Integration of distributed terminology resources to facilitate subject cross-browsing for library portal systems

By
Libo Si

Submitted in partial fulfilment of the requirements for the award of
Doctor of Philosophy of Loughborough University

Supervisors: Dr. Ann O'Brien
Dr. Steve Probets
Department of Information Science
Loughborough University

Nov 2009

ACKNOWLEDGEMENTS

I would like to praise and thank my Lord, Jesus Christ, in whom I trust, for he "arms me with strength and makes my way perfect" (Psalms 18:32). I was a self-funded PhD student, but in him, I am not in want throughout my PhD. Also, I would like to thank to the brothers and sisters in the Chinese congregation of Loughborough Holywell Free Church. Through them, I began to know my living Lord, Jesus Christ, who has given me all the strength, wisdom, understanding, insight, patience, and blessing for my PhD work.

Deep gratitude to my supervisors Dr. Ann O'Brien and Dr. Steve Probets, for their kind advice, guidance and encouragement. I also would like to thank all the staff at the Department of Information Science, Loughborough University for their kind support.

I am grateful to my parents, for their financial support at the beginning of my PhD research, and their encouragement.

Many thanks to my wife for her enduring love in Christ.

Many thanks to the expert interviewees and evaluators for kindly providing valuable data for my data collection work.

Abstract:

With the increase in the number of distributed library information resources, users may have to interact with different user interfaces, learn to switch their mental models between these interfaces, and familiarise themselves with controlled vocabularies used by different resources. For this reason, library professionals have developed library portals to integrate these distributed information resources, and assist end-users in cross-accessing distributed resources via a single access point in their own library. There are two important subject-based services that a library portal system might be able to provide. The first is a federated search service, which refers to a process where a user can input a query to cross-search a number of information resources. The second is a subject cross-browsing service, which can offer a knowledge navigation tree to link subject schemes used by distributed resources.

However, the development of subject cross-searching and browsing services has been impeded by the heterogeneity of different KOS (Knowledge organisation system) used by different information resources. Due to the lack of mappings between different KOS, it is impossible to offer a subject cross-browsing service for a library portal system.

In order to improve interoperability between controlled vocabularies used by different resources, a number of terminology services have been developed, such as HILT Terminology Service (TS), STAR TS, OCLC TS. Basically, a terminology service can provide mappings between different vocabularies to different library portal systems, and then help library portal systems facilitate subject cross-browsing and searching. However, it is worth noting that in most current cