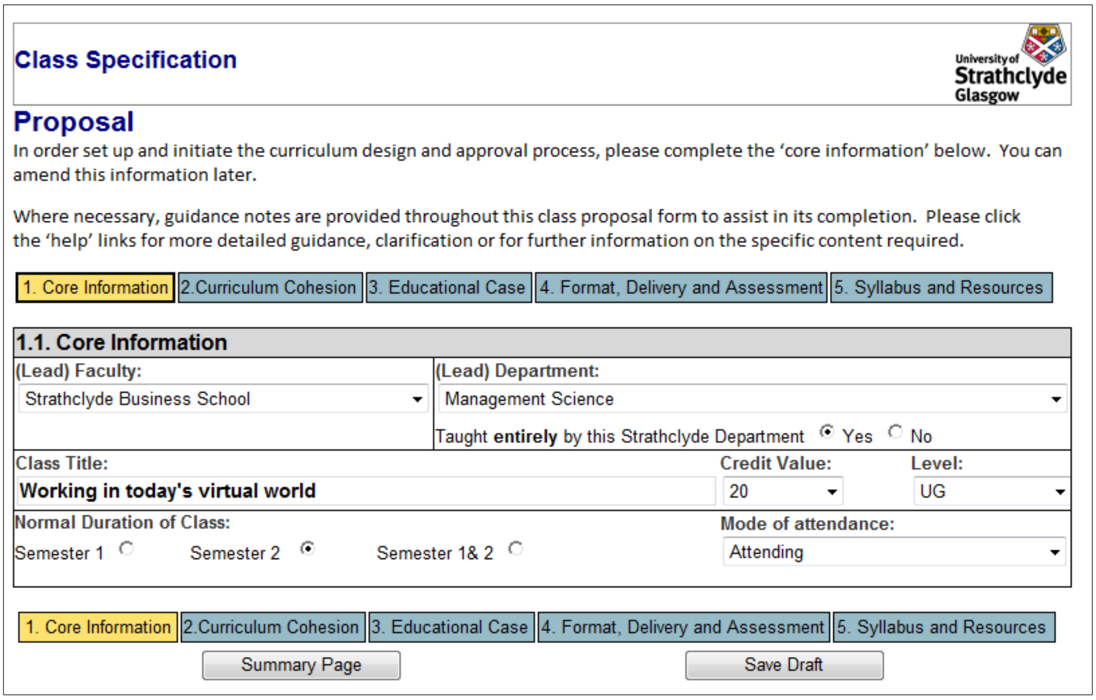
Time stamp: The time stamp should record the exact time at which the participant experiences a significant event, thus ensuring quick identification for stimulated recall. Example input format for an event at 6 minutes 45 seconds, 00:06:45.

Time stamp	Brief description of significant event	Optional notes on stimulated recall
00:00:42	Department list not up-to-date.	
00:01:11	"Curriculum" an ambiguous term, as is class.	
00:01:56	Should not be UG and PG. Should be level 1, 2, 3, etc.	
00:02:27	"What does "Open" mean?"	
00:02:30	Semester based options not applicable to some Engineering courses.	
00:05:02	Need for class codes, and/or dummy codes to support curriculum designer in course drafting process. C-CAP defficient here?	
00:07:01	Much of the form is to do with "New" modules. Doesn't cater for class amendments.	
00:13:00	Use of "Help". Business case is "way over the top for modules".	
00:14:32	Who would be reviewing this information?	
00:16:38	"Computer labs" should be included in Activity types. "Site visits". "Group work", "team working", "project work". "Crits" - almost like a viva.	
00:19:45	10 credit class should be a total of 100 hours.	
00:21:15	Examples of learning objectives from different disciplines in the Help section of C-CAP to support improved learning objective (Section 4.2), i.e. standardise practice with other academic colleagues, assist curriculum designer in drafting learning objectives that are sufficiently specific and measure performance, state critereon and conditions.	
00:22:55	Section 4.3. More assessment options required, e.g. "Presentation", "Web pages", and "Other".	
00:34:10	Confusion over "Duration" in section 4.3.	
00:24:45	Deadline unclear and inappropriate in some circumstances.	
00:26:00	Assessment deadlines is a dead concept in Engineering. No fixed deadlines. Flexibility required.	
00:26:58	Specificity in assessment design and dates will necessitate continual editing throughout its lifetime in order to reflect practical changes.	
00:28:00	Assessment and hours issue. Reducing time to balance at 100 hours. STIMULATED RECALL.	

43

00:28:18	Failure of C-CAP to support constructive alignment of assessment with learning objectives.	
00:29:00	Section 4.5. requires feedback examples.	
00:30:23	Section 4.6 is confusing. Terminology of formative and summative confusing and unclear.	
00:33:10	How brief should section 5.1 be?	
00:34:35	Student expected to purchase the recommended reading. Inclusion of MyPlace demonstrates that this section is far too ambiguous in its current form.	

10. Appendix E: C-CAP system interface (evaluation system)



Class Specification

University of Strathclyde Glasgow

Proposal

In order set up and initiate the curriculum design and approval process, please complete the 'core information' below. You can amend this information later.

Where necessary, guidance notes are provided throughout this class proposal form to assist in its completion. Please click the 'help' links for more detailed guidance, clarification or for further information on the specific content required.

1.1. Core Information

(Lead) Faculty: Strathclyde Business School

(Lead) Department: Management Science

Taught **entirely** by this Strathclyde Department ● Yes ○ No

Class Title: **Working in today's virtual world**

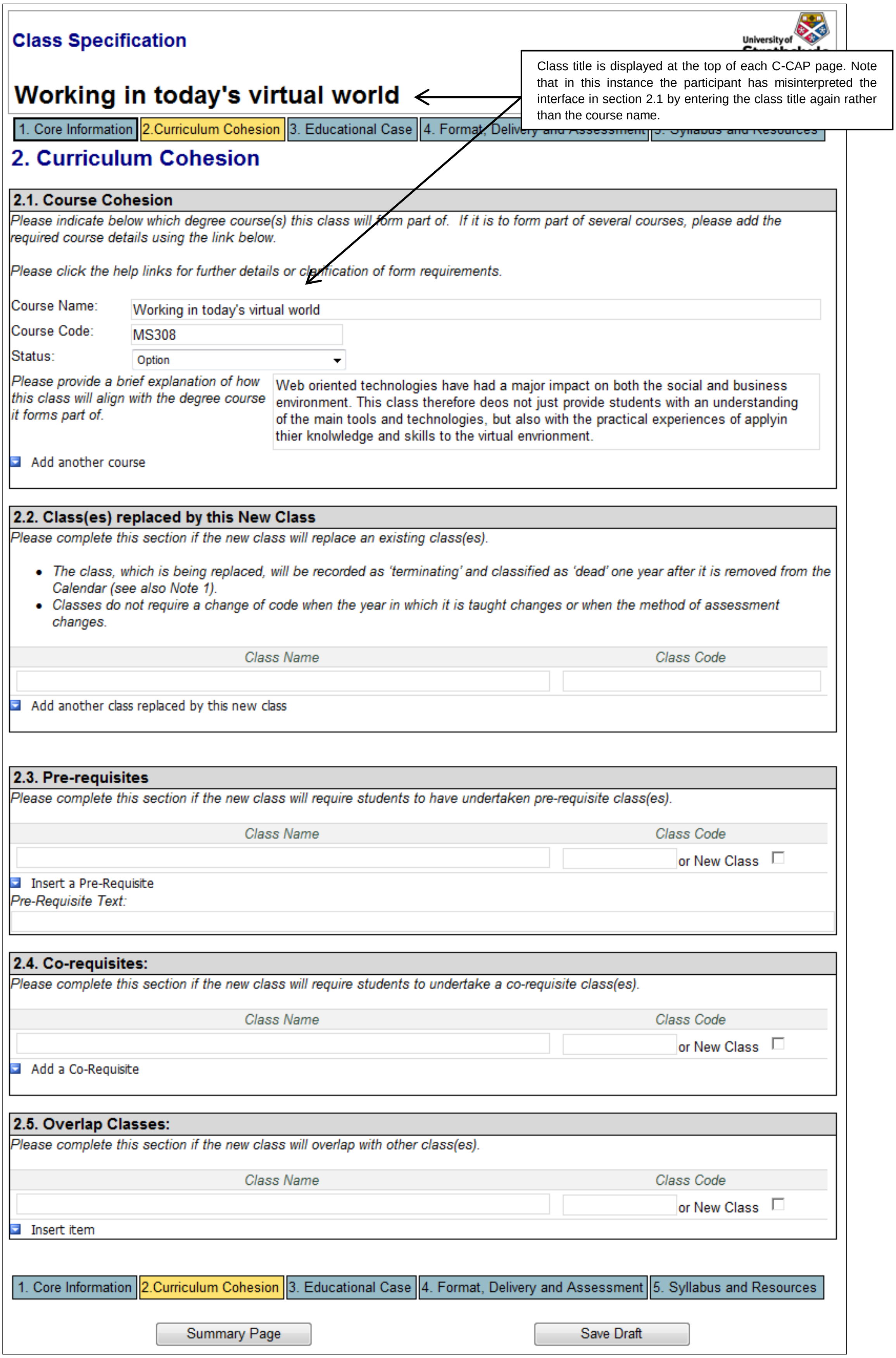
Credit Value: 20

Level: UG

Normal Duration of Class: Semester 1 ○ Semester 2 ● Semester 1& 2 ○

Mode of attendance: Attending

Figure 9: Section 1.1 of C-CAP (Core Information).



Class title is displayed at the top of each C-CAP page. Note that in this instance the participant has misinterpreted the interface in section 2.1 by entering the class title again rather than the course name.

Class Specification

Working in today's virtual world

2. Curriculum Cohesion

2.1. Course Cohesion

Please indicate below which degree course(s) this class will form part of. If it is to form part of several courses, please add the required course details using the link below.

Please click the help links for further details or clarification of form requirements.

Course Name: Working in today's virtual world
Course Code: MS308
Status: Option

Please provide a brief explanation of how this class will align with the degree course it forms part of. Web oriented technologies have had a major impact on both the social and business environment. This class therefore deos not just provide students with an understanding of the main tools and technologies, but also with the practical experiences of applyin thier knolwledge and skills to the virtual envrionment.

Add another course

2.2. Class(es) replaced by this New Class

Please complete this section if the new class will replace an existing class(es).

- *The class, which is being replaced, will be recorded as 'terminating' and classified as 'dead' one year after it is removed from the Calendar (see also Note 1).*
- *Classes do not require a change of code when the year in which it is taught changes or when the method of assessment changes.*

Class Name	*Class Code*

Add another class replaced by this new class

2.3. Pre-requisites

Please complete this section if the new class will require students to have undertaken pre-requisite class(es).

Class Name	*Class Code*
	or New Class □

Insert a Pre-Requisite

Pre-Requisite Text:

2.4. Co-requisites:

Please complete this section if the new class will require students to undertake a co-requisite class(es).

Class Name	*Class Code*
	or New Class □

Add a Co-Requisite

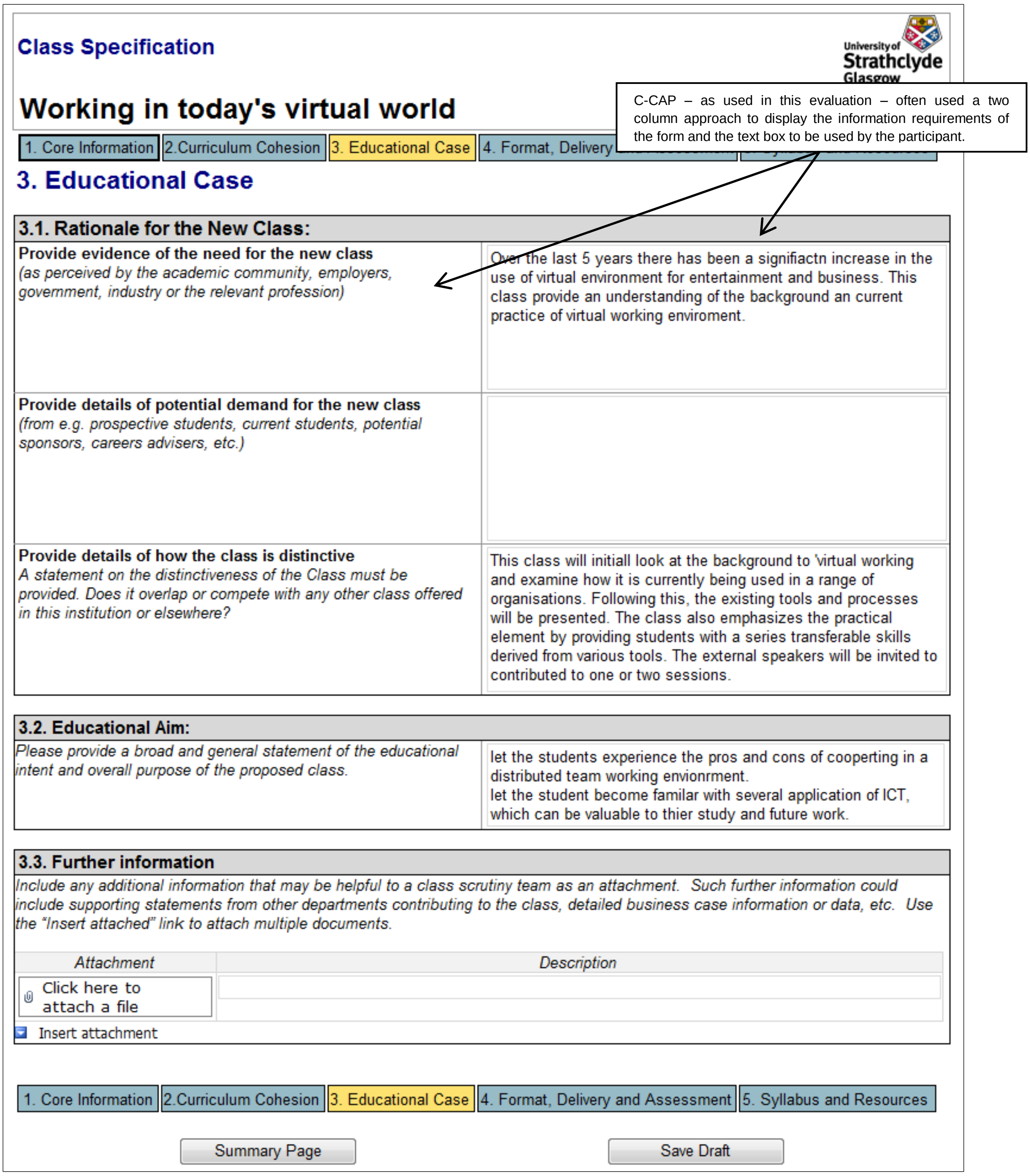
2.5. Overlap Classes:

Please complete this section if the new class will overlap with other class(es).

Class Name	*Class Code*
	or New Class □

Insert item

Figure 10: Section 2 of the C-CAP system (Curriculum cohesion).


University of Strathclyde Glasgow
1. Core Information
2.Curriculum Cohesion
3. Educational Case
4. Format, Delivery and Assessment
5. Syllabus and Resources
Summary Page
Save Draft

C-CAP – as used in this evaluation – often used a two column approach to display the information requirements of the form and the text box to be used by the participant.

Class Specification

Working in today's virtual world

3. Educational Case

3.1. Rationale for the New Class:

Provide evidence of the need for the new class *(as perceived by the academic community, employers, government, industry or the relevant profession)*	Over the last 5 years there has been a signifiactn increase in the use of virtual environment for entertainment and business. This class provide an understanding of the background an current practice of virtual working enviroment.
Provide details of potential demand for the new class *(from e.g. prospective students, current students, potential sponsors, careers advisers, etc.)*	
Provide details of how the class is distinctive *A statement on the distinctiveness of the Class must be provided. Does it overlap or compete with any other class offered in this institution or elsewhere?*	This class will initiall look at the background to 'virtual working and examine how it is currently being used in a range of organisations. Following this, the existing tools and processes will be presented. The class also emphasizes the practical element by providing students with a series transferable skills derived from various tools. The external speakers will be invited to contributed to one or two sessions.

3.2. Educational Aim:

Please provide a broad and general statement of the educational intent and overall purpose of the proposed class.	let the students experience the pros and cons of cooperting in a distributed team working envionrment. let the student become familar with several application of ICT, which can be valuable to thier study and future work.

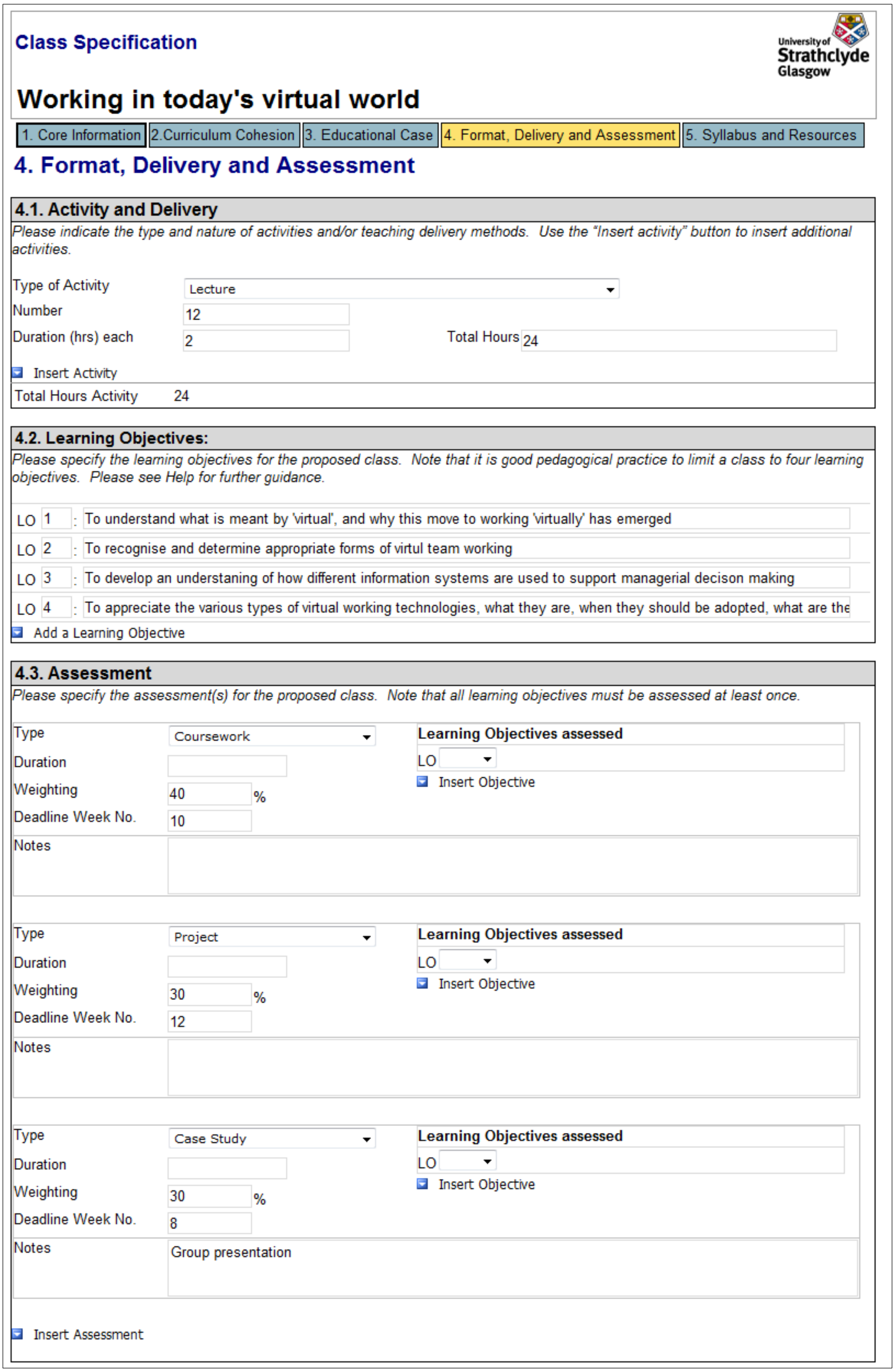
3.3. Further information

Include any additional information that may be helpful to a class scrutiny team as an attachment. Such further information could include supporting statements from other departments contributing to the class, detailed business case information or data, etc. Use the "Insert attached" link to attach multiple documents.

Attachment	Description
Click here to attach a file	

Insert attachment

Figure 11: Section 3 of the C-CAP system (Education case).



University of Strathclyde Glasgow

Class Specification

Working in today's virtual world

1. Core Information
2.Curriculum Cohesion
3. Educational Case
4. Format, Delivery and Assessment
5. Syllabus and Resources

4. Format, Delivery and Assessment

4.1. Activity and Delivery

Please indicate the type and nature of activities and/or teaching delivery methods. Use the "Insert activity" button to insert additional activities.

Type of Activity: Lecture
Number: 12
Duration (hrs) each: 2
Total Hours 24

Insert Activity

Total Hours Activity 24

4.2. Learning Objectives:

Please specify the learning objectives for the proposed class. Note that it is good pedagogical practice to limit a class to four learning objectives. Please see Help for further guidance.

LO 1 : To understand what is meant by 'virtual', and why this move to working 'virtually' has emerged
LO 2 : To recognise and determine appropriate forms of virtul team working
LO 3 : To develop an understaning of how different information systems are used to support managerial decison making
LO 4 : To appreciate the various types of virtual working technologies, what they are, when they should be adopted, what are the

Add a Learning Objective

4.3. Assessment

Please specify the assessment(s) for the proposed class. Note that all learning objectives must be assessed at least once.

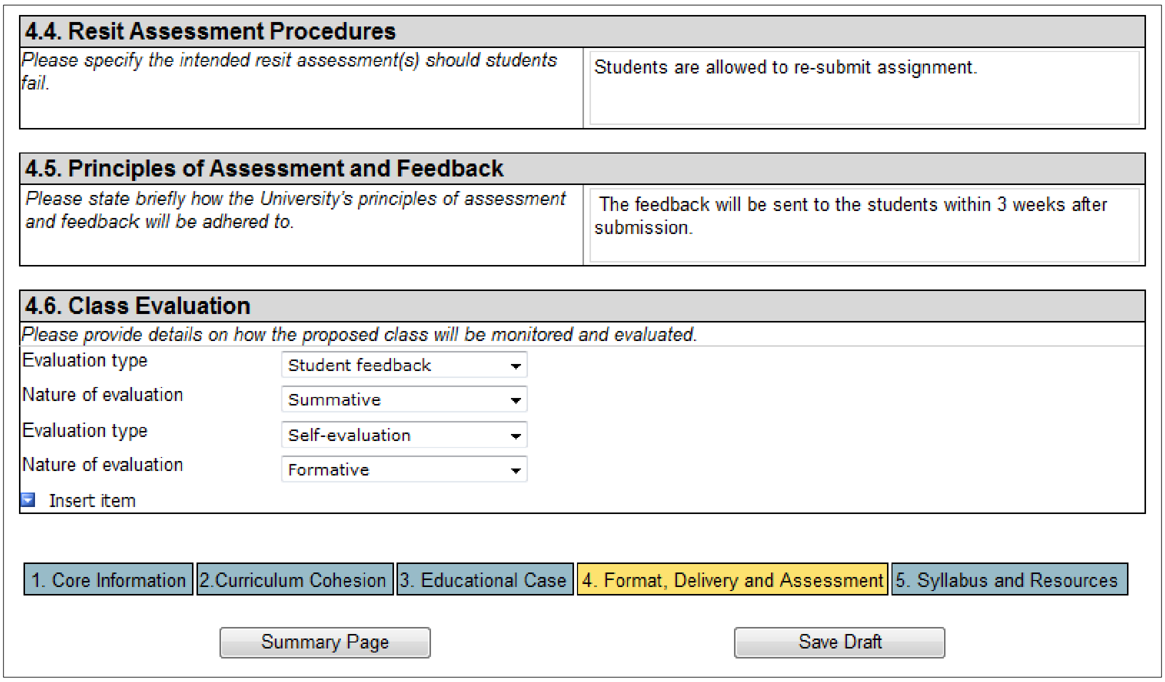
Type: Coursework
Duration:
Weighting: 40 %
Deadline Week No. 10
Notes:
Learning Objectives assessed
LO
Insert Objective

Type: Project
Duration:
Weighting: 30 %
Deadline Week No. 12
Notes:
Learning Objectives assessed
LO
Insert Objective

Type: Case Study
Duration:
Weighting: 30 %
Deadline Week No. 8
Notes: Group presentation
Learning Objectives assessed
LO
Insert Objective

Insert Assessment

Figure 12: Section 4 of the C-CAP system (Format, delivery and assessment). 1 of 2 screen shots.



4.4. Resit Assessment Procedures

Please specify the intended resit assessment(s) should students fail.	Students are allowed to re-submit assignment.

4.5. Principles of Assessment and Feedback

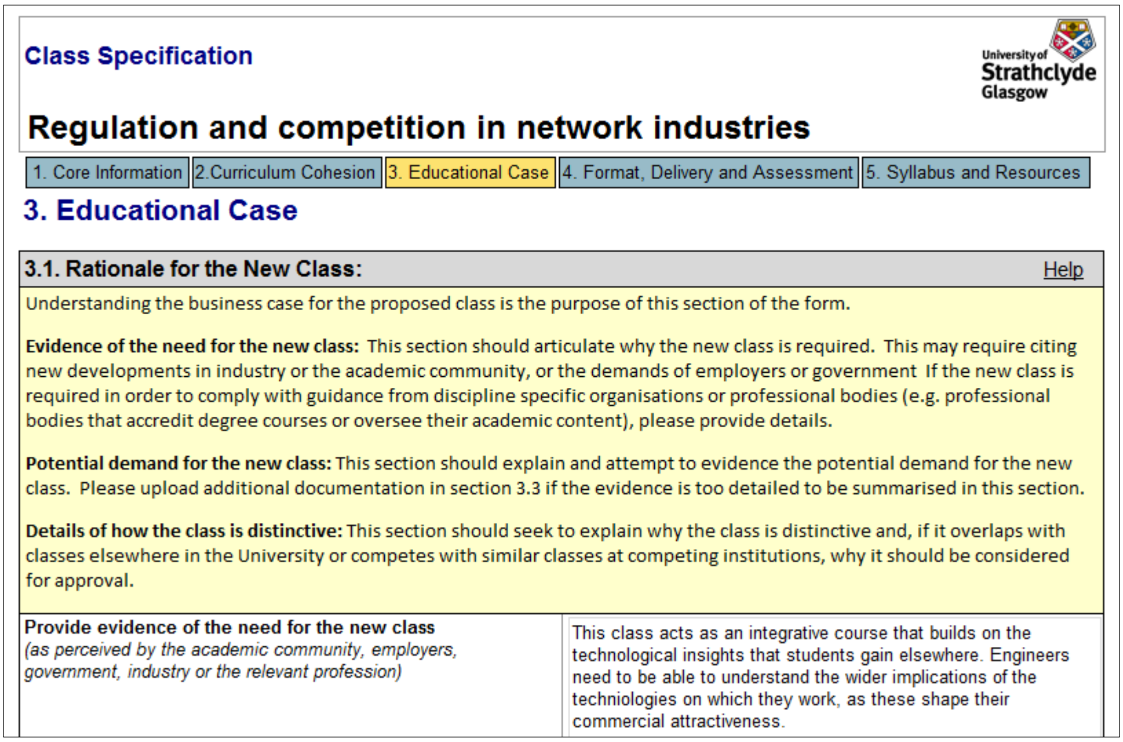
Please state briefly how the University's principles of assessment and feedback will be adhered to.	The feedback will be sent to the students within 3 weeks after submission.

4.6. Class Evaluation

Please provide details on how the proposed class will be monitored and evaluated.

Evaluation type	Student feedback
Nature of evaluation	Summative
Evaluation type	Self-evaluation
Nature of evaluation	Formative

Figure 13: Section 4 of the C-CAP system (continued). 2 of 2 screen shots.



Class Specification

Regulation and competition in network industries

University of Strathclyde Glasgow
1. Core Information
2.Curriculum Cohesion
3. Educational Case
4. Format, Delivery and Assessment
5. Syllabus and Resources

3. Educational Case

3.1. Rationale for the New Class:

Understanding the business case for the proposed class is the purpose of this section of the form.

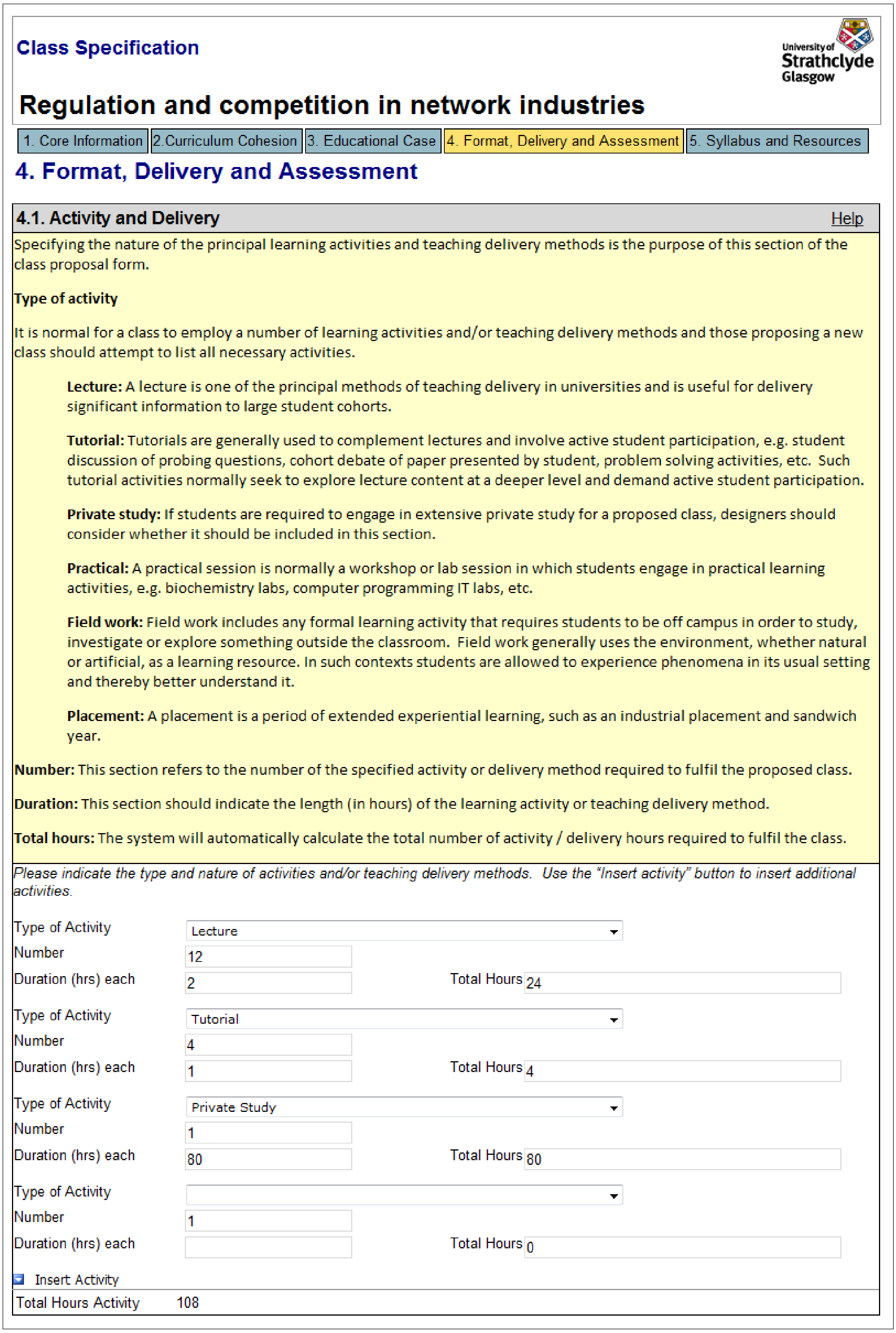
Evidence of the need for the new class: This section should articulate why the new class is required. This may require citing new developments in industry or the academic community, or the demands of employers or government If the new class is required in order to comply with guidance from discipline specific organisations or professional bodies (e.g. professional bodies that accredit degree courses or oversee their academic content), please provide details.

Potential demand for the new class: This section should explain and attempt to evidence the potential demand for the new class. Please upload additional documentation in section 3.3 if the evidence is too detailed to be summarised in this section.

Details of how the class is distinctive: This section should seek to explain why the class is distinctive and, if it overlaps with classes elsewhere in the University or competes with similar classes at competing institutions, why it should be considered for approval.

Provide evidence of the need for the new class *(as perceived by the academic community, employers, government, industry or the relevant profession)*	This class acts as an integrative course that builds on the technological insights that students gain elsewhere. Engineers need to be able to understand the wider implications of the technioliogies on which they work, as these shape their commercial attractiveness.

Figure 14: Example of the expandable / collapsible help screens available within C-CAP.


University of Strathclyde Glasgow
1. Core Information
2.Curriculum Cohesion
3. Educational Case
4. Format, Delivery and Assessment
5. Syllabus and Resources
Help
Insert Activity

Class Specification

Regulation and competition in network industries

4. Format, Delivery and Assessment

4.1. Activity and Delivery

Specifying the nature of the principal learning activities and teaching delivery methods is the purpose of this section of the class proposal form.

Type of activity

It is normal for a class to employ a number of learning activities and/or teaching delivery methods and those proposing a new class should attempt to list all necessary activities.

Lecture: A lecture is one of the principal methods of teaching delivery in universities and is useful for delivery significant information to large student cohorts.

Tutorial: Tutorials are generally used to complement lectures and involve active student participation, e.g. student discussion of probing questions, cohort debate of paper presented by student, problem solving activities, etc. Such tutorial activities normally seek to explore lecture content at a deeper level and demand active student participation.

Private study: If students are required to engage in extensive private study for a proposed class, designers should consider whether it should be included in this section.

Practical: A practical session is normally a workshop or lab session in which students engage in practical learning activities, e.g. biochemistry labs, computer programming IT labs, etc.

Field work: Field work includes any formal learning activity that requires students to be off campus in order to study, investigate or explore something outside the classroom. Field work generally uses the environment, whether natural or artificial, as a learning resource. In such contexts students are allowed to experience phenomena in its usual setting and thereby better understand it.

Placement: A placement is a period of extended experiential learning, such as an industrial placement and sandwich year.

Number: This section refers to the number of the specified activity or delivery method required to fulfil the proposed class.

Duration: This section should indicate the length (in hours) of the learning activity or teaching delivery method.

Total hours: The system will automatically calculate the total number of activity / delivery hours required to fulfil the class.

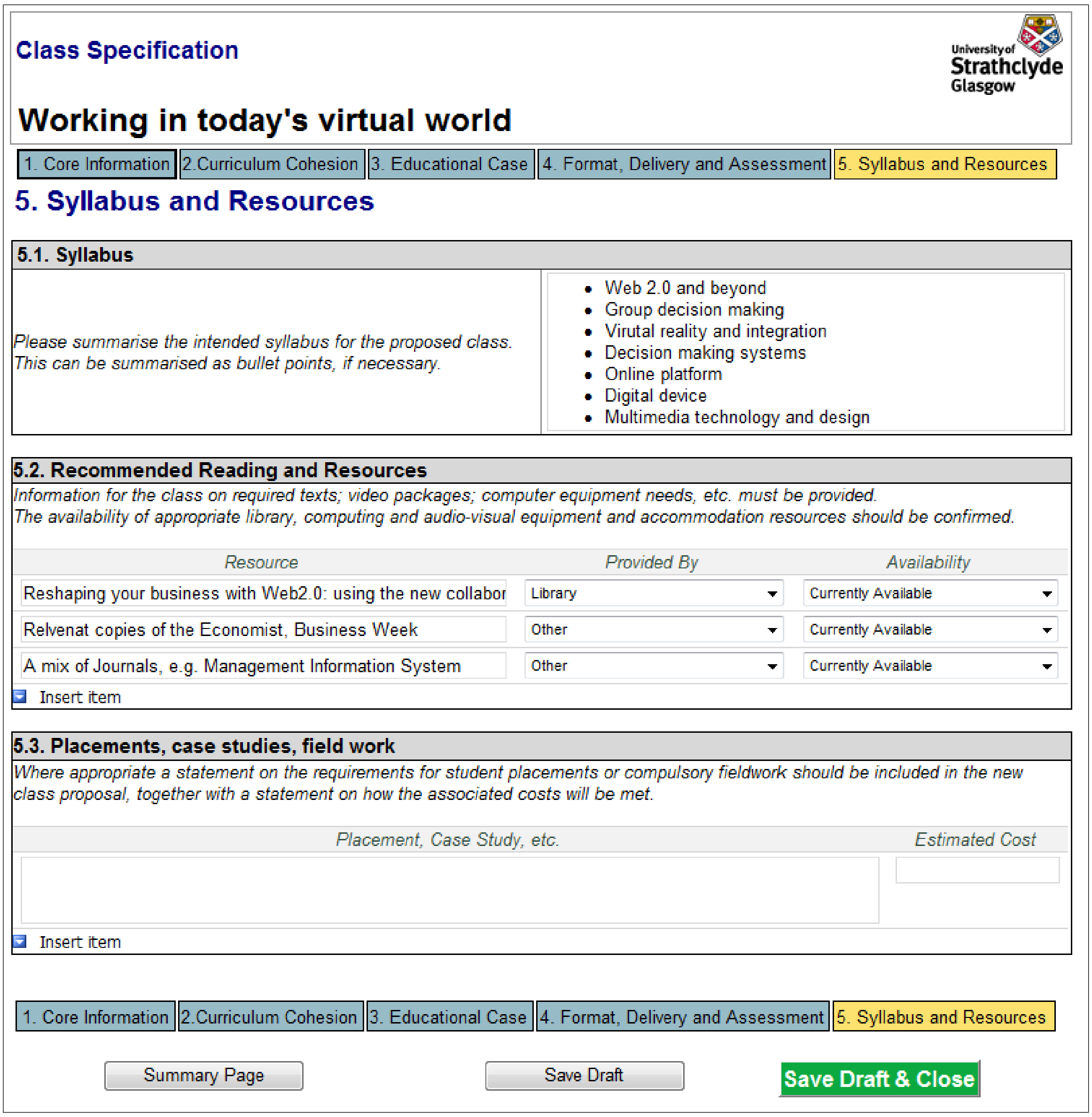
Please indicate the type and nature of activities and/or teaching delivery methods. Use the "Insert activity" button to insert additional activities.

Type of Activity	Lecture		
Number	12		
Duration (hrs) each	2	Total Hours	24
Type of Activity	Tutorial		
Number	4		
Duration (hrs) each	1	Total Hours	4
Type of Activity	Private Study		
Number	1		
Duration (hrs) each	80	Total Hours	80
Type of Activity			
Number	1		
Duration (hrs) each		Total Hours	0

Insert Activity

Total Hours Activity 108

Figure 15: Example of help / guidance detail available in expandable / collapsible help sections in C-CAP.



Class Specification

Working in today's virtual world

University of Strathclyde Glasgow
1. Core Information
2.Curriculum Cohesion
3. Educational Case
4. Format, Delivery and Assessment
5. Syllabus and Resources

5. Syllabus and Resources

5.1. Syllabus

Please summarise the intended syllabus for the proposed class. This can be summarised as bullet points, if necessary.

- Web 2.0 and beyond
- Group decision making
- Virutal reality and integration
- Decision making systems
- Online platform
- Digital device
- Multimedia technology and design

5.2. Recommended Reading and Resources

Information for the class on required texts; video packages; computer equipment needs, etc. must be provided.
The availability of appropriate library, computing and audio-visual equipment and accommodation resources should be confirmed.

Resource	Provided By	Availability
Reshaping your business with Web2.0: using the new collabor	Library	Currently Available
Relvenat copies of the Economist, Business Week	Other	Currently Available
A mix of Journals, e.g. Management Information System	Other	Currently Available

Insert item

5.3. Placements, case studies, field work

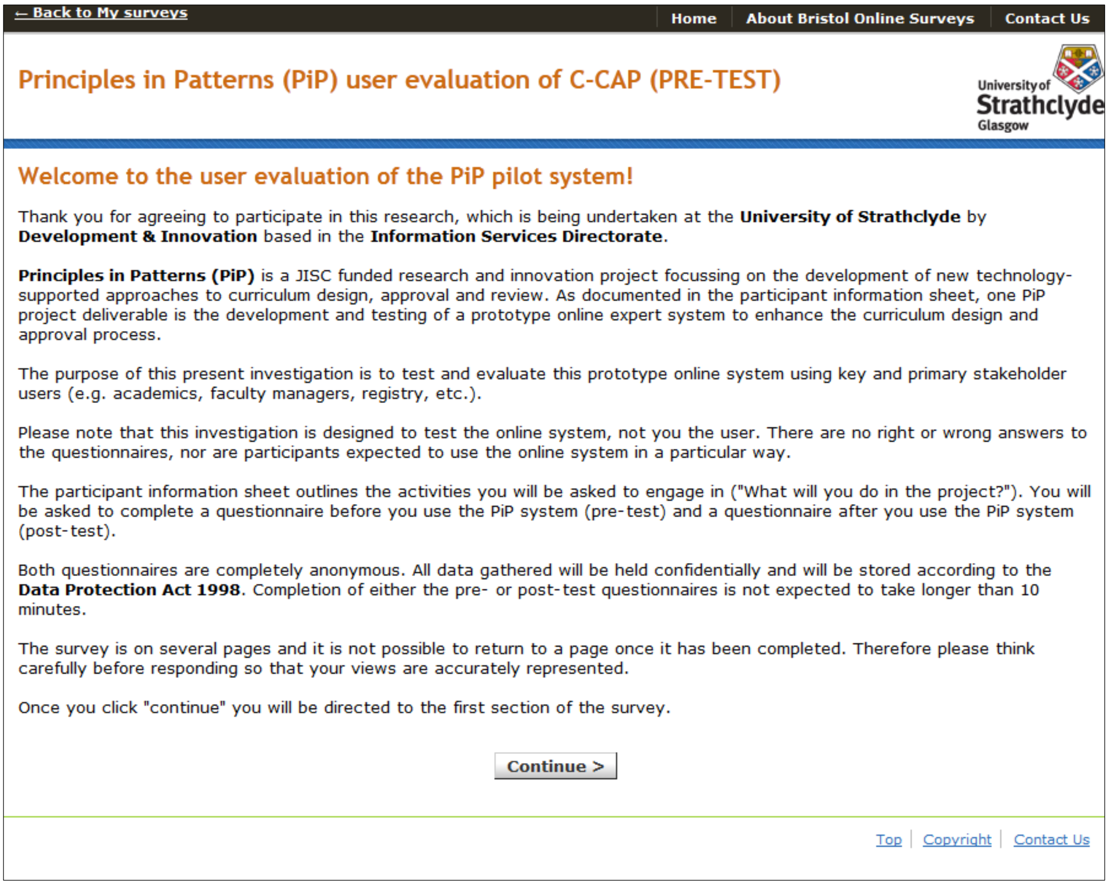
Where appropriate a statement on the requirements for student placements or compulsory fieldwork should be included in the new class proposal, together with a statement on how the associated costs will be met.

Placement, Case Study, etc.	Estimated Cost

Insert item

Figure 16: Section 5 of the C-CAP system (Syllabus and resources).

11. Appendix F: Pre-session questionnaire instrument in BOS


← Back to My surveys
Home
About Bristol Online Surveys
Contact Us
University of Strathclyde Glasgow

Principles in Patterns (PiP) user evaluation of C-CAP (PRE-TEST)

Welcome to the user evaluation of the PiP pilot system!

Thank you for agreeing to participate in this research, which is being undertaken at the **University of Strathclyde** by **Development & Innovation** based in the **Information Services Directorate**.

Principles in Patterns (PiP) is a JISC funded research and innovation project focussing on the development of new technology-supported approaches to curriculum design, approval and review. As documented in the participant information sheet, one PiP project deliverable is the development and testing of a prototype online expert system to enhance the curriculum design and approval process.

The purpose of this present investigation is to test and evaluate this prototype online system using key and primary stakeholder users (e.g. academics, faculty managers, registry, etc.).

Please note that this investigation is designed to test the online system, not you the user. There are no right or wrong answers to the questionnaires, nor are participants expected to use the online system in a particular way.

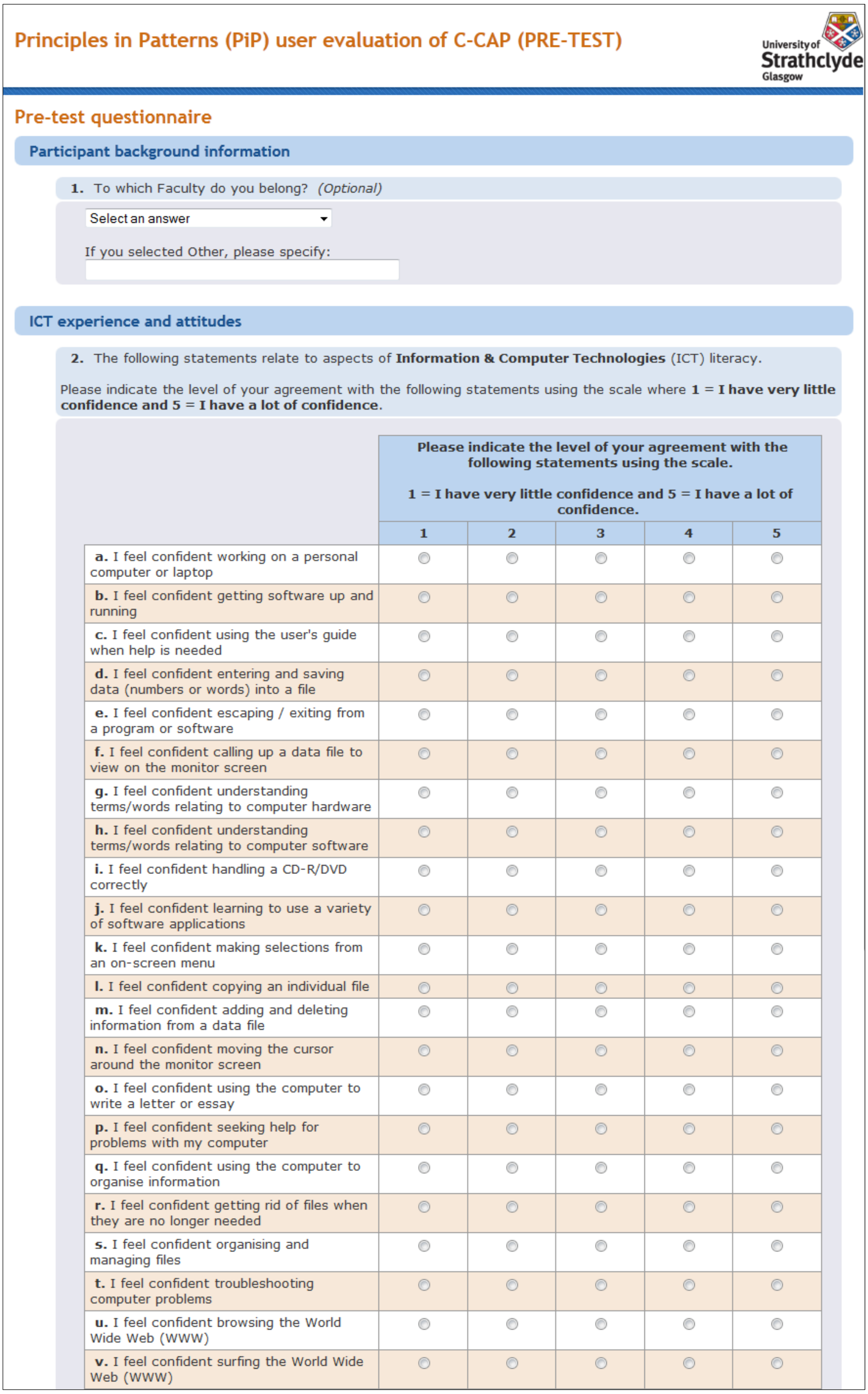
The participant information sheet outlines the activities you will be asked to engage in ("What will you do in the project?"). You will be asked to complete a questionnaire before you use the PiP system (pre-test) and a questionnaire after you use the PiP system (post-test).

Both questionnaires are completely anonymous. All data gathered will be held confidentially and will be stored according to the **Data Protection Act 1998**. Completion of either the pre- or post-test questionnaires is not expected to take longer than 10 minutes.

The survey is on several pages and it is not possible to return to a page once it has been completed. Therefore please think carefully before responding so that your views are accurately represented.

Once you click "continue" you will be directed to the first section of the survey.

Figure 17: Pre-session questionnaire instrument, page 1.



Principles in Patterns (PiP) user evaluation of C-CAP (PRE-TEST)

Pre-test questionnaire

Participant background information

1. To which Faculty do you belong? *(Optional)*

Select an answer ▾

If you selected Other, please specify:

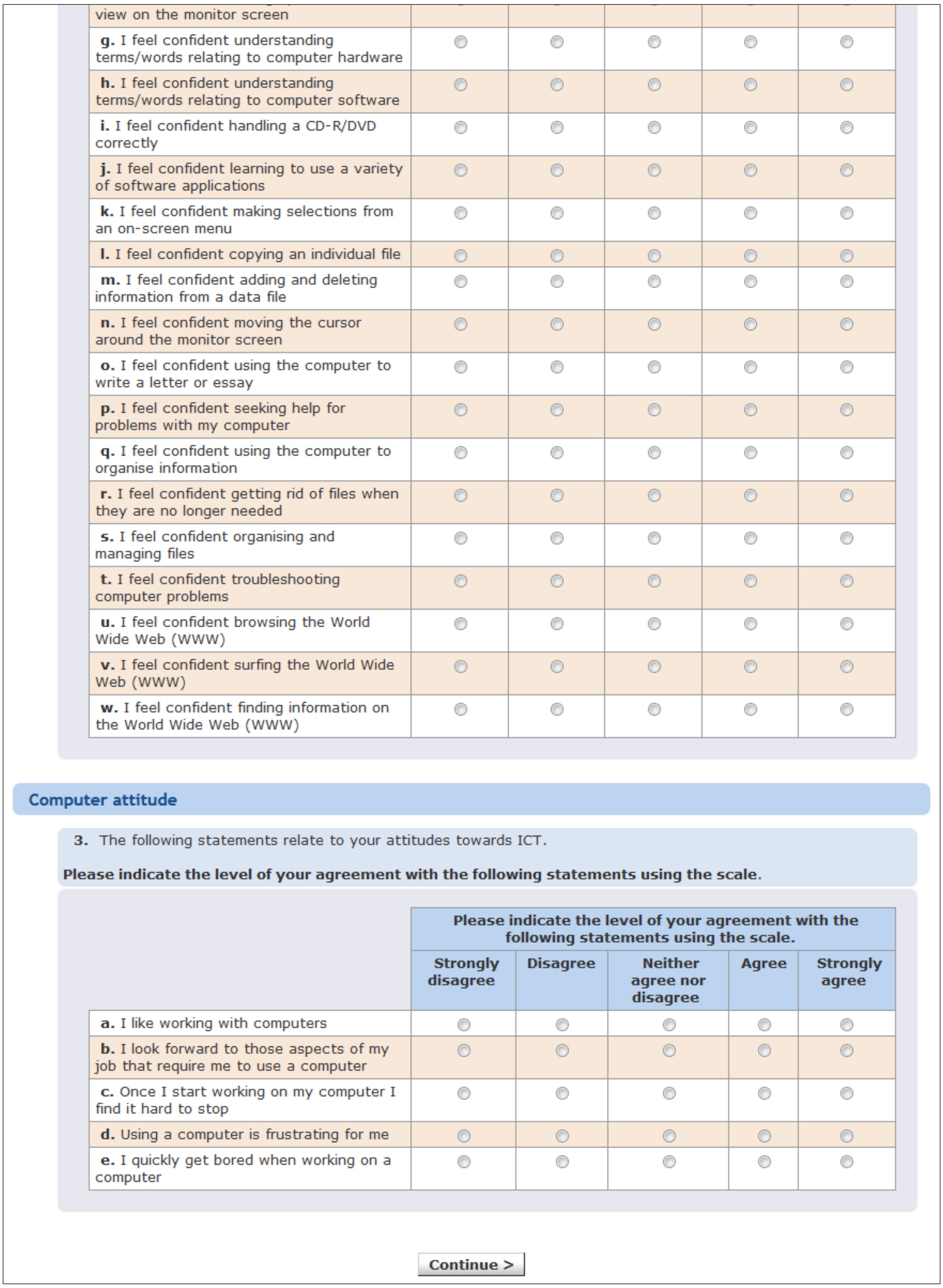
ICT experience and attitudes

2. The following statements relate to aspects of **Information & Computer Technologies** (ICT) literacy.

Please indicate the level of your agreement with the following statements using the scale where **1 = I have very little confidence and 5 = I have a lot of confidence**.

	Please indicate the level of your agreement with the following statements using the scale. 1 = I have very little confidence and 5 = I have a lot of confidence.				
	1	**2**	**3**	**4**	**5**
a. I feel confident working on a personal computer or laptop	○	○	○	○	○
b. I feel confident getting software up and running	○	○	○	○	○
c. I feel confident using the user's guide when help is needed	○	○	○	○	○
d. I feel confident entering and saving data (numbers or words) into a file	○	○	○	○	○
e. I feel confident escaping / exiting from a program or software	○	○	○	○	○
f. I feel confident calling up a data file to view on the monitor screen	○	○	○	○	○
g. I feel confident understanding terms/words relating to computer hardware	○	○	○	○	○
h. I feel confident understanding terms/words relating to computer software	○	○	○	○	○
i. I feel confident handling a CD-R/DVD correctly	○	○	○	○	○
j. I feel confident learning to use a variety of software applications	○	○	○	○	○
k. I feel confident making selections from an on-screen menu	○	○	○	○	○
l. I feel confident copying an individual file	○	○	○	○	○
m. I feel confident adding and deleting information from a data file	○	○	○	○	○
n. I feel confident moving the cursor around the monitor screen	○	○	○	○	○
o. I feel confident using the computer to write a letter or essay	○	○	○	○	○
p. I feel confident seeking help for problems with my computer	○	○	○	○	○
q. I feel confident using the computer to organise information	○	○	○	○	○
r. I feel confident getting rid of files when they are no longer needed	○	○	○	○	○
s. I feel confident organising and managing files	○	○	○	○	○
t. I feel confident troubleshooting computer problems	○	○	○	○	○
u. I feel confident browsing the World Wide Web (WWW)	○	○	○	○	○
v. I feel confident surfing the World Wide Web (WWW)	○	○	○	○	○

Figure 18: Pre-session questionnaire instrument, page 2. Includes CSE instrument [30].



view on the monitor screen					
g. I feel confident understanding terms/words relating to computer hardware	○	○	○	○	○
h. I feel confident understanding terms/words relating to computer software	○	○	○	○	○
i. I feel confident handling a CD-R/DVD correctly	○	○	○	○	○
j. I feel confident learning to use a variety of software applications	○	○	○	○	○
k. I feel confident making selections from an on-screen menu	○	○	○	○	○
l. I feel confident copying an individual file	○	○	○	○	○
m. I feel confident adding and deleting information from a data file	○	○	○	○	○
n. I feel confident moving the cursor around the monitor screen	○	○	○	○	○
o. I feel confident using the computer to write a letter or essay	○	○	○	○	○
p. I feel confident seeking help for problems with my computer	○	○	○	○	○
q. I feel confident using the computer to organise information	○	○	○	○	○
r. I feel confident getting rid of files when they are no longer needed	○	○	○	○	○
s. I feel confident organising and managing files	○	○	○	○	○
t. I feel confident troubleshooting computer problems	○	○	○	○	○
u. I feel confident browsing the World Wide Web (WWW)	○	○	○	○	○
v. I feel confident surfing the World Wide Web (WWW)	○	○	○	○	○
w. I feel confident finding information on the World Wide Web (WWW)	○	○	○	○	○

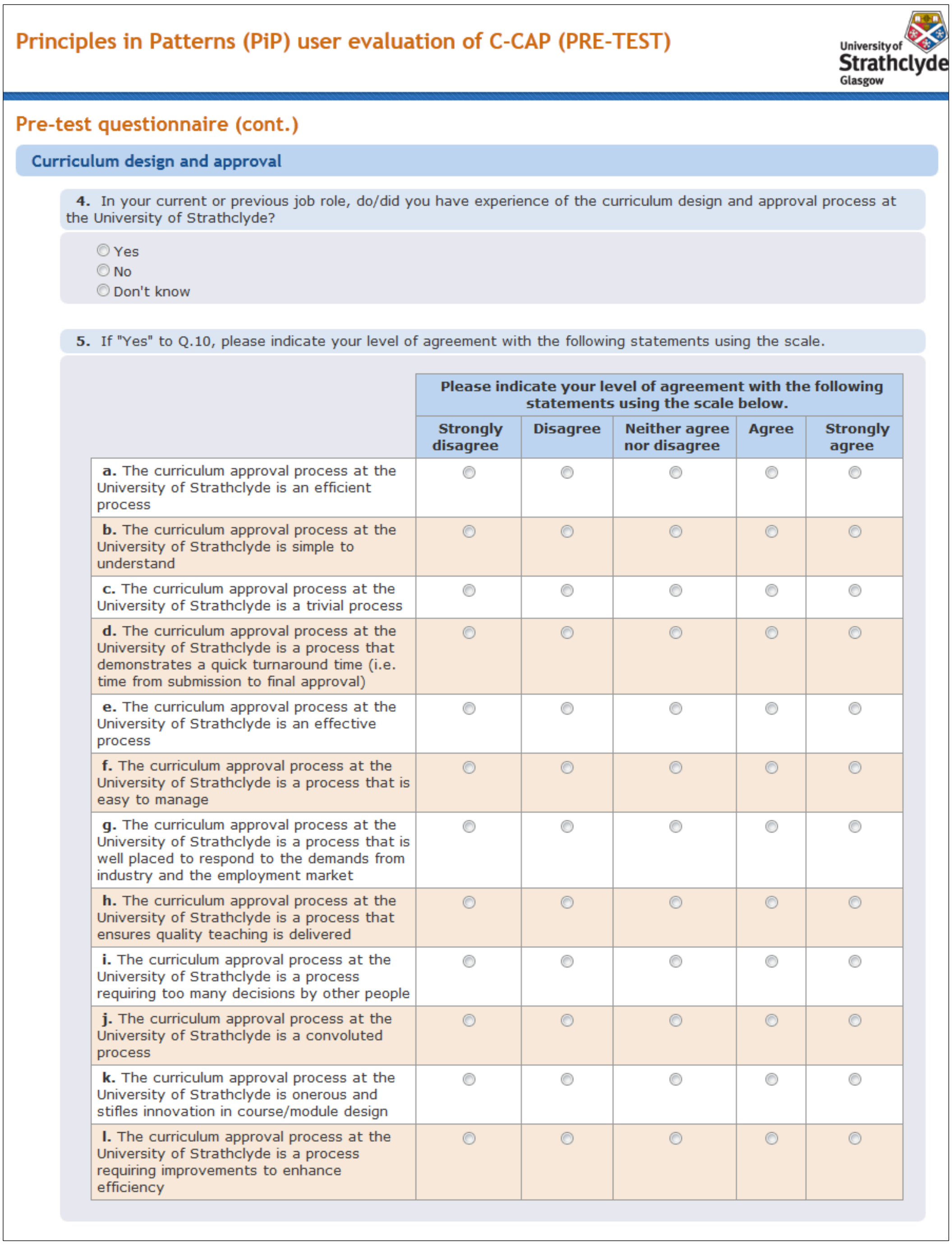
Computer attitude

3. The following statements relate to your attitudes towards ICT.

Please indicate the level of your agreement with the following statements using the scale.

	Please indicate the level of your agreement with the following statements using the scale.				
	Strongly disagree	**Disagree**	**Neither agree nor disagree**	**Agree**	**Strongly agree**
a. I like working with computers	○	○	○	○	○
b. I look forward to those aspects of my job that require me to use a computer	○	○	○	○	○
c. Once I start working on my computer I find it hard to stop	○	○	○	○	○
d. Using a computer is frustrating for me	○	○	○	○	○
e. I quickly get bored when working on a computer	○	○	○	○	○

Figure 19: Pre-session questionnaire instrument, page 2 continued.



Principles in Patterns (PiP) user evaluation of C-CAP (PRE-TEST)

Pre-test questionnaire (cont.)

Curriculum design and approval

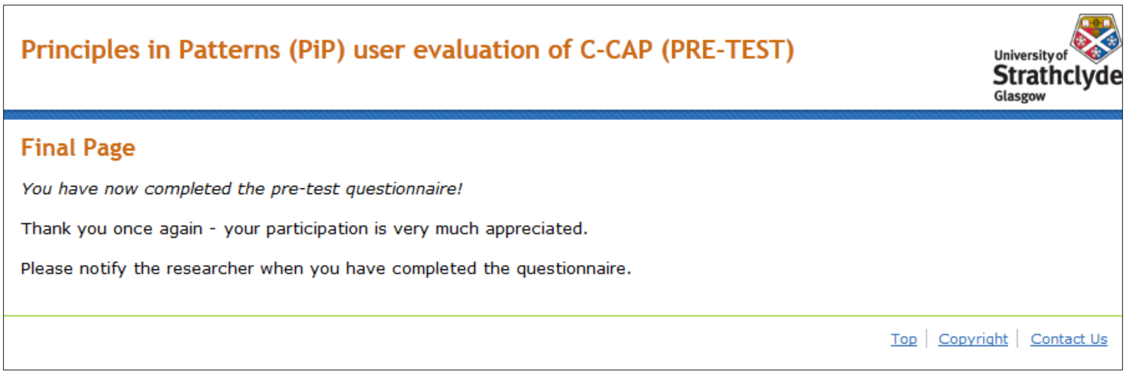
4. In your current or previous job role, do/did you have experience of the curriculum design and approval process at the University of Strathclyde?

○ Yes
○ No
○ Don't know

5. If "Yes" to Q.10, please indicate your level of agreement with the following statements using the scale.

	Please indicate your level of agreement with the following statements using the scale below.				
	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
a. The curriculum approval process at the University of Strathclyde is an efficient process	○	○	○	○	○
b. The curriculum approval process at the University of Strathclyde is simple to understand	○	○	○	○	○
c. The curriculum approval process at the University of Strathclyde is a trivial process	○	○	○	○	○
d. The curriculum approval process at the University of Strathclyde is a process that demonstrates a quick turnaround time (i.e. time from submission to final approval)	○	○	○	○	○
e. The curriculum approval process at the University of Strathclyde is an effective process	○	○	○	○	○
f. The curriculum approval process at the University of Strathclyde is a process that is easy to manage	○	○	○	○	○
g. The curriculum approval process at the University of Strathclyde is a process that is well placed to respond to the demands from industry and the employment market	○	○	○	○	○
h. The curriculum approval process at the University of Strathclyde is a process that ensures quality teaching is delivered	○	○	○	○	○
i. The curriculum approval process at the University of Strathclyde is a process requiring too many decisions by other people	○	○	○	○	○
j. The curriculum approval process at the University of Strathclyde is a convoluted process	○	○	○	○	○
k. The curriculum approval process at the University of Strathclyde is onerous and stifles innovation in course/module design	○	○	○	○	○
l. The curriculum approval process at the University of Strathclyde is a process requiring improvements to enhance efficiency	○	○	○	○	○

Figure 20: Pre-session questionnaire instrument, page 4.



Principles in Patterns (PiP) user evaluation of C-CAP (PRE-TEST)

University of Strathclyde Glasgow

Final Page

You have now completed the pre-test questionnaire!

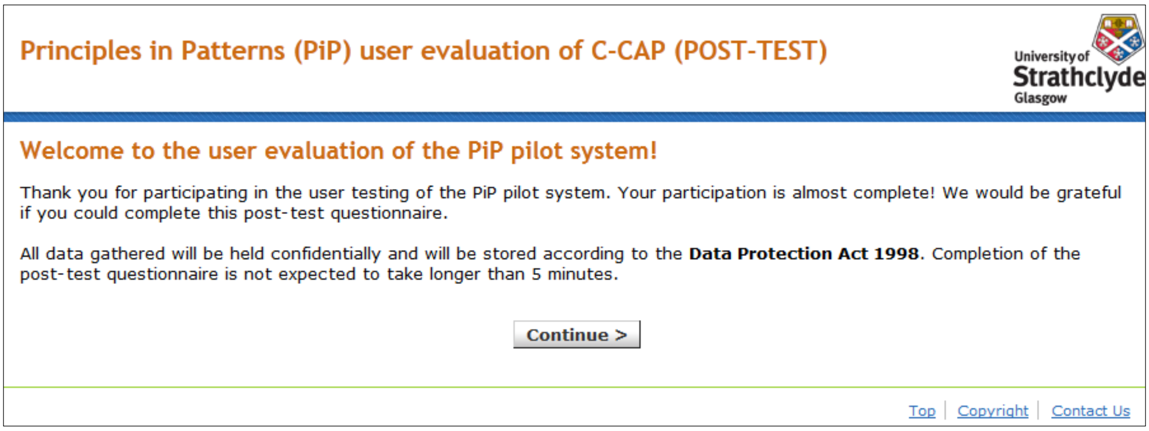
Thank you once again - your participation is very much appreciated.

Please notify the researcher when you have completed the questionnaire.

Top | Copyright | Contact Us

Figure 21: Pre-session questionnaire instrument, page 5.

12. Appendix G: Post-session questionnaire instrument in BOS



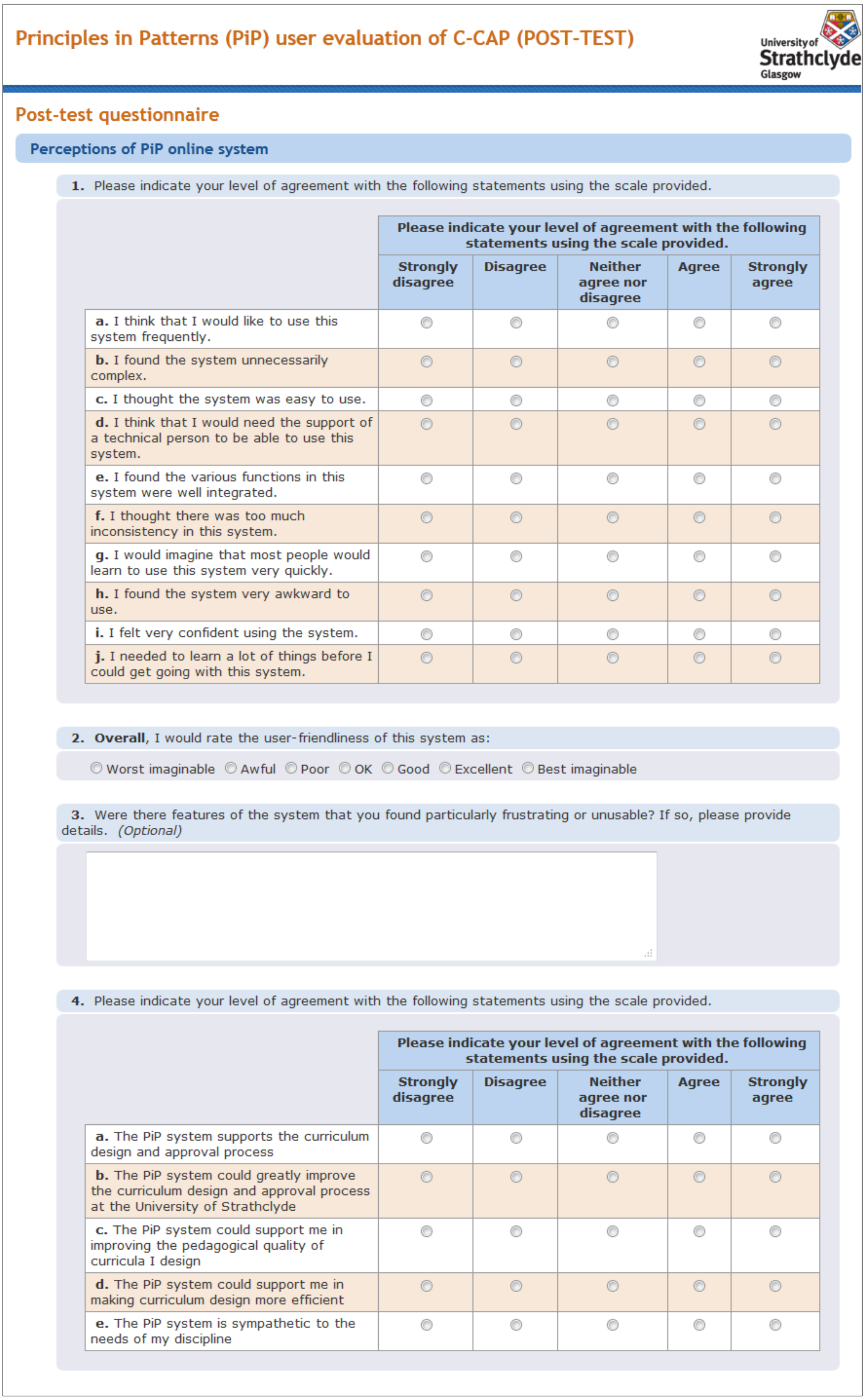
Principles in Patterns (PiP) user evaluation of C-CAP (POST-TEST)

Welcome to the user evaluation of the PiP pilot system!

Thank you for participating in the user testing of the PiP pilot system. Your participation is almost complete! We would be grateful if you could complete this post-test questionnaire.

All data gathered will be held confidentially and will be stored according to the **Data Protection Act 1998**. Completion of the post-test questionnaire is not expected to take longer than 5 minutes.

Figure 22: Post-session questionnaire instrument, page 1.



Principles in Patterns (PiP) user evaluation of C-CAP (POST-TEST)

Post-test questionnaire

Perceptions of PiP online system

1. Please indicate your level of agreement with the following statements using the scale provided.

	Please indicate your level of agreement with the following statements using the scale provided.				
	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
a. I think that I would like to use this system frequently.	○	○	○	○	○
b. I found the system unnecessarily complex.	○	○	○	○	○
c. I thought the system was easy to use.	○	○	○	○	○
d. I think that I would need the support of a technical person to be able to use this system.	○	○	○	○	○
e. I found the various functions in this system were well integrated.	○	○	○	○	○
f. I thought there was too much inconsistency in this system.	○	○	○	○	○
g. I would imagine that most people would learn to use this system very quickly.	○	○	○	○	○
h. I found the system very awkward to use.	○	○	○	○	○
i. I felt very confident using the system.	○	○	○	○	○
j. I needed to learn a lot of things before I could get going with this system.	○	○	○	○	○

2. Overall, I would rate the user-friendliness of this system as:

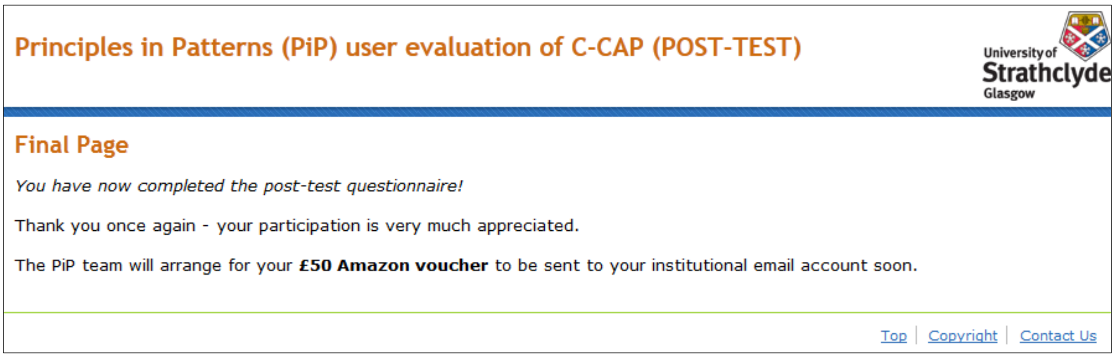
○ Worst imaginable ○ Awful ○ Poor ○ OK ○ Good ○ Excellent ○ Best imaginable

3. Were there features of the system that you found particularly frustrating or unusable? If so, please provide details. *(Optional)*

4. Please indicate your level of agreement with the following statements using the scale provided.

	Please indicate your level of agreement with the following statements using the scale provided.				
	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
a. The PiP system supports the curriculum design and approval process	○	○	○	○	○
b. The PiP system could greatly improve the curriculum design and approval process at the University of Strathclyde	○	○	○	○	○
c. The PiP system could support me in improving the pedagogical quality of curricula I design	○	○	○	○	○
d. The PiP system could support me in making curriculum design more efficient	○	○	○	○	○
e. The PiP system is sympathetic to the needs of my discipline	○	○	○	○	○

Figure 23: Post-session questionnaire instrument (page 2), including SUS and ARS questions [32], [33].



Principles in Patterns (PiP) user evaluation of C-CAP (POST-TEST)

University of Strathclyde Glasgow

Final Page

You have now completed the post-test questionnaire!

Thank you once again - your participation is very much appreciated.

The PiP team will arrange for your **£50 Amazon voucher** to be sent to your institutional email account soon.

Top | Copyright | Contact Us

Figure 24: Post-session questionnaire instrument, page 3.

13. Appendix H: Table of heuristic issues derived from protocol analysis

Table 9: Example table of heuristic issues to be resolved in C-CAP, as partially derived from the super-node "System issues" and its sub-nodes.

Issue #	Issue description	Issue severity
1	Credit value should be 10 or 20. (Core information)	2
2	Specify academic level (Core information), not UG or PG	4
3	What's the definition of "Open".	3
4	Use "compulsory" rather than "mandatory".	1
5	A drop down menu for existing course codes.	2
6	Section 2.2 onwards – easier to insert the class code first.	1
7	Auto-populating elements of section 2. (curriculum cohesion)	1
8	Page error preventing form submission.	3
9	Entering "Not applicable" – wide unneeded boxes	4
10	Teaching across different sites – not visible on first screen. Insufficient options?	2
11	Need to generate read only version for printing. Faster to write for drafting, etc.	1
12	Extra options in drop down menu for Activity type (section 4). Some departments set homework and it is a defined task. List to add: computer lab, "Other – please specify", need for notes field / further information field	6
13	Ability to click "all" learning objectives assessed.	5
14	No deadline should be associated with examination.	3
15	Ability to accommodate anomalous assessment situation whereby students get examination exemptions.	1
16	Change class evaluation options; too difficult to understand.	6
17	Unclear how to delete an item.	1
18	Recommended reading and resources needs addressing. Too confused and conflates too many resources. Rooms should be taken for granted? Need for bibliographic elements.	5
19	Class session types (Activity 4.1) – add to list: computer lab, "Other – please specify", etc.	3
20	Idea of assessment duration problematic / type - how long if included?	8
21	Additional assessment types	4
22	Problem understanding difference between class and course in section 2.1	3
23	No deadlines for courseworks – flexibility required.	3
24	Need for dummy course codes – how to amend an existing class, e.g. search by module code?	3
25	Section 3.3 vague – examples required.	1
26	Larger syllabus box?	1
27	Save and submit error (e.g. "Some rules were not applied").	1
28	Departmental name corrections in Core Information required.	1
29	Help – learning outcome examples for disciplines.	1
30	Insert item – purpose of button unclear.	2
31	Module leader details – personal ownership required.	1
32	Core information screen – need for note of compulsory information.	1
33	Insert buttons unresponsive (e.g. Add a learning objective). Requires clicking twice.	2
34	Retention of blank learning objectives during constructive alignment.	1
35	Section 4.3 – adding of assessment weighting and removal of warning.	1
36	Check weightings to be consistent (section 4.3)	1

code{4}lib JOURNAL

Issue 43, 2019-02-14

Improving the discoverability and web impact of open repositories: techniques and evaluation

In this contribution we experiment with a suite of repository adjustments and improvements performed on Strathprints, the University of Strathclyde, Glasgow, institutional repository powered by EPrints 3.3.13. These adjustments were designed to support improved repository web visibility and user engagement, thereby improving usage. Although the experiments were performed on EPrints it is thought that most of the adopted improvements are equally applicable to any other repository platform. Following preliminary results reported elsewhere, and using Strathprints as a case study, this paper outlines the approaches implemented, reports on comparative search traffic data and usage metrics, and delivers conclusions on the efficacy of the techniques implemented. The evaluation provides persuasive evidence that specific enhancements to technical aspects of a repository can result in significant improvements to repository visibility, resulting in a greater web impact and consequent increases in content usage. COUNTER usage grew by 33% and traffic to Strathprints from Google and Google Scholar was found to increase by 63% and 99% respectively. Other insights from the evaluation are also explored. The results are likely to positively inform the work of repository practitioners and open scientists.

by George Macgregor

Introduction

Significant resource has been invested over the past decade to expose rich digital collections using a variety of repository technologies. This investment has resulted in unprecedented usage of institutional repositories, as evidenced in the UK by services such as IRUS-UK which, at time of writing, has recorded 146,398,650 COUNTER compliant downloads from participating repositories since 2013 [1]. However, many institutions continue to demonstrate limited commitment to ensuring their scholarly content is exposed optimally. This also extends to a failure to ensure their repository is as usable as possible. In fact, many repositories have not undergone development beyond their original establishment and maintenance of its scholarly collection. The reasons for this inertia are complex and it is not the purpose of this paper to explore them. However, it is sufficient to state that such institutions may attempt to promote their repository content but if few attempts have been made to optimise for discovery, then these repositories may find themselves under exposed [2] and under used.

Significant future challenges are facing Open Access repositories, as well as the open science movement more generally [3]. Competing scholarly platforms, many of which are proprietary, appear to be growing in popularity yet demonstrate poor support for open standards or prevalent open science technical protocols, as well as low levels of integration with open scholarly infrastructure. It is therefore imperative that user expectations of repositories are better met and improvements to the index penetration and exposure of their scholarly content demonstrated. Only by doing this will scholarly Open Access repositories validate their continued relevance in open scholarly communication.

In this contribution we experiment with a suite of repository adjustments and improvements performed on Strathprints [4], the University of Strathclyde institutional repository powered by EPrints 3.3.13. These adjustments were designed to support improved repository web visibility and user engagement, thereby improving usage. Although the experiments were performed on EPrints it is thought that most of the adopted improvements are equally applicable to any other repository platform. Following preliminary results reported elsewhere [5], and using Strathprints as a case study, this paper will outline the approaches implemented, report on comparative search traffic data and usage metrics, and deliver conclusions on the efficacy of the techniques implemented. The results are likely to positively inform the work of repository practitioners and open scientists.

Background

Given the importance of institutional repositories in promoting open scholarly communication and the discovery of open research content, it is perhaps surprising to note that only a limited amount of prior work has been documented on repository discoverability approaches and their evaluation. Many contributions note the importance of repository discoverability and report on some of the factors that should be addressed [6], but few then evaluate the impact of these factors. Most recently, however, the Code4Lib Journal published a contribution on the use of microdata within institutional repositories as a “low barrier” means of better exposing contents to Google [7]. This work described the implementation of Schema.org within DSpace. It is a notable contribution owing to the fact that repository support for Schema.org is a feature of the COAR Next Generation Repositories agenda [8]. Pekala reported generally positive results but conceded that demonstrating its impact was difficult.

Kelly and Nixon reported on the use of general SEO techniques on three separate UK repositories [2]. This work relied on analytics services and tested early data indicating the importance of blogs in driving repository web traffic. The authors reported mixed results and therefore concluded that further work was required in order to refine their methodology and better understand search engine behaviour. In a poster presented at the 2017 Repository Fringe Conference, the present author evaluated the preliminary results derived from a series of repository enhancements designed to improve web impact and discoverability. While some encouraging evidence was reported about the impact of specific repository enhancements, the small nature of the evaluation prohibited any wider conclusions to be drawn.

Others have focused on hypothesised impediments to repository discoverability. For example, Tonkin et al. explored the significance of repository coversheets in disrupting the bot crawling potential of repositories in some cases [9], a practice also considered by Anurag Acharya of Google Scholar as undesirable [10]. Better supporting Google Scholar indexing was addressed by Arlitsch and O'Brien, who noted variable indexing coverage of repositories on GS and evaluated the effects of adjusting in-page metadata on GS indexing penetration. Arlitsch and O'Brien highlight the dangers of paying insufficient attention to discoverability and propose corrective actions for repository managers to perform.

Promoting repository discovery

Whilst many of the prominent repository platforms (e.g. EPrints, DSpace, Digital Commons, OJS, etc.) now provide basic out-of-the-box support for discovery and interoperability with key academic tools, including meeting Google Scholar inclusion guidelines, there remains wide variation on the relative visibility and discoverability of repository content. The question of repository discoverability is therefore something which has attracted significant attention at the University of Strathclyde as the institution seeks to ensure its internationally significant research [11], much of it available open access via Strathprints, can be found easily.

Strathprints is powered by EPrints (version 3.3.13). To improve repository web visibility and user engagement, thereby improving usage, a series of technical changes were made to Strathprints in spring 2016 and their impact monitored during 2016/2017, and again in 2017/2018. Process improvements were also implemented. The changes could be said to fall into one of two categories: improvements, and; adjustments. "Improvements" were changes that resulted in substantive modifications to repository functionality, while "adjustments" included actions that sought to refine existing aspects of the repository. As noted below, much of the motivation for these improvements and adjustments came from the broader literature on web publication best practice and SEO; although some were gleaned from the repository best practice literature [10].

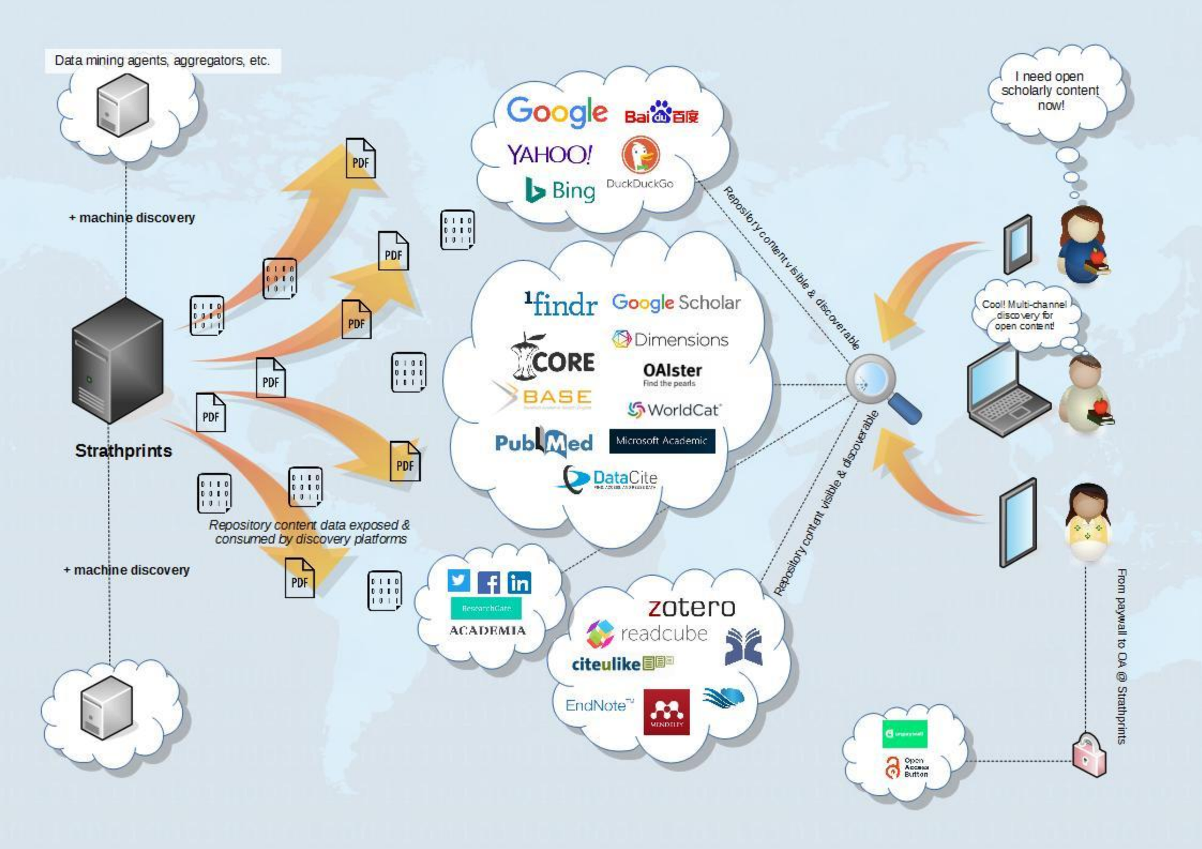

Data mining agents, aggregators, etc.
+ machine discovery
Strathprints
PDF
Repository content data exposed & consumed by discovery platforms
+ machine discovery
Google
Baidu 百度
YAHOO!
DuckDuckGo
Bing
1findr
Google Scholar
Dimensions
CORE
OAIster
Find the pearls
BASE
WorldCat
PubMed
Microsoft Academic
DataCite
ResearchGate
ACADEMIA
zotero
readcube
citeulike
EndNote
MENDELEY
unpaywall
Open Access Button
Repository content visible & discoverable
Repository content visible & discoverable
I need open scholarly content now!
Cool! Multi-channel discovery for open content!
From paywall to OA @ Strathprints

Figure 1. Discovery paths for content stored in the Strathprints repository.

Technical changes

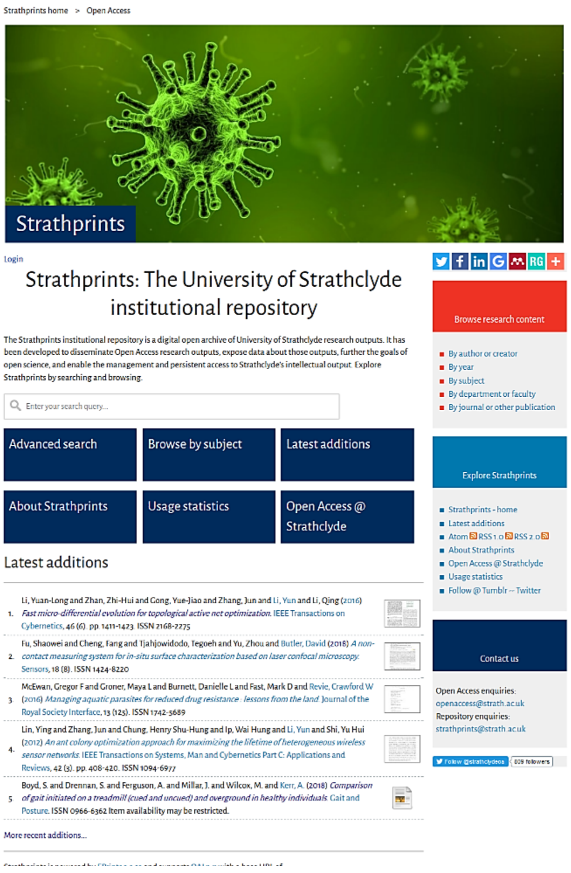
Improvements

The principal improvements made included:

- Implementation of a refreshed Strathprints user interface (UI). Many repositories continue to demonstrate low levels of usability [12], [13]. Low levels of usability can result in the users' abandonment of a website or of system sessions [14], [15], [16]. An heuristic evaluation [17] of Strathprints user interface (UI) was therefore undertaken in early 2016 to direct UI changes intended to improve usability and user engagement (Figure 2 & 3);
- Following heuristic evaluation, a "mobile first", responsive re-engineering of Strathprints was implemented, thereby triggering important signals in PageRank [18] and, later, heavier weighting in the Google "Penguin" updates [19] (Figure 2 & 3);
- "White hat" improvements [20] to the way Strathprints functions. This included improvements to internal linking (e.g. navigation, hyperlink labels, etc.) and content improvements including promotion of user interaction through support for the Core Recommender and AltMetric. Both of these

improvements stimulate additional user interaction. For example, in the case of the Core Recommender this is achieved by referring users to alternative but related additional Strathprints content, recommended to the user on the basis of the repository item they are currently browsing;

- Improved integration with social tools, including growth in social interactions which are the result of Tweets about recently deposited Strathprints content;
- Implementation of a "connector-lite" configuration actioned to cultivate Strathprints as a full-text destination for users and machines alike [21]. Within the currently scholarly communication landscape it is not uncommon for institutional repositories to now operate in parallel with the local Current Research Information System (CRIS). This so-called "connected" configuration enables metadata and digital content exchange from the CRIS to the repository. It is a configuration that applies to Strathprints, which is no longer a point of entry for staff wishing to deposit content in Strathprints; instead users deposit via the CRIS which then automatically writes metadata and content to Strathprints. "Connector-lite", however, enables greater control over what is written to Strathprints by the institutional CRIS [21].



Strathprints home > Open Access

Strathprints

Login

Strathprints: The University of Strathclyde institutional repository

The Strathprints institutional repository is a digital open archive of University of Strathclyde research outputs. It has been developed to disseminate Open Access research outputs, expose data about those outputs, further the goals of open science, and enable the management and persistent access to Strathclyde's intellectual output. Explore Strathprints by searching and browsing.

Enter your search query...

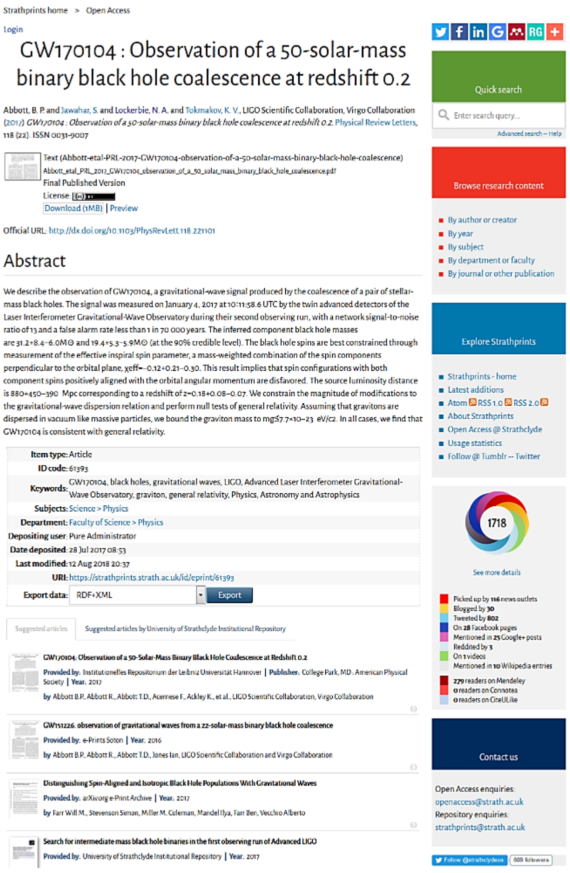
Latest additions

1. Li, Yuan-Long and Zhan, Zhi-Hui and Gong, Yue-Jiao and Zhang, Jun and Li, Yun and Li, Qing (2016) *Fast micro-differential evolution for topological active net optimization.* IEEE Transactions on Cybernetics, 46 (6). pp. 1411-1423. ISSN 2168-2275
2. Fu, Shaowei and Cheng, Fang and Tjahjowidodo, Tegoeh and Yu, Zhou and Butler, David (2018) *A non-contact measuring system for in-situ surface characterization based on laser confocal microscopy.* Sensors, 18 (8). ISSN 1424-8220
3. McEwan, Gregor F and Groner, Maya L and Burnett, Danielle L and Fast, Mark D and Revie, Crawford W (2016) *Managing aquatic parasites for reduced drug resistance : lessons from the land.* Journal of the Royal Society Interface, 13 (125). ISSN 1742-5689
4. Lin, Ying and Zhang, Jun and Chung, Henry Shu-Hung and Ip, Wai Hung and Li, Yun and Shi, Yu Hui (2012) *An ant colony optimization approach for maximizing the lifetime of heterogeneous wireless sensor networks.* IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 42 (3). pp. 408-420. ISSN 1094-6977
5. Boyd, S. and Drennan, S. and Ferguson, A. and Millar, J. and Wilcox, M. and Kerr, A. (2018) *Comparison of gait initiated on a treadmill (cued and uncued) and overground in healthy individuals.* Gait and Posture. ISSN 0966-6362 Item availability may be restricted.

More recent additions...

Open Access enquiries: openaccess@strath.ac.uk
Repository enquiries: strathprints@strath.ac.uk

Figure 2. Strathprints UI (homepage).



Strathprints home > Open Access

Login

GW170104 : Observation of a 50-solar-mass binary black hole coalescence at redshift 0.2

Abbott, B. P. and Jawahar, S. and Lockerbie, N. A. and Tokmakov, K. V., LIGO Scientific Collaboration, Virgo Collaboration (2017) *GW170104 : Observation of a 50-solar-mass binary black hole coalescence at redshift 0.2.* Physical Review Letters, 118 (22). ISSN 0031-9007

Text (Abbott-etal-PRL-2017-GW170104-observation-of-a-50-solar-mass-binary-black-hole-coalescence)
Abbott_etal_PRL_2017_GW170104_observation_of_a_50_solar_mass_binary_black_hole_coalescence.pdf
Final Published Version
License:
Download (1MB) | Preview

Official URL: http://dx.doi.org/10.1103/PhysRevLett.118.221101

Abstract

We describe the observation of GW170104, a gravitational-wave signal produced by the coalescence of a pair of stellar-mass black holes. The signal was measured on January 4, 2017 at 10:11:58.6 UTC by the twin advanced detectors of the Laser Interferometer Gravitational-Wave Observatory during their second observing run, with a network signal-to-noise ratio of 13 and a false alarm rate less than 1 in 70 000 years. The inferred component black hole masses are 31.2+8.4−6.0M⊙ and 19.4+5.3−5.9M⊙ (at the 90% credible level). The black hole spins are best constrained through measurement of the effective inspiral spin parameter, a mass-weighted combination of the spin components perpendicular to the orbital plane, χeff=−0.12+0.21−0.30. This result implies that spin configurations with both component spins positively aligned with the orbital angular momentum are disfavored. The source luminosity distance is 880+450−390 Mpc corresponding to a redshift of z=0.18+0.08−0.07. We constrain the magnitude of modifications to the gravitational-wave dispersion relation and perform null tests of general relativity. Assuming that gravitons are dispersed in vacuum like massive particles, we bound the graviton mass to mg≤7.7×10−23 eV/c2. In all cases, we find that GW170104 is consistent with general relativity.

Item type:	Article
ID code:	61393
Keywords:	GW170104, black holes, gravitational waves, LIGO, Advanced Laser Interferometer Gravitational-Wave Observatory, graviton, general relativity, Physics, Astronomy and Astrophysics
Subjects:	Science > Physics
Department:	Faculty of Science > Physics
Depositing user:	Pure Administrator
Date deposited:	28 Jul 2017 08:53
Last modified:	12 Aug 2018 20:37
URI:	https://strathprints.strath.ac.uk/id/eprint/61393
Export data:	RDF+XML Export

Suggested articles | Suggested articles by University of Strathclyde Institutional Repository

GW170104: Observation of a 50-Solar-Mass Binary Black Hole Coalescence at Redshift 0.2
Provided by: Institutionelles Repositorium der Leibniz Universität Hannover | **Publisher:** College Park, MD : American Physical Society | **Year:** 2017
by Abbott B.P., Abbott R., Abbott T.D., Acernese F., Ackley K., et al., LIGO Scientific Collaboration, Virgo Collaboration

GW151226: observation of gravitational waves from a 22-solar-mass binary black hole coalescence
Provided by: e-Prints Soton | **Year:** 2016
by Abbott B.P., Abbott R., Abbott T.D., Jones Ian, LIGO Scientific Collaboration and Virgo Collaboration

Distinguishing Spin-Aligned and Isotropic Black Hole Populations With Gravitational Waves
Provided by: arXiv.org e-Print Archive | **Year:** 2017
by Farr Will M., Stevenson Simon, Miller M. Coleman, Mandel Ilya, Farr Ben, Vecchio Alberto

Search for intermediate mass black hole binaries in the first observing run of Advanced LIGO
Provided by: University of Strathclyde Institutional Repository | **Year:** 2017

Picked up by 116 news outlets
Blogged by 30
Tweeted by 802
On 28 Facebook pages
Mentioned in 25 Google+ posts
Reddited by 3
On 1 videos
Mentioned in 10 Wikipedia entries
279 readers on Mendeley
0 readers on Connotea
0 readers on CiteULike

Open Access enquiries: openaccess@strath.ac.uk
Repository enquiries: strathprints@strath.ac.uk

Figure 3. Strathprints UI (abstract pages).

Adjustments

A series of adjustments were made to fine tune the search engine friendliness of Strathprints and to enhance user experience. A number of these related to delivering page speed improvements for Strathprints, in line with trends within search agents to factor speed in results rankings ([18], [19], [20], [22]).

- Adjustments to the file-naming conventions used for deposited full-text files in order to render them more crawler friendly. Descriptive file-names can lead to better and more effective crawling of files. Moreover, words contained in file-names factor in retrieval algorithms and may be highlighted to users in results pages, so accurate naming is necessary to facilitate 'known-item' searching by users. A descriptive file naming convention with proactive use of hyphens to separate words in the filename [18] was therefore adopted. The broad approach to naming was as follows:

  {Author surname(s)}{Journal/conference acronym}{Year of publication}{Selected uninterrupted words from title of article using hyphens for spacing}.pdf

  So, for example, a file pertaining to the present article would be named:

  Macgregor-C4L-2019-Improving-the-discoverability-and-web-impact-of-open-repositories-techniques-and-evaluation.pdf

- Gradual cleaning of broken links within Strathprints thereby improving the "content health" of Strathprints and, again, triggering important signals in PageRank [18]. Like many repositories of its type, Strathprints has been operating in one form or another for over 10 years and during that time has accumulated its fair share of "link rot";
- "Minification" of all relevant repository files (e.g. CSS, JS, etc.) to deliver increased page loading speeds. Minification refers to the process of

removing superfluous or redundant data without affecting how the resource is processed by browsers, e.g. code comments, formatting, white space characters, unused code, using shorter variable names, etc. This superfluous data may aid the human readability of the code but is not needed for the code to execute efficiently.

- Rationalisation of all CSS and Javascript (JS) files in order to remove unused rules and variables. This can be performed manually but there are automatic online tools (e.g. PurifyCSS, UnCSS! Online) which can analyse websites to determine which CSS rules are actually being applied to a given website, thus allowing redundant rules to be deleted. Similarly, there are code quality tools for JS (e.g. JSHint).
- Asynchronous loading of JS resources: Render-blocking JS is probably the single most difficult obstacle to overcome when attempting to deliver repository speed improvements (see [23] for further details). A repository like Strathprints, like most others, will require the loading of many JS resources in order to deliver important functionality. For Strathprints this includes native JS resources but also third-party JS such as the Google JSAPI, AltMetric API, analytics from Google Analytics and AddThis, as well as for any EPrints plugins that have been installed from the EPrints Bazaar. However, some simple experimentation can deduce whether it is necessary for JS to be loaded at the same time as the page itself since in many cases JS can actually be deferred until after page rendering [23]. HTML5 introduced the async attribute to be used with <script>. This Boolean attribute indicates that the browser should, if possible, execute the script asynchronously. For example:

```
<script type="text/javascript" async="async"
        src="https://www.google.com/jsapi"><!--padder--></script>
```

- GZIP compression: gzip is a file format and software application used for file compression and decompression. All modern browsers support and automatically negotiate gzip compression for all HTTP requests and, where used, gzip can compress the size of the transferred response by up to 90%. This significantly reduces the amount of time needed to download resources, reduces data usage for users, and improves the first render time repository pages. Enabling gzip, however, is an infrastructural task as it necessitates adjusting the repository server configuration so that it returns "gzipped" content to compliant browsers. gzip implementation is described in more details at [23].
- Revisiting image optimisation: The question of optimising images for delivery over the web will vary from repository to repository and, in fact, many repositories have very little visual content at all. Strathprints uses large banner images which, when not sufficiently compressed, were found to negatively influence page loading times [22]. All image resources were therefore compressed and optimised accordingly.
- Migration to InnoDB as the MySQL storage engine in order to improve repository performance: EPrints generally runs on MySQL, using MyISAM as the default storage engine, but table locking was found to be a DB performance issue thereby inhibiting the execution of simultaneous queries. InnoDB demonstrates concurrency, locking only the row(s) which are relevant to the DB query, leaving the rest of the table available for CRUD operations.
- Deployment of Google Data Highlighter: We noted earlier that exposing contents to Google could be improved through the implementation of Schema.org [7]. It was not possible in this instance to re-engineer EPrints in order to expose Schema.org interoperable data, although this may be something to be explored in future. Instead Google Data Highlighter – a pattern matching tool for structured data on websites – was deployed as a substitute [24].

Data and results

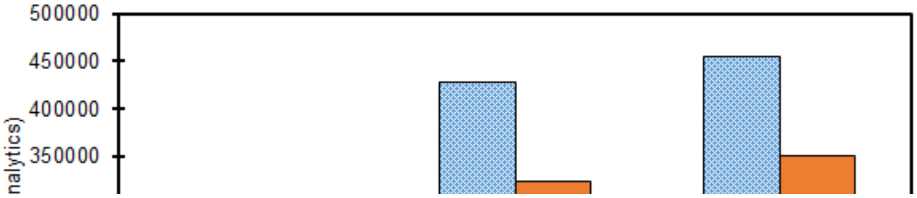
The impact of the repository changes were monitored and measured using a variety of metrics, including search traffic data from Google Search Console [25], COUNTER compliant usage data from IRUS-UK [1], Google Analytics [26] tracking data and routine statistical data from Strathprints itself. The periods examined were the year up to end July 2016 (Year 1 – Y1), prior to the changes being implemented; and the years up to end July 2017 (Year 2 – Y2), after improvements were deployed, and end July 2018 (Year 3 – Y3), after the adjustments were implemented.

Note that COUNTER usage data [1] refers to the international COUNTER 'Code of Practice', which sets standards on how electronic content usage is calculated thereby allowing content publishers to provide consistent, credible usage data. This data can then be used to accurately understand real world usage and provide usage comparisons across multiple services or websites.

Traffic

Web traffic, as measured by Google Analytics (GA), grew by 150,408 in Y2 to 428,407, equivalent to a 54% improvement when compared to Y1. A 52% improvement in unique traffic was also observed during the same period (Figure 4(a)). An increase in traffic in Y3 was less than Y2 (6%) but was still in excess of Y2 (n = 454,318), meaning that the total percentage growth in traffic during the entire reporting period was 63% and 65% for traffic and unique traffic respectively.

As might be expected, Google was found to be the largest referral source, accounting for 55% of all traffic in Y3; but thereafter Google Scholar was found to be the most significant referral source, accounting for 25% of all web traffic in Y3 and growing by 99% during the entire reporting period. Traffic in Y2 grew by 48% (n = 83,045) and 34% (n = 111,563) in Y3. 77% of all this traffic in both Y2 and Y3 was unique. This is at variance with previously reported results emerging from a preliminary evaluation [3], in which GS traffic was found to have declined slightly as a proportion of total web traffic (by 3%). In fact, this present evaluation, using a more comprehensive dataset, found the percentage of total traffic to Strathprints from GS to have increased by 5%, with almost all of these gains achieved during Y3. Thus, the percentage traffic gains achieved from GS during the reporting period (99%) grew even quicker than the broader gains achieved from other web traffic sources (63%). This can be observed data charted in Figure 4(b). Repositories serving more content enjoy deeper indexing by Google Scholar (GS) [10] and, combined with the other improvements and adjustments, may be a possible explanation for the GS improvements.



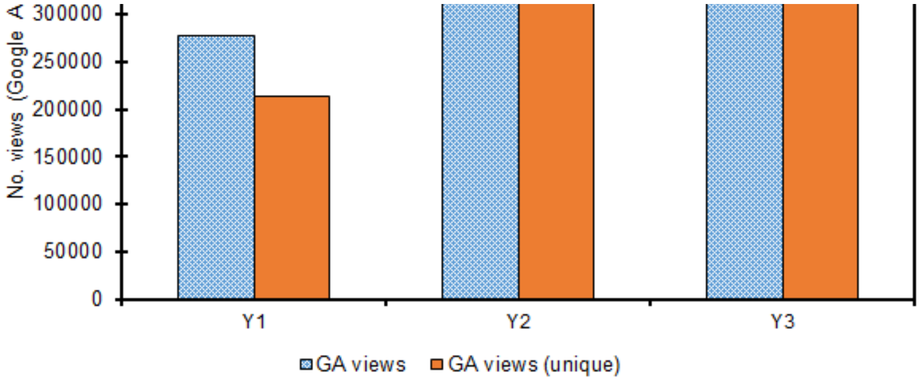


Figure 4(a). Volume of referral traffic (views and unique views) as calculated by Google Analytics (GA) in Y1, Y2 & Y3.

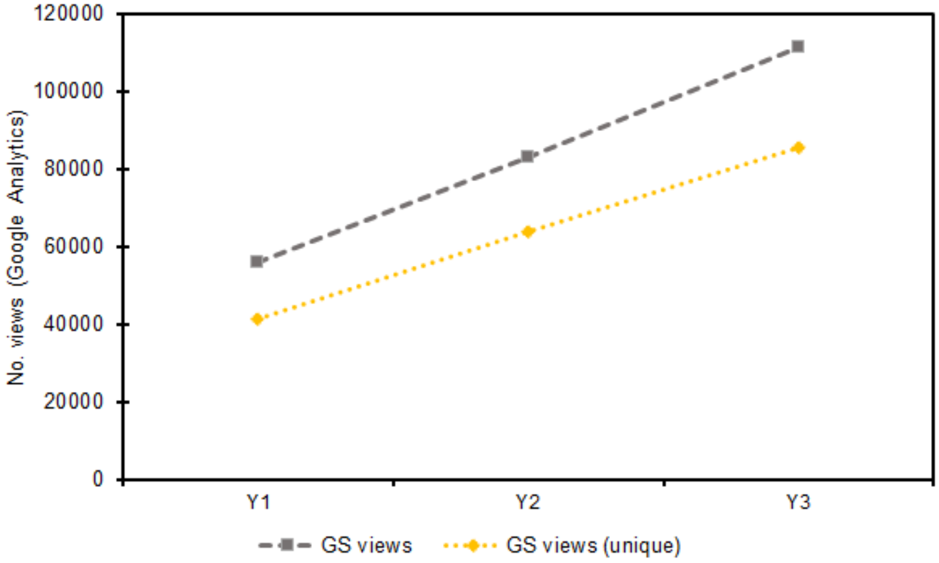


Figure 4(b). Volume of referral traffic from Google Scholar (GS) for views and unique view, as calculated by Google Analytics (GA) in Y1, Y2 & Y3.

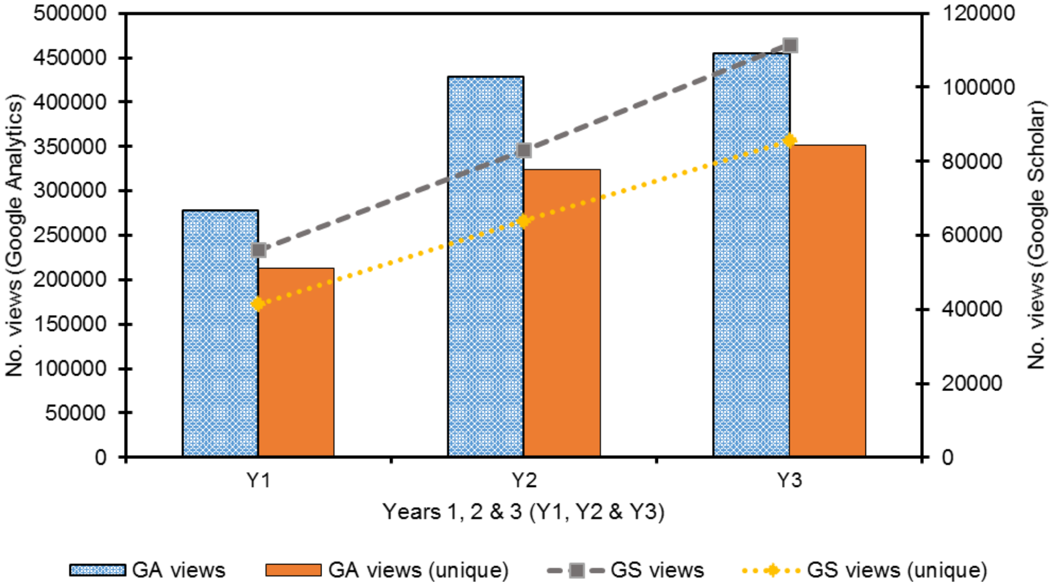


Figure 4(c). Volume of Google & Google Scholar referral traffic (views) in Y1, Y2 & Y3.

The principal referral sources remained largely unchanged during the reporting period, with both Google and GS referring the majority of the traffic. However, the proportion of the overall traffic referred to Strathprints by Google and GS grew by 18% between Y2 and Y3 such that 80% of all repository traffic was referred by either Google or GS. The remaining 20% comprised a long tail of services. The nature of this traffic growth can perhaps be better observed in Figure 4(c) when the data for Figure (a) and (b) are overlaid, with GS demonstrating steeper growth relative to other traffic.

Table 1 summarises the top ten referral sources (with local sources excluded). A 29% decrease in Bing referrals between Y2 and Y3 is noted, as is a larger decrease for Yahoo! (which shares the Bing index). Reasons for this are suggested later in this paper but essentially relate to search interference arising from the institutional CRIS. However, given the overall small contribution to traffic made by Bing and Yahoo! – and the far larger increases in referral traffic from other sources (including within the long tail) – this decrease is more than cancelled out. An interesting observation relates to the increase in referrals from social sources, such as Twitter and Facebook. Again, this traffic remains small in relation to the volume of total traffic but extraordinary percentage increases can nevertheless be observed. For example, traffic from Twitter increased by 3700% between Y1 and Y3 as improved social media interaction opportunities were implemented.

Referral source	Y1	Y2	Y3
Google	152890	185491	251705
Google Scholar	57319	83045	111563
Bing	10794	10411	7405
Twitter	173	1414	6556
Android Google Search	0	0	2274
Baidu	3234	2657	2209
Glgoo	878	1048	2077
Yahoo	3628	1351	1436
Facebook	533	634	1108
Ebsco (EDS)	482	433	485

Table 1. Summarised web traffic referral sources as measured by GA with local sources excluded.

Discovery

A more appropriate measure of repository discoverability lies in search metrics. Google Search Console was used to gather search data during the reporting period, thereby allowing the effect of the repository changes to be examined on Google search queries. Search Console makes the distinction between data pertaining to "impressions" and "clicks". Impressions are defined as occurring when "A link URL record […] appears in a search result for a user", while a click is "any click that sends the user to a page outside of Google Search" [25].

Improvements in impressions and clicks were observed in Y2 at 52% (n = 5,795,781) and 23% (n = 169,720) respectively when compared to the Y1 period. This upwards trend continued in Y3 at 61% (n = 9,357,582) and 25% (n = 212,148), and a general upwards trend in impressions and clicks can be observed in the graph profile of Figure 5, with impressions and clicks demonstrating particular growth from early 2017 onwards. The total percentage growth in impressions and clicks during the entire reporting period was 146% and 53% respectively. Figure 6 provides a summary of the increase in clicks, impressions and COUNTER usage, with steeper increases in impressions and clicks noted between Y2 and Y3.

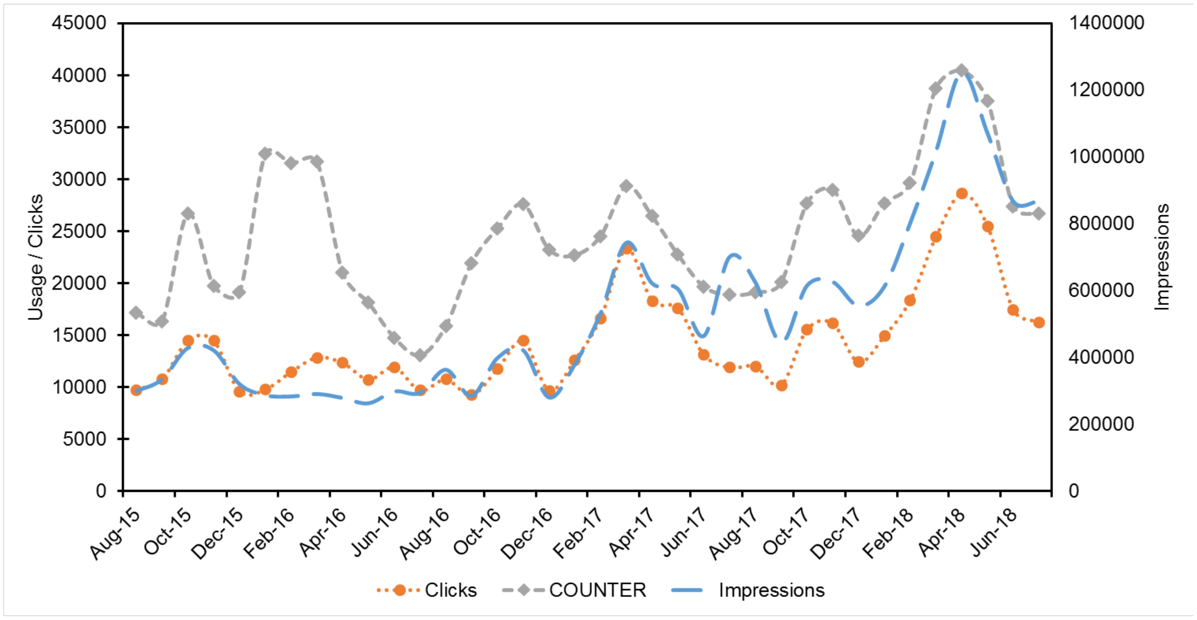
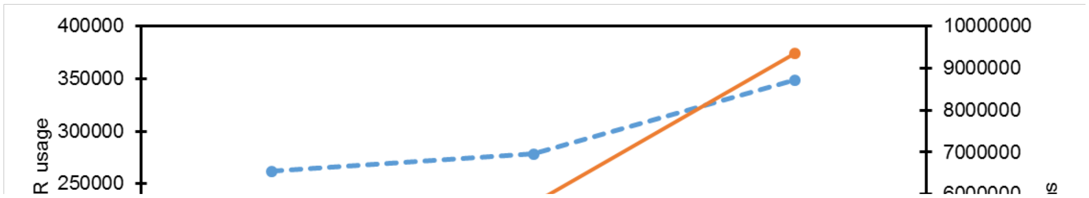


Figure 5. Strathprints COUNTER usage during Y1, 2 & 3 and Google clicks & impressions during the same period.



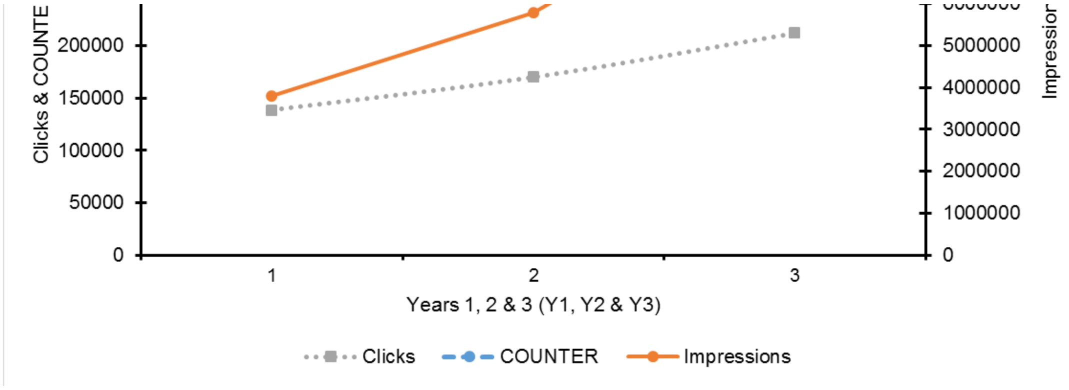


Figure 6. Charted data on observed clicks, impressions & COUNTER usage during reporting period.

During the full period examined (i.e. Y1-Y3), Strathprints demonstrated a 33% growth in COUNTER compliant usage. This growth in usage was observed despite only a 19% growth in full-text deposits during the same period. The pattern of this usage appears more nuanced when considered on an annual basis. For example, Y2 and Y3 observed a 6% and 25% increase respectively in COUNTER usage, with the number of deposits in Y3 actually declining by 19% while increasing by a similar proportion in Y2. Usage therefore generally increased greater than the number of deposits but in the first year this was not observed, possibly owing to the latency of search tool indexes during Y1. It is also noteworthy recalling that Google search referrals and GS traffic demonstrated growth well in excess of the 19% full-text deposit rate, as per Figure 4. In other words, the percentage of users being referred increased at a greater rate than the percentage growth rate of full-text.

To determine whether a correlation between clicks and COUNTER usage was present, Pearson's correlation coefficient was calculated for each year in the reporting period. Pearson's correlation coefficient provides a measure of the linear correlation between two variables by using a value between -1 and 1 to denote the strength of correlation. It can be reported that a correlation was detected, ranging from a weak relationship in Y1 ($r = 0.26$) to a moderate positive correlation in Y2 ($r = 0.65$). For those readers familiar with statistics, this correlation was confirmed via the t statistic ($t = 2.68$, $df = 11$, $p < 0.05$). A strengthening of the positive correlation was further observed in Y3 ($r = 0.97$), also confirmed by the t statistic and a higher level of statistical significance ($t = 12.72$, $df = 11$, $p < 0.001$).

Computing the coefficient of determination (r2) allows us to better understand the proportion of the variance in the dependent variable (i.e. COUNTER usage) which is predictable from the independent variable (i.e. Google clicks). Computing the coefficient of determination revealed data to be more nuanced (Figures 7, 8 & 9). r2 was stronger in Y2 (r2 = 0.419) than Y1 (r2 = 0.069); clearly a significantly higher value but indicating that only circa 42% of the unique variance in COUNTER usage could be directly attributed to Google clicks. However, this variance narrowed considerably for Y3 (r2 = 0.934) with a strong linear relationship between variables noted such that 94% of the unique variance in COUNTER usage could be directly attributed to Google clicks. This narrowing in variation can also be observed from Figure 5, with data points grouping more closely to the regression line.

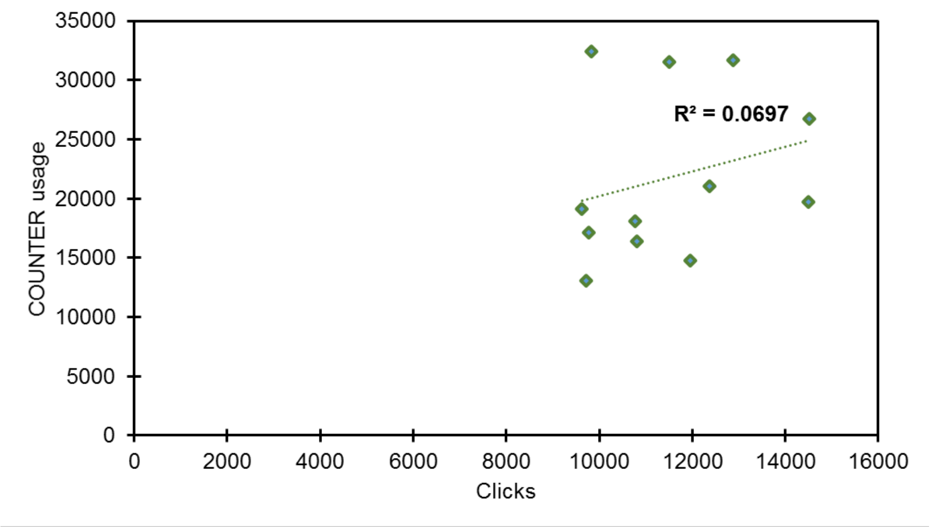
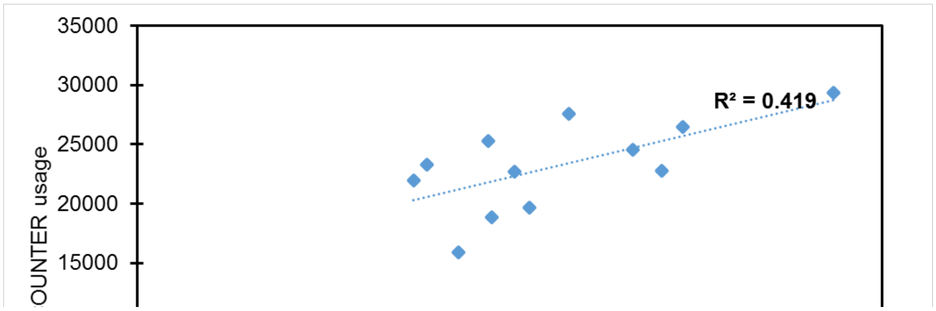


Figure 7. Coefficient of determination (r squared) for Y1 (clicks and COUNTER usage).



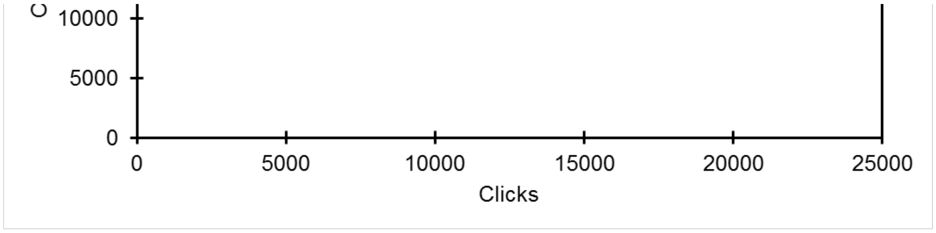


Figure 8. Coefficient of determination (r squared) for Y2 (clicks and COUNTER usage).

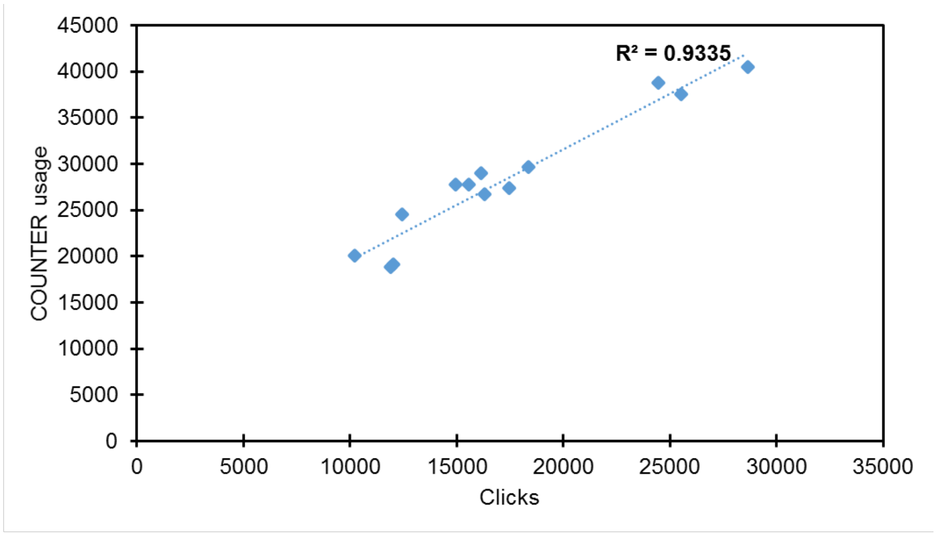


Figure 9. Coefficient of determination (r squared) for Y3 (clicks and COUNTER usage).

By exposing their content to disparate search services, and the nature of repository content itself, repositories encourage – and are conducive to – "horizontal" information seeking strategies [19]. These types of information seeking strategy typically correspond with the relatively high "bounce rates" that repositories experience. Bounce rates are calculated by GA as "a session that triggers only a single request to the Analytics server, such as when a user opens a single page on your site and then exits without triggering any other requests […] during that session" [26]. The bounce rate in this study remained relatively unchanged, fluctuating across reporting periods at circa 75%. However, the average time users spent on Strathprints upon arrival increased, up from 01:13 in Y1 to 01:54 in Y2 and then 01:59 in Y3. Although users were continuing to bounce, they were typically spending longer on Strathprints, indicative perhaps that improvements to the UI and Strathprints functionality was enough to persuade users to defer their bounce. In other words, it was possible to improve users' "dwell time" on Strathprints by 61% between Y1 and Y3.

Dwell time therefore suggests itself as a more accurate indicator of repository engagement than bounce rates, which experienced only marginal change during the reporting period. Bounce rates are not necessarily a reliable metric within models of information seeking behaviour. For example, a user might spend 25 mins reading content on a repository, taking notes and chaining references, but then they might leave. In this example the user "bounced" because they failed to navigate to another page on the repository. But, in repository terms, the user spent 25 mins consuming repository content and found that content sufficiently useful that they "dwelled" for 25 mins. Dwell time is therefore critical to understanding repository engagement. Interestingly, it is for this reason that many search services, Google and Bing included, factor "dwell time" into their relevance rankings [27], [28]. Like PageRank more generally, the way in which search tools calculate dwell time, or the weighting it is assigned in computing algorithms such as PageRank, is unknown; but it is clearly a variable in calculating relevance and is therefore a metric institutions and repository managers should monitor. Similarly, the significance of dwell time in this evaluation is impossible to calculate. It is only possible to state that it would have positively influenced the visibility of Strathprints in the search results of services such as Google and Bing.

Conclusion and future work

In this contribution we experimented with a suite of repository adjustments and improvements performed on an EPrints powered repository. These adjustments were designed to support improved repository web visibility and user engagement thereby improving usage and should be considered within the wider context of the COAR Next Generation Repositories agenda. The evaluation provides persuasive evidence that specific enhancements to technical aspects of a repository can result in significant improvements to repository visibility, resulting in a greater web impact and consequent increases in content usage. The results suggest that both web and search traffic and COUNTER usage can be significantly improved on the most important search and discovery tools, with strong correlations between Google search visibility and repository COUNTER usage demonstrated and variation narrowing particularly in Y3. 94% of the unique variance in COUNTER usage was found to be directly attributed to Google clicks. Strathprints also demonstrated a 33% increase in COUNTER compliant usage during the years examined. Across the entire reporting period total traffic to Strathprints grew by 63%, with Google impressions and clicks increasing by 146% and 53% respectively. GS traffic was also found to have generated a traffic growth 99%, accounting for 25% of all web traffic to Strathprints in Y3. User dwell time was also found to have increased, suggesting longer interaction sessions by users.

Of course, as with any experiments attempting to effect change on third party systems, it is impossible to control for all variables hypothesised to influence web visibility. It is not claimed that every known variable has been addressed in this instance. The approach adopted here of delivering repository adjustments and improvements was a holistic one, and was intended to address as many as possible. The approach could therefore be described as pursuing the accumulation of marginal gains; identifying numerous minor optimisations that can be implemented which, when taken in aggregate, effect further significant improvements. There are also limitations to be noted on the use of search data from Google Search Console which, for obvious reasons, provides data on Google searches only. However, as the majority of referral traffic to Strathprints comes via Google this seemed an acceptable compromise to be made in this instance. Future similar studies should nevertheless explore additional sources of search data to improve the accuracy of conclusions drawn, especially as Google cannot be relied upon to be the preeminent web search engine indefinitely. We intend to continue monitoring our data into Y4 with the hope of exploring how additional adjustments could improve visibility on other search discovery tools, thereby providing the basis for greater longitudinal analysis.

Although the experiments were performed on EPrints it is thought that most of the adopted improvements are equally applicable to most repository platforms. There is, in fact, potential for others to improve the impact of the approach. For example, it was noted in the literature that coversheets are considered to be disruptive to the bot crawling potential of repositories and it has been suggested that repositories disable such repository functionality [29]. Based on local experimentation and the need to ensure accurate attribution of repository outputs, coversheets remained enabled in Strathprints and continue to remain enabled. This therefore highlights a possible limitation. However, there are also potential additional improvements to be gained by other repositories willing to develop their own alternative approaches (e.g. watermarking attribution details) or disabling coversheets altogether. Furthermore, owing to the existence of Strathprints within a connected CRIS configuration, the present author noted issues of the CRIS front-end interfering with the visibility of Strathprints in some cases. Again, this interference was almost impossible to quantify and appeared to particularly affect Bing and Yahoo! Searches; but for those repositories operating outside of a CRIS environment or functioning as the de facto CRIS front-end, considerable additional opportunities are available vis-à-vis promoting the discoverability and web impact of repository content.

References

[1] IRUS-UK [Internet]. 2018 [cited 2018 Jun 30]. Available from: http://www.irus.mimas.ac.uk/

[2] Kelly B, Nixon W. SEO analysis of institutional repositories: What's the back story? In: Open Repositories 2013 [Internet]. University of Bath; 2013 [cited 2017 Jul 19]. Available from: http://opus.bath.ac.uk/35871/

[3] Macgregor G. The long read: Why do institutional repositories remain one of the only viable options for Green Open Access? [Internet]. Open Access @ Strathclyde. 2016 [cited 2017 Jun 29]. Available from: https://perma.cc/G52J-2FSG

[4] Macgregor G. Reviewing repository discoverability?: approaches to improving repository visibility and web impact. In: Repository Fringe 2017 [Internet]. John McIntyre Conference Centre, University of Edinburgh; 2017 [cited 2018 Aug 3]. Available from: https://strathprints.strath.ac.uk/61333/

[5] Macgregor G. Reviewing repository discoverability with Strathprints [Internet]. Open Access @ Strathclyde. 2017 [cited 2018 Aug 29]. Available from: https://perma.cc/A3R9-W2JV

[6] Tmava AM, Alemneh DG. Enhancing Content Visibility in Institutional Repositories: Overview of Factors that Affect Digital Resources Discoverability [Poster] [Internet]. iConference, 2013, Fort Worth, Texas, United States. 2013 [cited 2018 Aug 13]. Available from: https://digital.library.unt.edu/ark:/67531/metadc146593/

[7] Pekala S. Microdata in the IR: A Low-Barrier Approach to Enhancing Discovery of Institutional Repository Materials in Google. Code4Lib Journal [Internet]. 2018 Feb 5 [cited 2018 Aug 13];(39). Available from: https://journal.code4lib.org/articles/13191

[8] COAR. Next Generation Repositories: Behaviours and Technical Recommendations of the COAR Next Generation Repositories Working Group [Internet]. Göttingen: COAR; 2017 Nov. Available from: https://www.coar-repositories.org/files/NGR-Final-Formatted-Report-cc.pdf

[9] Tonkin EL, Taylor S, Tourte GJL. Cover sheets considered harmful. Information Services & Use [Internet]. 2013 Jan 1 [cited 2018 Aug 29];33(2):129–37. Available from: https://doi.org/10.3233/ISU-130705

[10] Acharya A. Indexing repositories: pitfalls and best practices [Internet]. Proceedings of Open Repositories 2015. 2015. Available from: http://purl.dlib.indiana.edu/iudl/media/6537033b6s

[11] Shirlaw D. University of Strathclyde research rankings rocket [Internet]. Glasgow City of Science and Innovation – News. 2014 [cited 2018 Aug 14]. Available from: https://perma.cc/9CNK-8Z53

[12] Zhang T, Maron D, Charles C. Usability evaluation of a research repository and collaboration website. Journal of Web Librarianship [Internet]. 2013 Jan 1; Available from: http://docs.lib.purdue.edu/lib_fsdocs/51

[13] McKay D, Burriss S. Improving the Usability of Novel Web Software: An Industrial Case Study of an Institutional Repository. In: Web Information Systems Engineering – WISE 2008 Workshops [Internet]. Springer, Berlin, Heidelberg; 2008 [cited 2017 Jul 18]. p. 102–11. (Lecture Notes in Computer Science). Available from: https://link.springer.com/chapter/10.1007/978-3-540-85200-1_12

[14] Wang J, Senecal S. Measuring Perceived Website Usability. Journal of Internet Commerce [Internet]. 2007 Aug 8;6(4):97–112. Available from: https://doi.org/10.1080/15332860802086318

[15] Pendell KD, Bowman MS. Usability Study of a Library's Mobile Website: An Example from Portland State University. Information Technology and Libraries [Internet]. 2012 Jun 12 [cited 2018 Aug 3];31(2):45–62. Available from: https://ejournals.bc.edu/ojs/index.php/ital/article/view/1913

[16] Everard A, McCoy S. Effect of Presentation Flaw Attribution on Website Quality, Trust, and Abandonment. Australasian Journal of Information Systems [Internet]. 2010 Mar 1 [cited 2018 Aug 3];16(2). Available from: http://journal.acs.org.au/index.php/ajis/article/view/516

[17] Nielsen J, Molich R. Heuristic Evaluation of User Interfaces. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems [Internet]. New York, NY, USA: ACM; 1990. p. 249–256. (CHI '90). Available from: http://doi.acm.org/10.1145/97243.97281

[18] Google. Search Engine Optimization (SEO) Starter Guide [Internet]. 2018 [cited 2018 Aug 3]. Available from: https://perma.cc/8CT3-UAV5

[19] Kloboves K. Continuing to make the web more mobile friendly [Internet]. Official Google Webmaster Central Blog. 2016 [cited 2018 Jul 19]. Available from: https://webmasters.googleblog.com/2016/03/continuing-to-make-web-more-mobile.html

[20] Moreno L, Martínez P. Overlapping factors in search engine optimization and web accessibility. 2013 Jun [cited 2018 Aug 3]; Available from: https://e-archivo.uc3m.es/handle/10016/20175

[21] Macgregor G. Feeding the beast: workloads in a hybrid IR / CRIS environment [Internet]. Open Access @ Strathclyde. 2017 [cited 2017 Jul 19]. Available from: https://perma.cc/DL7U-9VCE

[22] Wang Z, Phan D. Using page speed in mobile search ranking [Internet]. Official Google Webmaster Central Blog. 2018 [cited 2018 Aug 3]. Available from: https://perma.cc/8QKP-NE5S

[23] Macgregor G. Demonstrating the need for speed: improving page loading and rendering in repositories [Internet]. Open Access @ Strathclyde. 2017 [cited 2018 Aug 3]. Available from: https://perma.cc/DCM7-TS7B

[24] Google. About Data Highlighter [Internet]. 2018 [cited 2018 Aug 24]. Available from: https://perma.cc/92DY-MFZP

[25] Google. Google Search Console [Internet]. 2018 [cited 2018 Aug 10]. Available from: https://www.google.com/webmasters/tools/home

[26] Google. Google Analytics [Internet]. 2018 [cited 2018 Aug 13]. Available from: https://marketingplatform.google.com/about/analytics/

[27] Microsoft. How To Build Quality Content [Internet]. 2011 [cited 2018 Aug 24]. Available from: https://perma.cc/X2U7-QMPJ

[28] Shewan D. Dwell Time: The Most Important Metric You're Not Measuring [Internet]. 2017 [cited 2018 Aug 24]. Available from: https://perma.cc/5H5E-BTE2

[29] Tonkin E, Taylor S, Tourte G, web-support@bath.ac.uk. Cover sheets considered harmful. In: 17th International Conference on Electronic Publishing [Internet]. Blekinge: University of Bath; 2013 [cited 2015 Sep 18]. Available from: http://www.bth.se/com/elpub2013.nsf/

Data statement

Data underpinning this work are available under a CC-BY license at: https://doi.org/10.5281/zenodo.1411207

About the author

George Macgregor (g3om4c@gmail.com) is the Institutional Repository Manager at the University of Strathclyde in Glasgow, Scotland (UK). George's interests are in structured open data (esp. within Semantic Web and repository contexts), information retrieval, distributed digital repositories and human-computer interaction (HCI).

Web: https://purl.org/g3om4c

ORCID: https://orcid.org/0000-0002-8482-3973

Subscribe to comments: For this article | For all articles

publications

Article

Enhancing Content Discovery of Open Repositories: An Analytics-Based Evaluation of Repository Optimizations

George Macgregor [1,2]

[1] Scholarly Publications and Research Data, IS Information Management, University of Strathclyde, Curran Building, 101 St James Road, Glasgow G4 0NS, UK; george.macgregor@strath.ac.uk

[2] iSchool, Department of Computer and Information Sciences, University of Strathclyde, Glasgow G4 0NS, UK

Received: 25 November 2019; Accepted: 22 January 2020; Published: 6 February 2020

Abstract: Ensuring open repositories fulfil the discovery needs of both human and machine users is of growing importance and essential to validate the continued relevance of open repositories to users, and as nodes within open scholarly communication infrastructure. Following positive preliminary results reported elsewhere, this submission analyses the longer-term impact of a series of discovery optimization approaches deployed on an open repository. These approaches were designed to enhance content discovery and user engagement, thereby improving content usage. Using Strathprints, the University of Strathclyde repository as a case study, this article will briefly review the techniques and technical changes implemented and evaluate the impact of these changes by studying analytics relating to web impact, COUNTER usage and web traffic over a 4-year period. The principal contribution of the article is to report on the insights this longitudinal dataset provides about repository visibility and discoverability, and to deliver robust conclusions which can inform similar strategies at other institutions. Analysis of the unique longitudinal dataset provides persuasive evidence that specific enhancements to the technical configuration of a repository can generate substantial improvements in its content discovery potential and ergo its content usage, especially over several years. In this case study, COUNTER usage grew by 62%. Increases in Google 'impressions' (266%) and 'clicks' (104%) were a notable finding too, with high levels of statistical significance found in the correlation between clicks and usage ($t = 14.30, df = 11, p < 0.0005$). Web traffic to Strathprints from Google and Google Scholar (GS) was found to increase significantly with growth on some metrics exceeding 1300%. Although some of these results warrant further research, the article nevertheless demonstrates the link between repository optimization and the need for open repositories to assume a proactive development path, especially one that prioritises web impact and discovery.

Dataset: Data supporting this work are available under a Creative Commons Attribution (CC-BY) license at: https://doi.org/10.5281/zenodo.3146553.

Keywords: institutional repositories; open repositories; resource discovery; Open Access; content visibility; repository optimization; search engine optimization; information retrieval; open science; web traffic

1. Introduction

Institutional and subject-based repositories have become essential nodes within global open scholarly communications infrastructure [1]. Such repositories typically deliver a set of services to academic or disciplinary communities to ensure that digital content generated by community members is assured long-term management and dissemination [2]. The content dissemination potential of repositories is well noted and remains a core motivation of open science movement [3].

More than ever, users of repository content expect to discover open content easily, normally via search, and for their own content (typically scholarly content deposited in an open repository) to be equally discoverable. Repositories are, and have been, well placed to meet these needs but cannot remain static, isolated systems, removed from the changing technical expectations of discovery tools. This article contributes to the discussion surrounding user discovery needs and provides evidence that content discovery requires prioritization.

Better meeting user expectations is crucial to preserving the relevance of repositories as nodes within open science infrastructure. The emergence of proprietary scholarly communications platforms represents a significant future challenge for open repositories. Such platforms are increasingly demonstrating popularity within research institutions yet simultaneously often demonstrate poor support for open standards or prevalent open science technical protocols. Low levels of integration with existing open scholarly infrastructure is also recognised to be a frequent challenge [4–6]. Ensuring that repositories can continue to expose content as optimally as possible to search and discovery agents, and in a manner superior to alternative platforms, is therefore a key tenet of repositories and central to their relevance to users. Understanding the way in which this can be technically achieved is important too; COAR's conceptions of Next Generation Repositories [7] has delivered an important development path for repositories to follow in coming years. This includes the promotion of repository 'behaviours' upon which functionality supporting better content discovery can be built, but also better support for social networking integrations and peer review or annotation within the global repository network. However, the need to gather evolving evidence on visibility and discovery remains a necessity to direct new or unexpected streams of technical work or to steer institutional decision making in instances where HEIs are confronted with choices about selecting or migrating scholarly communications platforms.

Using Strathprints, the University of Strathclyde repository as a case study, this article uses analytics on web impact, COUNTER usage, web traffic and other indicators over a 4-year time-frame. The principal contribution, described in Sections 4 and 5, is to report on the insights this longitudinal dataset yields about repository visibility and discoverability, and to deliver robust conclusions which can then inform similar strategies at other institutions. The data presented were captured following the embedding of several technical adjustments and enhancements to Strathprints, which have been documented in more detail in previous work [8], and are especially relevant to both institutional and subject-based repositories. These adjustments and technical enhancements are reviewed in Section 3, within which the methodology is detailed. Data are described in Section 4 as is its collection and analysis. Related work is considered in the following section.

2. Repository Visibility and Discovery: Related Work

Previous work has noted the importance of repositories in promoting open scholarly communication and the discovery of open research content, e.g., [9–12]. The importance of repository visibility as a precursor to the discovery of this content has been addressed by the work of Aguillo [13], especially via the 'Ranking Web of Repositories' which attempts to monitor and rank repositories according to their visibility in Google Scholar indexes [14]. Such is the importance of visibility in generating repository discovery and eventual impact that the concept has been enshrined in the German DINI Certificate, which promotes best practice in standardization, interoperabity and service quality as a means of achieving superior repository visibility [15]. However, translating this visibility into content discovery remains a less understood area of research.

Arlitsch [16] provides a useful contribution on the role of search engine optimization (SEO), the importance of 'white hat' adjustments and its role in promoting repository indexing by common search engines, as well as academically focused discovery tools like Google Scholar. Related works by Askey and Arlitsch [17] have reported on the growing importance of white hat changes, such as a migration to HTTPS, as a contributory factor in ranking repository content within Google's PageRank, with SEO toolkits also developed to support digital library service administrators in 'getting found' [18].

It is noteworthy that the same group of researchers are responsible for the Repository Analytics and Metrics Portal (RAMP) which seeks to improve the quality of usage and traffic analytics [19].

Contributions have also emerged from individuals closer to the systems which refer much of the web traffic repositories seek, such as Google Scholar. Acharya [20], for example, delivers recommendations on repository optimization from a position of authority, noting how common technical failings inhibit satisfactory Google Scholar crawling and indexing. Acharya highlights various optimizations which can be performed on repositories in order to ensure improved Google Scholar crawling and indexing penetration. One of Acharya's recommendations pertains to 'coversheets', the influence of which Tonkin et al. [21] explore in their survey of coversheet usage within the UK. Coversheets are additional pages which are typically prepended to the first page of any document served by a repository. A coversheet typically provides further information about the nature of the item downloaded, such as attribution information, full bibliographic reference details, copyright statement, and so forth. Acharaya [20] has noted that coversheets can disrupt the automated metadata extraction techniques used by Google Scholar as their technique is based on interpreting the first page of academic documents, most of which tend to follow a typical format and layout. Tonkin et al.'s analysis of the literature concluded that coversheets should be avoided as they can impede discovery but, lacking any supporting evaluative data, they acknowledge that local decision making on the part of repository administrators and developers is necessary. Suffice to state that the negative crawling issues arising from coversheet use in repositories is an issue highlighted more recently by Acharya [20], thereby supporting Tonkin et al.'s recommendations.

Despite these contributions to the literature, and despite the importance of repositories and their infrastructure in exposing open research content, wider understanding about repository visibility and discoverability remains embryonic. Few studies have sought to codify and then evaluate the impact of their approaches and many restrict their analyses to anecdotal observations surrounding the logical visibility benefits native to the majority of repository platforms. Recent related work by the present author has gone some way to addressing this by studying and codifying specific technical adjustments and improvements which can be made to an open repository, followed by the observation of longitudinal web analytics and usage data in order to assess the efficacy of these changes [8]. The emergence of COUNTER-compliant repository usage statistics has been an important development in this regard by providing a new, additional source of reliable usage statistics.

The COUNTER Code of Practice establishes open international standards and protocols for the provision of service-generated online usage statistics, specifically for digital resource usage [22]. This ensures consistent counting and processing of usage, including control for the interpretation of robot visits, unusual usage patterns, etc. By specifying what constitutes 'usage', COUNTER enables disparate services to supply data which are directly comparable. Interestingly, recent work by Wood-Doughty et al. [23] has identified anomalies in usage data reported by commercial publishers (a so-called 'publisher effect'), suggesting a degree of divergence in the implementation of COUNTER where multiple agencies are responsible for reporting the statistics. However, the emergence of COUNTER-compliant usage data available from the UK national repository usage aggregation service, IRUS-UK, has provided new opportunities for understanding the nature of repository discoverability. By aggregating usage data for circa 200 repositories according to COUNTER, IRUS-UK provides a degree of authoritativeness in the figures it reports [24]. This provides repositories with comparable, authoritative, standards-based data and facilitates the profiling and benchmarking of repositories.

Preliminary experiments documented in [25] noted some encouraging evidence about the positive impact of certain repository enhancements, making use of IRUS-UK data, but the small nature of the study and dataset provided only indicative results. Results from a subsequent and more detailed study from the same stream of work [8] concluded that web traffic, search traffic and COUNTER usage could be improved on the most important search and discovery tools by deploying the specified technical changes. Strong correlations between Google search visibility and repository COUNTER usage were demonstrated, as were significant increases in web traffic, Google 'impressions' and 'clicks' and COUNTER usage.

3. Methodology

This article seeks to continue the aforementioned line of enquiry by validating the results reported in [8] through examination of a larger web impact and COUNTER usage dataset. This larger dataset encompasses a longer temporal period thereby compensating for the limited number of data points used in the aforementioned work. The dataset for this current article, described in detail within Section 4, captures data over a four-year period instead of three or two years, as in the less exhaustive studies. Analyses performed on such a large dataset better delivers reliable and actionable conclusions which can then inform repository discovery strategies elsewhere. The case study repository for this article, Strathprints[1], the University of Strathclyde institutional repository, is powered by EPrints (version 3.3.13). Though EPrints is the focus here, it is thought that most of the adopted technical changes are equally applicable to other repository platforms.

3.1. Implemented Repository Changes

Prominent repository platforms (e.g., EPrints, DSpace, Digital Commons, OJS, etc.) continue to demonstrate out-of-the-box support for discovery and interoperability with key academic tools, e.g., Google Scholar (GS), scholarly aggregators like CORE and BASE, etc. However, there nevertheless remains wide variation in the relative visibility and discoverability of repository content, even across similar or the same repository platforms, such that it is necessary to take steps towards repository optimization. To effect change in web visibility and user engagement, thereby improving usage, a series of technical 'improvements' and 'adjustments' were implemented on Strathprints in March 2016.

'Improvements' were changes that resulted in substantive modifications to repository functionality, while 'adjustments' included actions that sought to refine existing aspects of the repository. As this article is largely concerned with the effect of the technical changes and the resulting data, the nature of the adjustments and improvements are only summarised in Table 1 to provide context. Full details, including the motivation behind these changes, are instead available from [8]. Suffice to state that few of either the improvements or adjustments were onerous to implement and most are feasible to action by repository development managers. This is largely because the most significant pertain to the repository front-end thus making any serious software re-engineering unnecessary. For example, adherence to site speed best practice, such as asynchronous loading of resources, CSS and Javascript minification, GZIP compression, etc., all of which have become important signals for Google [26]. Similarly, ensuring a positive mobile experience for users has become a signal in PageRank, with a heavier weighting assigned in recent search engine updates [27,28].

[1] Strathprints: https//:strathprints.strath.ac.uk/

Table 1. Summary of technical 'adjustments' and 'improvements' implemented on Strathprints. Full details in [8].

Key Technical 'Adjustments'
Modification of file-naming conventions
'Minification' of all relevant repository source files
Rationalisation of all CSS and JavaScript (JS) files in order to remove unused rules and variables
Asynchronous loading of JS resources
Deployment of GZIP compression
Image optimization, e.g., compression, use of .webp, etc.
Migration to InnoDB as the MySQL storage engine
Deployment of Google Data Highlighter
Key Technical 'Improvements'
Repository user interface (UI) improvements
'Mobile first', responsive re-engineering of repository to align with new weighting in PageRank, etc.
'White hat' improvements, e.g., navigation, hyperlink labels, content improvements promoting user interaction
'Connector-lite' ecosystem implemented within repository-CRIS interactions

3.2. Data Collection

A variety of metrics were monitored in order to measure the influence of the technical 'adjustments' and 'improvements' to Strathprints, including search traffic data from Google Search Console[2], COUNTER compliant usage data from IRUS-UK[3], Google Analytics[4] (GA) tracking data and routine statistical data from Strathprints itself.

Search metrics offer an appropriate measure of repository content discoverability. Google Search Console was therefore used to capture search data during the reporting period, thereby enabling the effect of the technical adjustments and improvements to be explored on Google search queries. The distinction between 'impressions' and 'clicks' is recognised by Search Console and is reflected in its search data. Impressions are stated as arising when "A link to a URL record ... appears in a search result for a user", while a click is "any click that sends the user to a page outside of Google Search" [29]. Data pertaining to clicks and impressions were extracted from Search Console and compiled in a .csv file.

Reporting from Google Analytics can provide rich data on web traffic and its sources. Site content behaviour reports were generated for the relevant periods within Google Analytics, with 'acquisition' used as a secondary dimension to capture 'source', thereby providing data on traffic referral sources as well as typical data on number of page views, unique page views, page path and so forth. Data were exported to .csv for further analysis.

COUNTER compliant usage data from IRUS-UK was generated via an 'item report 1'. The item report 1 provides details of the number of successful item download requests by month and by repository identifier. Data relating to item URL, title, author(s), item type and total downloads by month and in total for the period selected are included in this report and were also exported to a .csv file.

All data were captured for the year up to March 2016, representing Year 1 (Y1 = 2015/2016). This ensured a data baseline for repository web impact prior to the implementation of the technical changes. Data were then monitored for the same periods during Year 2 (Y2 = 2016/2017), Year 3 (Y2 = 2017/2018) and Year 4 (Y4 = 2018/2019), with data collection ending on 31 March 2019.

The usage of repositories can be cyclical in nature, with usage reflecting the periods when researchers and students tend to be busiest. It is therefore typical to observe increases in usage during

[2] Google Search Console: https://www.google.com/webmasters/tools/home
[3] IRUS-UK: https://irus.jisc.ac.uk/
[4] Google Analytics: https://analytics.google.com/

academic semesters and declines during summer, spring and winter vacation periods. Data from the present author's prior work used data that followed these cyclical patterns. Increases in usage, followed immediately by near commensurate declines, can be observed from the chart in Figure 1, which displays the total usage of 88 IRUS-UK member repositories between August 2016 and July 2019. For this reason the analysis in this current article employs an alternative temporal segmentation thereby controlling for any data variation potentially arising from these established usage patterns. Altering the segmentation controls for any confirmation bias emerging from prior analyses and better tests whether observations in these prior analyses hold true when usage periods are modified. For example, in this instance the year up to March 2016 is examined, and the same period in each subsequent year. Related prior work instead analysed data based on a typical academic calendar year (years up to end July) [8] and years up to end June [25]. Examining an alternative temporal segmentation may limit direct comparisons with specific data points within prior analyses but is a justifiable compromise to ensure effects are observable where data are segmented differently.

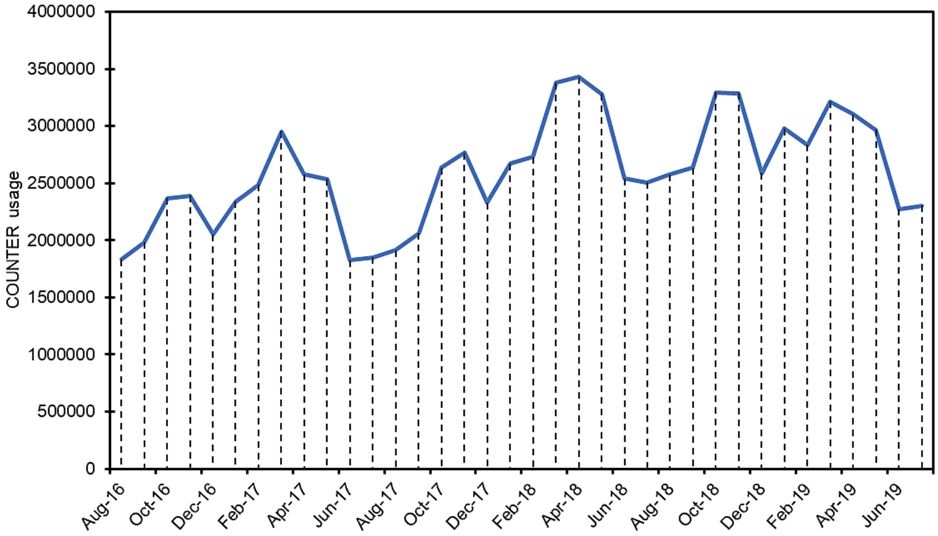


Figure 1. COUNTER-compliant usage data for 88 IRUS-UK member repositories 08/2016–07/2019.

4. Results

4.1. Analytics

Measurement of web traffic and unique web traffic was performed using Google Analytics (GA). Data are set out in Table 2.

Traffic in Y2 increased by 68,824 to 365,024, equating to a 23% improvement when compared to Y1. A 22% improvement in unique traffic was also observed (*n* = 276,042). Y3 also yielded a 23% increase in traffic on Y2 (*n* = 450,520), with percentage growth in unique traffic equivalent to 26% (*n* = 346,851). The increase in traffic and unique traffic for Y4 was lower than Y3 at 9% and 10% respectively.

These increases in traffic initially appear to be lower than those reported previously [8] which, for example, reported a Y2 traffic increase of 54%, from 150,408 to 428,407, considerably higher than the 23% improvement reported here. Similar disparities can be observed for Y3 data too. However, it should be noted that the alternative segmentation of annual web impact data have altered the spread of traffic data across years, making direct comparisons to previous results problematic. Indeed, while [8] reported a plateauing of traffic (6%) and unique traffic (8%) in Y3, this article instead reports a considerable percentage increase at 23% and 26% for Y3, with plateauing of traffic (9%) and unique traffic (11%) observed in Y4. This means that total percentage growth during the entire reporting period of this present study was more significant, at 65% and 69% for traffic and unique traffic respectively.

This actually exceeds previously reported results but highlights the difficulties which can arise from studying different 'annual segments' of data.

Google was again found to be the single largest referral source during the reporting period, accounting for 56% of all repository traffic in Y4. Over the entire reporting period this referral traffic (including unique traffic) increased by circa 1500% (Table 2). The most significant referral source thereafter was found to be Google Scholar (GS), equivalent to 26% of all web traffic by Y4 and growing by 1920% during the entire reporting period (Table 2). Much of this massive percentage growth can be observed in Y2, owing to a low baseline in GS traffic during Y1 but with significant increases observed in Y3 and Y4 also.

To verify the influence of outlying data points it is worthwhile briefly reviewing the extent of data variability using some common measures of central tendency. Table 3 sets out measures[5] for the total traffic data detailed above in Table 2 ('Current data—A') alongside the same measures for data reported in previous work [8], labelled in Table 3 as 'Prior data—B'. Data used for 'Prior data—B' are publicly available [30].

Table 2. Data table of total and unique web traffic to Strathprints during Y1–Y4, alongside total and unique traffic referred via Google and Google Scholar (GS).

	Total	Unique	Google	Unique Google	GS	Unique GS
Y1	296,200	226,791	17,436	13,274	6208	4827
Y2	365,024	276,042	164,550	130,565	72,179	55,294
Y3	450,520	346,851	230,953	182,227	104,051	80,786
Y4	489,140	383,117	274,983	217,826	125,405	94,305
Total Y1–Y4	1,600,884	1,232,801	687,922	543,892	307,843	235,212
% growth (Y2)	23.24	21.72	843.74	883.61	1062.68	1045.51
% growth (Y3)	23.42	25.65	40.35	39.57	44.16	46.1
% growth (Y4)	8.57	10.46	19.06	19.54	20.52	16.73
% growth (Exc. Y1)	34	38.79	73.74	70.55	67.11	66.83
Total % growth (Y1–Y4)	65.14	68.93	1477.1	1541	1920.05	1853.7

Table 3. Measures of central tendency for total and unique web traffic to Strathprints during Y1–Y4 ('Current data—A'), alongside total and unique traffic referred via Google and Google Scholar (GS). Data also include measures for 'Prior data—B' using data reported in [8] for comparison. Bottom row, 'Current data—A*', are 'Current data—A' data excluding outlying Y1 data.

Current Data—A	**Total**	**Unique**	**GS**	**Unique GS**	**Google**	**Unique Google**
Mean (*M*)	400,221	308,200.3	76,960.75	58,803	171,980.5	135,973
Standard deviation (*SD*)	86,594.41	70,161.76	51,992.13	39,451.94	112,585.5	89,300.31
Prior Data—B	**Total**	**Unique**	**GS**	**Unique GS**	**Google**	**Unique Google**
Mean (*M*)	386,908	296,311	83,569.33	63,691.33	196,783.67	154,834.67
Standard deviation (*SD*)	95,203.59	73,250.7	27,735.22	22,046.71	50,429.38	38,672.46
Current Data—A*	**Total**	**Unique**	**GS**	**Unique GS**	**Google**	**Unique Google**
Mean (*M*)	434,894.67	335,336.67	100,545	76,795	223,495.33	176,872.67
Standard deviation (*SD*)	63,516.21	54,458.23	26,785.65	19,809.36	55,592.94	43,876.21

A higher mean and lower standard deviation for total (M_A = 400,221; SD_A = 86,594. M_B = 386,908; SD_B = 95,203) and unique traffic (M_A = 308,200; SD_A = 70,162. M_B = 296,311; SD_B = 73,251) can initially be observed within 'Current data (A). When Google and GS are considered separately, however, we notice the opposite, with lower mean traffic and higher levels of variability around the mean, highlighting the low baselines in Y1 for both Google and GS.

[5] Interquartile range has been omitted owing to the small number of cases.

By excluding Y1's outlying data from these measures, as we have done in the bottom row of Table 3, we can note a higher mean, and less variability around the mean, for total (M_* = 434,895; SD_* = 63,516) and unique traffic (M_* = 335,337; SD_* = 54,458). Similarly, higher means and lower deviations for Strathprints traffic and unique traffic from Google Scholar can be observed. Interestingly, while higher means are observable for traffic and unique traffic from Google, a slightly higher standard deviation is found when compared to 'Prior data—B'.

It is significant to note from Table 2 that the traffic gains to Strathprints from GS during the reporting period experienced a more rapid rate of growth when compared to the general population of other web traffic sources. Even if we were to consider the large growth observed in Y1–Y2 as anomalous and were to exclude it from data as an outlier, a 74% and 70% increase in GS referral traffic and unique traffic respectively can still be observed between Y2 and Y4. This exceeds the growth rates in total (34%) and unique total traffic (39%) by some margin. Rapid growth in referral traffic from Google itself can also be found to have increased by 67% and 69% for traffic and unique traffic respectively. This is clearly lower than the figures for GS but nevertheless exceeds the growth rates observed in the wider pool of referral sources and may explain the higher standard deviation noted in 'Current data—A*'. The especially steep increase in GS traffic and unique traffic can perhaps best be observed by the profile of the chart presented in Figure 2.

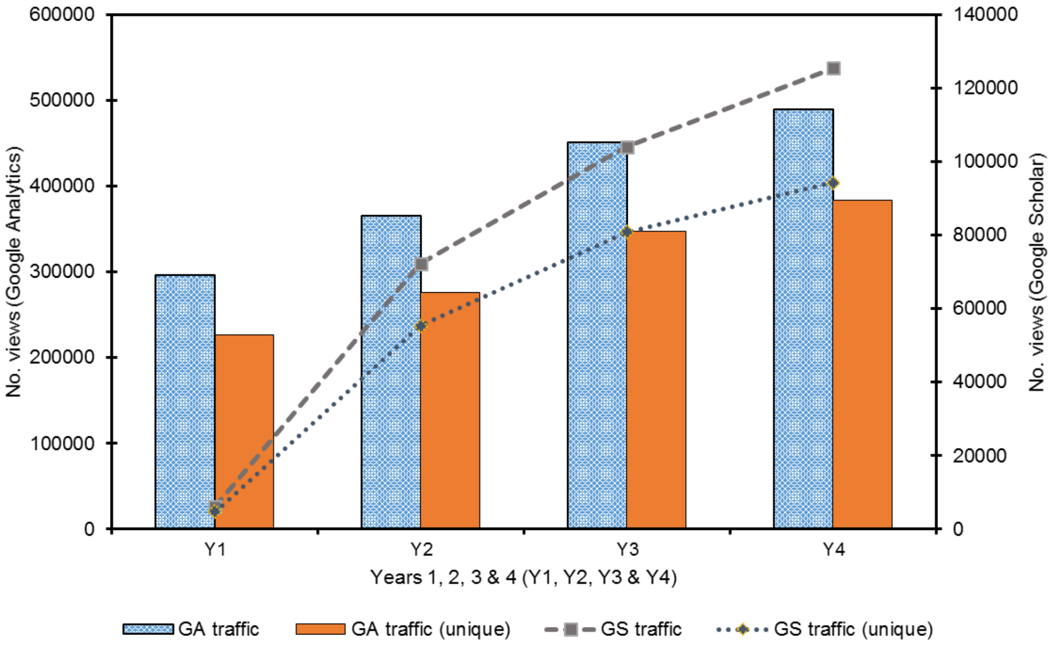


Figure 2. Volume of Google and Google Scholar referral traffic , including unique traffic in Y1, Y2, Y3 and Y4.

4.2. Repository Content Discovery and Usage

Improvements in impressions and clicks were observed in Y2 at 16% (*n* = 4,537,744) and 23% (*n* = 153,539) respectively when compared to the Y1 period. This upwards trend accelerated in subsequent reporting years. In Y3 a 69% (*n* = 7,687,550) and 21% (*n* = 185,232) increase in impressions and clicks respectively can be observed, followed by an 86% (*n* = 14,290,059) and 61% (*n* = 298,020) increase in Y4. This general upwards trend in impressions and clicks, including the aforementioned acceleration in Y3 and Y4, can be observed in Figure 3.

Data are contained in Table 4. The total percentage growth in impressions and clicks during the entire reporting period was 266% and 104% respectively. Figure 4 summarises the increase in clicks, impressions and COUNTER usage; sharper increases in impressions and clicks can be noted between Y2 and Y4.

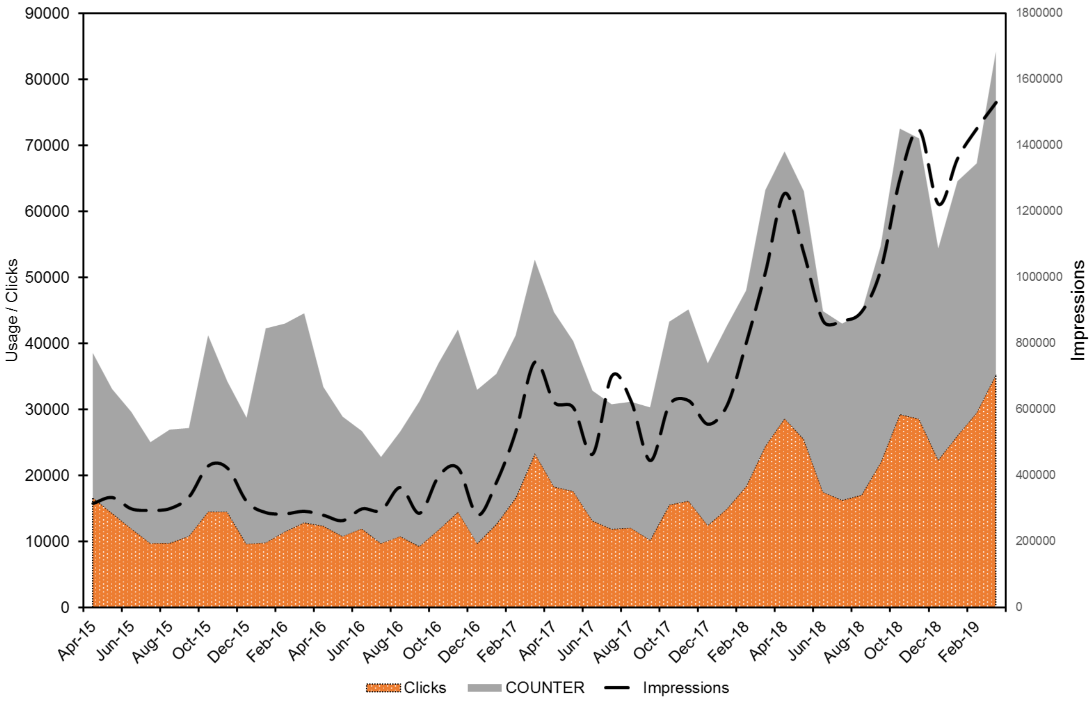


Figure 3. Strathprints COUNTER usage during Y1–Y4 alongside Google clicks and impressions during the same period.

Strathprints demonstrated a 62% growth in COUNTER compliant usage during the full period examined (i.e., Y1–Y4). It is noteworthy that this growth was observed despite only a 23% growth in full-text deposits during the same period. Even where embargoed content is factored into total full-text deposits, growth remained lower (54%) than the overall increase in usage. As noted in previous work [8], usage appears to demonstrate a more nuanced pattern when it is examined on a year by year basis. Usage in Y1–Y2 is particularly notable since it deviates considerably from the results reported previously and indicates that in the first year of observation Strathprints actually demonstrated negative growth, albeit minor. Conversely, Y4 yielded a 43% increase in COUNTER usage with only a 20% increase in full-text deposits recorded. Similarly, Y3 yielded an 18% increase in usage but experienced negative growth in full-text deposits (−22%).

It might be assumed that patterns in usage follow an exponential growth model, based on the volume of content deposited over time. In other words, that any increase in usage is directly proportional to increases in the volume of content deposited. This may indeed be true in some examples–and further research is encouraged in this respect; however, in this particular study, a weak exponential relationship was observed via exponential regression ($r^2 = 0.47$) with poor curve fitting notable (Figure 5), indicating the limited influence content deposit growth has on overall usage. Fitting with other common models such as linear, power or logarithmic was similarly weak.

It is apposite to highlight data from the previous section that Google search referrals and GS traffic increased well in excess of the full-text deposit rate, at 266% and 104% respectively; ergo the percentage of users being referred increased at a higher rate than the rate of full-text deposit during the reporting period. This is relevant because, based on these observations, it suggests that the rapid growth in search referrals from Google and GS has been a key factor influencing the increase in COUNTER usage.

To determine whether a correlation between Google clicks and COUNTER usage was present, Pearson's correlation coefficient was calculated for each year in the reporting period. A correlation was detected, ranging from a weak relationship in Y1 ($r = 0.11$) to a moderate positive correlation in Y2 ($r = 0.65$). Y1 and Y2 were followed by a strengthening of the relationship in Y3 ($r = 0.87$) and Y4 ($r = 0.97$). This strengthening of the positive correlation was confirmed via the t statistic for both Y3 ($t = 5.72, df = 11, p < 0.0005$) and Y4, at a far higher level of statistical significance ($t = 14.30, df = 11, p < 0.0005$).

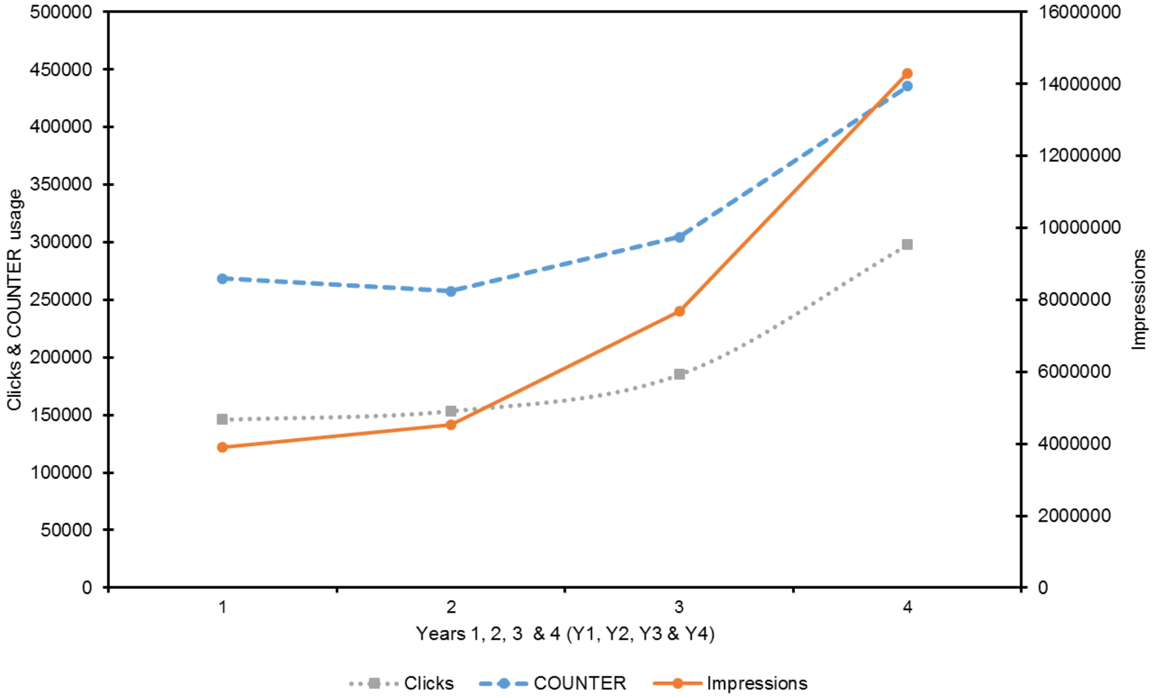


Figure 4. Charted data on observed clicks, impressions and COUNTER usage during Y1, Y2, Y3 and Y4.

Table 4. Data table of Strathprints COUNTER usage during and Google clicks and impressions during Y1–Y4. Volume of full-text OA deposits and volume of combined full-text and embargoed deposits.

	Impressions	Clicks	Usage	Deposits (OA)	Deposits (OA and Emb.)
Sub-total (Y1)	3,903,830	146,064	268,453	2326	2346
Sub-total (Y2)	4,537,744	153,539	257,560	2978	3074
Sub-total (Y3)	7,687,550	185,232	304,327	2314	3010
Sub-total (Y4)	14,290,059	298,020	435,467	2861	3620
Total (Y1–Y4)	**30,419,183**	**782,855**	**1,265,807**	**10,479**	**12,050**
% growth (Y2)	16.24	5.12	−4.06	28.03	31.03
% growth (Y3)	69.41	20.64	18.16	−22.3	−2.08
% growth (Y4)	85.89	60.89	43.09	23.64	20.27
Total % (Y1–Y4)	**266.05**	**104.03**	**62.21**	**23**	**54.31**

Computing the coefficient of determination (r^2) allows for better appreciation of the proportion of variance observed in the dependent variable (i.e., COUNTER usage) which is then predictable from the independent variable (i.e., Google clicks resulting from the changes implemented). In computing the coefficient of determination it was found that r^2 was significantly stronger in Y2 ($r^2 = 0.423$) than Y1 ($r^2 = 0.012$), but at such a low level that only 42% of variance in usage could be attributed to clicks. Variance narrowed considerably for Y3 ($r^2 = 0.766$) with a strong linear relationship between variables noted. This variance then narrowed again in Y4 ($r^2 = 0.953$), whereupon 95% of usage could be attributed to Google clicks. The incremental narrowing in variation between Y1 and Y4 can easily be observed from Figure 5, in which data points in Y3, and particularly Y4, are grouped more closely to the regression line.

An area that evades sufficient understanding in the data analysed thus far is the extent to which specific repository optimizations can also influence discovery on web search platforms that are not either Google or GS. This is largely because these discovery platforms lack any commensurate analytics. Acknowledging that the majority of repository traffic appears to originate from Google and GS, it is nevertheless possible to summarise the most common web traffic referral sources over the reporting period, as measured by GA and using the existing dataset, to establish whether changes could be observed in other platforms. Such data may lack the specificity typical of analyses earlier in this section but nevertheless enable a degree of inference about whether the optimizations have had an influence beyond Google and GS.

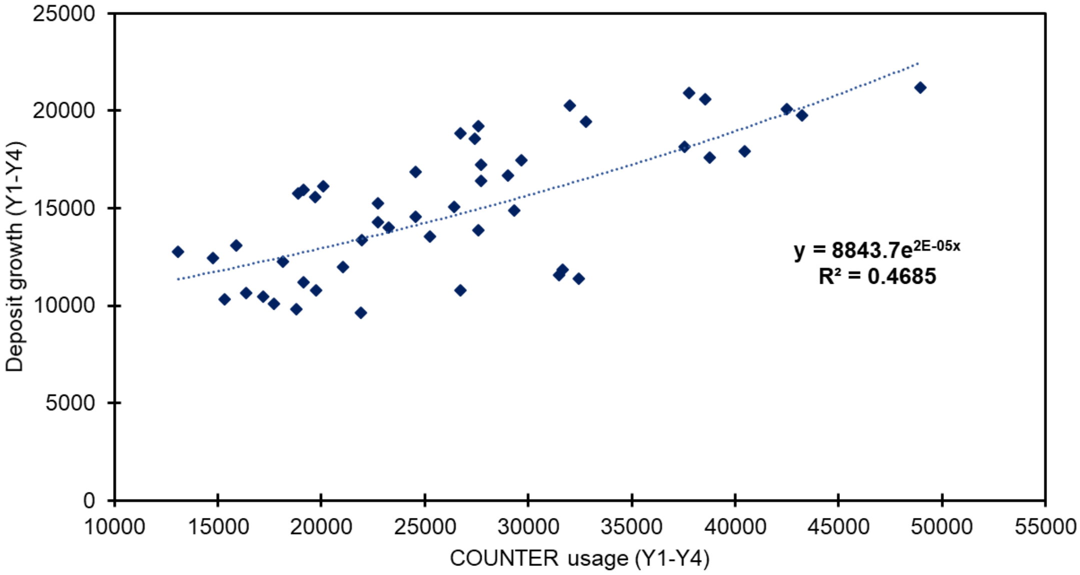


Figure 5. Exponential regression (r squared) for deposit growth and COUNTER usage (Y1–Y4).

Figures 6 and 7 chart the top ten web traffic referral sources during the reporting period, with local sources excluded (e.g., local university website searches, native searches on Strathprints, etc.). From Figure 6 it is possible to observe significant traffic growth from Google and GS. This is to be expected based on analyses earlier in this section, but little change can be observed in the other sources, such as Bing or Baidu, which display limited or zero growth. To better appreciate any modest change in traffic from these other sources, Figure 8 charts the same data but with data on Google and GS excluded. From this it is clear that variation in traffic can be observed across reporting years but no single profile suggests any sustained or significant growth. This would tend to infer that the technical improvements and adjustments implemented in this study demonstrate a Google-specific effect only. Traffic from other sources remained at such low volumes as to have a negligible impact on the overall volume of traffic received by Strathprints.

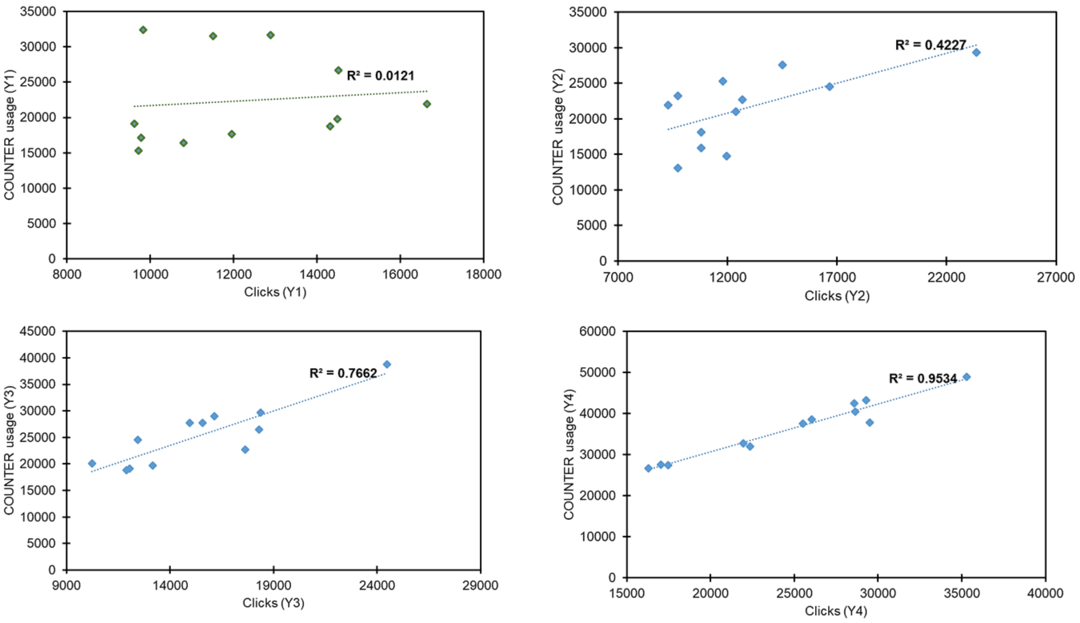


Figure 6. Coefficient of determination (r squared) for Y1, Y2, Y3 and Y4 between clicks and COUNTER usage.

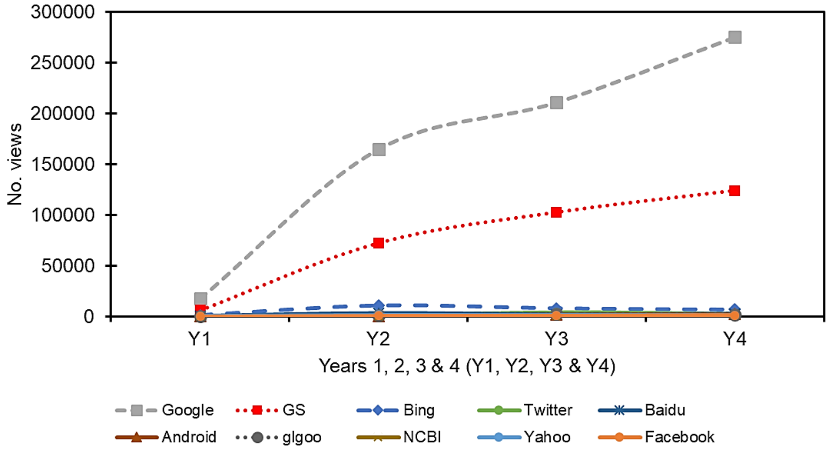


Figure 7. Top ten web traffic referral sources during Y1, Y2, Y3 and Y4. Local sources excluded.

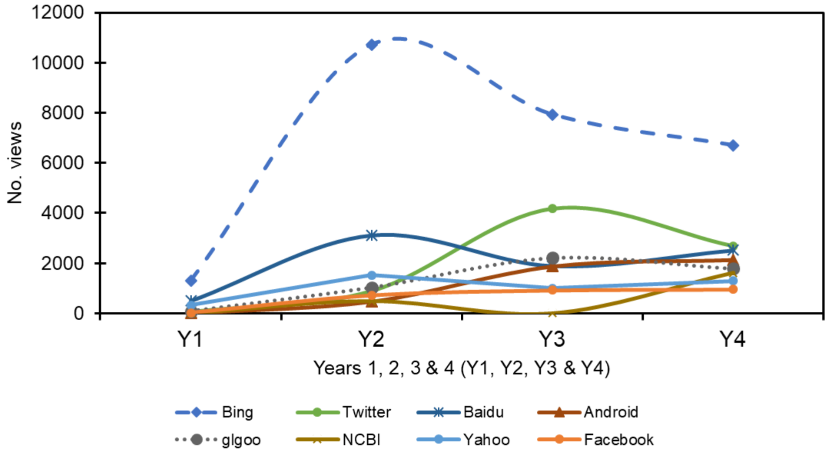


Figure 8. Top ten web traffic referral sources during Y1, Y2, Y3 and Y4, excluding Google, Google Scholar and local sources excluded.

5. Discussion

This article provides further analysis of the influence repository optimization approaches can have on the relative visibility, discovery and usage of an open repository. The nature of the longitudinal dataset used to track web traffic, usage and search metrics can be said to add additional weight to our findings and analysis. It corroborates previous evaluative studies [8] and reinforces prior evidence that specific technical enhancements to a repository can yield significant gains in web impact and usage.

Its dominance in search is such that Google is frequently found to be at the centre of many users' information seeking strategies [31]. The results from this study do not appear to challenge this continuing assertion, nor previously reported results [8], with 56% of all repository traffic referred by Google. Total web traffic was found to have increased by 65% during the period examined, with unique traffic growing 69%. Within this total, unique traffic from Google increased in excess of 70% during the reporting period, even where outlying data in Y1 were removed. Again, with Y1 excluded, 67% increases in total and unique traffic were noted for Google Scholar (GS). All of this was noted despite far lower rates of full-text deposit during the reporting period. The notion that usage was growing in line or exponentially as a result of deposit growth was also excluded in this instance. Temporal variations in time of data collection were nevertheless noted as influencing some of the results which suggests that future work, or replicative studies, should attempt analyses over different annual reporting lifecycles.

Y1 data were excluded from some of the web traffic analyses in Section 4.1 owing to their assumed anomalous appearance within subsequent data and underlying trending. It is worth revisiting this assumption here as the low baseline traffic detailed in Table 2 may have been outlying but not anomalous. Given the issues some repositories experience in achieving deep indexing by GS (e.g., [9,20]), and the low indexing recorded by some repositories in the recent Ranking Web of Repositories of July 2019 [32], it appears quite conceivable that the low traffic baseline for Strathprints was an accurate reflection of the GS indexing penetration of Strathprints prior to the technical changes in 2016. If this were the case then percentage increases of 1920% and 1854% in total and unique traffic respectively on GS were achieved during the reporting period, attributable to the technical improvements deployed, and reflect the rapid deep indexing of Strathprints by GS. It is relevant to highlight this since it also suggests that significant growth in traffic from GS is possible if steps are taken to optimize accordingly. Such high levels of indexing appear to be corroborated by recently published data in which Strathprints was placed in the top 5% of UK repositories and the top 10% of world repositories for number of records indexed by GS [32].

But while traffic originating from GS grew considerably–and GS indexing penetration also appears to be high–it is evident that the proportion of traffic originating from GS may actually be lower than those reported elsewhere. For example, [33], who previously examined the web traffic received by four repositories, found 48%–66% of traffic to be referred by GS, which is far greater than the 26% reported in this current study. Possible explanations for this GS traffic disparity could be positive rather than negative. For instance, it is conceivable that the technical strategies deployed on Strathprints were unusually successful in promoting traffic from competing search and discovery tools such that the proportion of GS traffic appears smaller than it otherwise might. In other words, it is less that traffic from GS is less than it should be and more that the changes implemented have yielded a far greater improvement in search tools relative to GS. This would correspond with prior observations [25]. Web traffic from Google certainly increased at a faster rate than GS; however, it should be noted that it also started from a higher baseline in Y1.

Another possible cause could be latency in detecting traffic resulting from the improved indexing of Strathprints by GS. This explanation posits that GS traffic will increase in forthcoming months and years as improvements in indexing depth and coverage translate into greater numbers of GS users being referred to Strathprints content over time. This hypothesis is something that can be easily verified by the present author and is a metric which will be monitored in future work, including any replicative studies.

A 62% increase in COUNTER compliant usage was reported despite far lower rates of full-text deposit, and even a decline in deposits during Y3. The rapid growth in search referrals from Google and GS was noted as a key driver in the overall increase in COUNTER usage during the reporting period as was their share of the total traffic Strathprints receives. This too was reflected in Google specific search metrics in which increases of 266% and 104% were observed in Google impressions and clicks respectively. The influence of Google clicks on COUNTER usage was verified via Pearson's correlation coefficient. This noted a strengthening of the relationship in every year, with high levels of statistical significance noted in years 3 and 4 (e.g., $p < 0.0005$) and r^2 demonstrating a strong linear relationship by Y4.

Accepting that correlation does not always equate to causation, the finding from this analysis that circa 95% of usage could be attributed to Google clicks warrants further scrutiny since it appears to demonstrate a potential disconnect with web traffic figures. Certainly a strong correlation exists—and this alone should provide a strong steer in how repositories are developed technically over coming years. The reported growth of Google and GS traffic clearly exceeded other traffic sources, and the increase in impressions and clicks was also significant. 56% of all web traffic may have arrived via Google but the predictive potential of this analysis seems slightly incongruous ($r^2 = 0.953$), suggesting that further data gathering or replication, preferably using different repositories, could be beneficial in verifying this finding. Indeed, a post hoc fallacy remains a risk since interference from possible extraneous variables remains difficult to discount given the research context. For example, overall global growth in web traffic during the reporting period was not explicitly controlled. The Cisco VNI global IP traffic forecast [34] predicts a compound annual growth rate of 26% between 2017 and 2022, which the traffic figures in this study appear to exceed; but without adequate experimental controls for such variables it is impossible to be definitive.

It is also necessary to state that the cumulative effect of a mounting corpus of full-text content (with full-text deposits accumulating year upon year) is not necessarily observable in a single year of observation. It is highly probable that content deposited in Y2 benefited usage metrics in subsequent years since factors critical in discovery and usage (e.g., search engine indexing, content aggregation, etc.) can take many months. Total percentage growth across all years (i.e., 62%) is therefore a more reliable indicator of the underlying pattern. Nevertheless, we should also note the limited influence content deposit growth appeared to have on overall usage, as corroborated by the weak exponential relationship between content that was noted between deposited and usage.

Recall that Acharya [20] and Tonkin et al. [21] reported the potentially negative consequences of coversheets on repository deposits. In this case study, automatically generated coversheets were enabled on Strathprints throughout the period of data collection. Given the enhancements to visibility and discovery which have been observed in this evaluation, it appears unsafe to conclude that the application of coversheets will always apply a negative drag on repository indexing. As this study has demonstrated, there are many variables which can potentially influence content discovery, coversheets are but one. Coversheets on Strathprints have since been disabled for local monitoring purposes but it seems necessary for future experimental work to verify the nature of their relationship to content discovery. Such work should seek to evaluate beyond Google Scholar since understanding surrounding coversheet usage currently appears to be influenced by a single academic discovery tool.

Limitations

Although it has been noted that Google accounted for the largest proportion of search traffic, the use of Google Search Console as a source of search metric data presents a data compromise by excluding metrics from other discovery tools. This decision was necessary owing to the lack of data available from other discovery tools and could therefore be described as a necessary limitation. The finding that there was little change in traffic volume from services other than Google and GS tends to infer that the repository optimizations deployed deliver a Google-specific benefit to repositories and may not provide the desired universal web impact or discovery improvements across other services.

A satisfactory explanation for this particular observation deserves further research since only one optimization (i.e., Google Data Highlighter) could be described as platform specific. All others were platform agnostic and reflected known 'white hat' best practice from the literature and platform inclusion guidelines.

There are of course limitations in the way this evaluation was approached and in the data collected. As we have noted already, experiments seeking to effect change on third party systems are immediately problematic since it becomes impossible to control for all variables hypothesised to influence web visibility. It is therefore not claimed that every known variable has been controlled in the work for this article; however, through exhaustive prior work [8], efforts have been taken to control as much as possible for all known variables. It is perhaps worth noting too that the brief nature of article precludes any additional data analysis; additional analyses were conducted but are not presented here owing to space limitations. Interested readers are nevertheless encouraged to download the raw data for analysis and potential new insights.

6. Conclusions

Section 5 highlighted several interesting discussion points but also raised several areas worthy of attention in future or replicative studies. These include the monitoring of traffic latency as a factor on improved GS indexing and better measuring the influence, if any, of coversheets on repository indexing more generally. However, any replication of this study should seek to improve the study design in certain key respects, especially improving the control of extraneous variables to avoid the possibility of correlation fallacy. A more productive design could include a collaborative study involving several repositories, whereby extant usage and web analytics are benchmarked across a number of disparate search agents and specific repositories used as a control. This would go some way to eliminating the potential influence of extraneous variables by confirming or refuting the observations noted in this article. Aspects of such an analysis could be performed using open data currently made available by IRUS-UK about active UK repositories, although collection of the necessary analytics on web traffic and search metrics requires invasive repository modifications.

The increasing importance of open repositories in fulfilling the discovery needs of both human and machine users is beyond doubt and it therefore remains essential to validate the continued relevance of repositories to users and their role as nodes within global scholarly communications infrastructure. Despite the limitations and some of the questions surrounding the findings, this article provides some persuasive evidence that open repositories should be managed in such a way as to enable routine technical enhancements to be deployed frequently and in response to intelligence and analytics pertaining to search, usage and web impact data. As noted in Section 1, repositories cannot remain static nodes in open scholarly communications infrastructure but instead active and responsive, driving content discovery, and usage and thereby better satisfying users' needs, while simultaneously addressing the challenges presented by proprietary systems. Analysis of the unique dataset presented in this article suggests that specific enhancements to the technical configuration of a repository can generate substantial improvements in its content discovery potential and ergo its content usage, especially when relevant metrics are monitored over several years. In this case study large increases were reported in COUNTER-compliant usage, key measures of web analytics and impact. Web traffic to Strathprints from Google and Google Scholar was also found to increase significantly with growth. Despite the noted limitations, the article demonstrates the link between repository optimization and the need for open repositories to assume a proactive development path, especially one that prioritises web impact and discovery.

Funding: This research received no external funding.

Acknowledgments: Thanks are extended to Martin Brändle (University of Zürich) for his helpful comments and suggestions in earlier drafts of this paper.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Pinfield, S.; Salter, J.; Bath, P.A.; Hubbard, B.; Millington, P.; Anders, J.H.S.; Hussain, A. Open-access repositories worldwide, 2005–2012: Past growth, current characteristics, and future possibilities. *J. Assoc. Inf. Sci. Technol.* **2014**, *65*, 2404–2421. doi:10.1002/asi.23131. [CrossRef]
2. Lynch, C.A. Institutional Repositories: Essential Infrastructure For Scholarship In The Digital Age. *Portal: Libr. Acad.* **2003**, *3*, 327–336. doi:10.1353/pla.2003.0039. [CrossRef]
3. McKiernan, E.C.; Bourne, P.E.; Brown, C.T.; Buck, S.; Kenall, A.; Lin, J.; McDougall, D.; Nosek, B.A.; Ram, K.; Soderberg, C.K.; et al. How open science helps researchers succeed. *eLife* **2016**, *5*, e16800. doi:10.7554/eLife.16800. [CrossRef] [PubMed]
4. de Castro, P. 7 Things You Should Know about Institutional Repositories, CRIS Systems, and Their Interoperability. 2017. Available online: https://perma.cc/69A4-TSL8 (accessed on 21May 2019).
5. Moore, S.; Gray, J.; Lämmerhirt, D.; Swan, A. *PASTEUR4OA Briefing Paper: Infrastructures for Open Scholarly Communication*; Technical Report; National Documentation Centre: Athens, Greece, 2016. Available online: http://pasteur4oa.eu/resources/229 (accessed on 20 December 2019).
6. Macgregor, G. Repository and CRIS interoperability issues within a 'connector lite' environment. In Proceedings of the 14th International Conference on Open Repositories (OR2019), Universität Hamburg, Hamburg, Germany, 10–13 June 2019. Available online https://strathprints.strath.ac.uk/68240/ (accessed on 27 December 2019).
7. COAR. *Next Generation Repositories: Behaviours and Technical Recommendations of the COAR Next Generation Repositories Working Group*; Technical Report; COAR: Göttingen, Germany, 2017. Available online https://www.coar-repositories.org/files/NGR-Final-Formatted-Report-cc.pdf (accessed on 21 May 2019).
8. Macgregor, G. Improving the discoverability and web impact of open repositories: Techniques and evaluation. *Code4lib J.* **2019**. Available online https://journal.code4lib.org/articles/14180 (accessed on 13 May 2019).
9. Arlitsch, K.; OBrien, P. Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Libr. Tech.* **2012**, *30*, 60–81. doi:10.1108/07378831211213210. [CrossRef]
10. Ferreras-Fernández, T.; Merlo-Vega, J.A.; García-Peñalvo, F.J. Impact of Scientific Content in Open Access Institutional Repositories: A Case Study of the Repository Gredos. In Proceedings of the First International Conference on Technological Ecosystem for Enhancing Multiculturality, TEEM '13, Salamanca, Spain, 14–15 November 2013; ACM: New York, NY, USA, 2013; pp. 357–363. doi:10.1145/2536536.2536590. [CrossRef]
11. Kelly, B.; Nixon, W. SEO Analysis of Institutional Repositories: What's the Back Story? Open Repositories 2013. Available online http://opus.bath.ac.uk/35871/ (accessed on 19 July 2019).
12. Pekala, S. Microdata in the IR: A Low-Barrier Approach to Enhancing Discovery of Institutional Repository Materials in Google. *Code4lib J.* **2018**. Available online https://journal.code4lib.org/articles/13191 (accessed on 13 August 2018).
13. Aguillo, I. Altmetrics of the Open Access Institutional Repositories: A Webometrics Approach. In Proceedings of the 23rd International Conference on Science and Technology Indicators (STI 2018), Leiden, The Netherlands, 12–14 September 2018; pp. 159–169.
14. Aguillo, I.F. TRANSPARENT RANKING: Institutional Repositories by Google Scholar. 2019. Available online http://repositories.webometrics.info/en/institutional (accessed on 26 December 2019).
15. Müller, U.; Scholze, F.; Arning, U.; Bange, D.; Beucke, D.; Hartmann, T.; Korb, N.; Meinecke, I.; Pampel, H.; Schirrwagen, J.; et al. *DINI Certificate for Open Access Repositories and Publication Services 2016*; Humboldt-Universität zu Berlin: Berlin, Germany, 2017. doi:10.18452/18178. [CrossRef]
16. Arlitsch, K. Driving Traffic to Institutional Repositories: How Search Engine Optimization can Increase the Number of Downloads from IR. *Zenodo* **2017**. doi:10.5281/zenodo.894564. [CrossRef]
17. Askey, D.; Arlitsch, K. Heeding the Signals: Applying Web best Practices when Google recommends. *J. Libr. Adm.* **2015**, *55*, 49–59. doi:10.1080/01930826.2014.978685. [CrossRef]
18. Arlitsch, K.; OBrien, P. Introducing the "Getting Found" Web Analytics Cookbook for Monitoring Search Engine Optimization of Digital Repositories. *Qual. Quant. Methods Libr. (QQML)* **2015**, *4*, 947–953. Available online https://scholarworks.montana.edu/xmlui/handle/1/9668 (accessed on 22 May 2019).
19. OBrien, P.; Arlitsch, K.; Mixter, J.; Wheeler, J.; Sterman, L.B. RAMP—The Repository Analytics and Metrics Portal. *Libr. Tech.* **2017**, *35*, 144–158. doi:10.1108/LHT-11-2016-0122. [CrossRef]

20. Acharya, A. *Indexing Repositories: Pitfalls and Best Practices*; Indiana University: Bloomington, IN, USA, 2015. Available online https://media.dlib.indiana.edu/media_objects/9z903008w (accessed on 5 September 2018).
21. Tonkin, E.L.; Taylor, S.; Tourte, G.J.L. Cover sheets considered harmful. *Inf. Serv. Use* **2013**, *33*, 129–137. doi:10.3233/ISU-130705. [CrossRef]
22. Bull, S.; Beh, E. Release 5 of the COUNTER Code of Practice. *Ser. Libr.* **2018**, *74*, 179–186. doi:10.1080/0361526X.2018.1447748. [CrossRef]
23. Wood-Doughty, A.; Bergstrom, T.; Steigerwald, D.G. Do Download Reports Reliably Measure Journal Usage? Trusting the Fox to Count Your Hens? *Coll. Res. Libr. (C andRl)* **2019**, *80*. doi:10.5860/crl.80.5.694. [CrossRef]
24. MacIntyre, R.; Needham, P.; Lambert, J.; Alcock, J. Measuring the Usage of Repositories via a National Standards-based Aggregation Service: IRUS-UK. In *New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science: Scale, Openness and Trust: Proceedings of the 19th International Conference on Electronic Publishing*; IOS Press: Amsterdam, The Netherlands, 2015; pp. 83–92. doi:10.3233/978-1-61499-562-3-83. [CrossRef]
25. Macgregor, G. Reviewing Repository Discoverability: Approaches to Improving Repository Visibility and Web Impact. 2017. Available online https://strathprints.strath.ac.uk/61333/ (accessed on 3 August 2018).
26. Wang, Z.; Phan, D. Using Page Speed in Mobile Search Ranking. 2018. Available online https://perma.cc/8QKP-NE5S (accessed on 3 August 2018).
27. Zhang, F. Rolling Out Mobile-First Indexing. 2018. Available online: https://docs.lib.purdue.edu/libf (accessed on 22 January 2020).
28. Jayasankar, S. Our Approach to Mobile-Friendly Search. 2015. Available online: https://perma.cc/5EQQFCGC (accessed on 22 January 2020).
29. Google. Google Search Console. 2019. Available online https://www.google.com/webmasters/tools/home (accessed on 13 May 2019).
30. Macgregor, G. Supporting dataset for: Repository optimisation and techniques to improve discoverability and web impact: An evaluation. *Dataset* **2018**, doi:10.5281/zenodo.1411207. [CrossRef]
31. Rowlands, I.; Nicholas, D.; Williams, P.; Huntington, P.; Fieldhouse, M.; Gunter, B.; Withey, R.; Jamali, H.R.; Dobrowolski, T.; Tenopir, C. The Google generation: the information behaviour of the researcher of the future. *Aslib Proc.* **2008**, *60*, 290–310. doi:10.1108/00012530810887953. [CrossRef]
32. CSIC. Transparent Ranking: Institutional Repositories by Google Scholar (May 2019) | Ranking Web of Repositories. 2019. Available online https://repositories.webometrics.info/en/institutional (accessed on 10 August 2018).
33. OBrien, P.; Arlitsch, K.; Sterman, L.B.; Mixter, J.; Wheeler, J.; Borda, S. Undercounting File Downloads from Institutional Repositories. *J. Libr. Adm.* **2016**, *56*, 1–24. Available online https://scholarworks.montana.edu/xmlui/handle/1/9943 (accessed on 13 August 2018). [CrossRef]
34. Cisco Systems, I. *Cisco Visual Networking Index: Forecast and Trends, 2017–2022*; White Paper; Cisco Systems, Inc.: San Jose, CA, USA, 2019. Available online https://perma.cc/9D9X-Y7MZ (accessed on 26 December 2019).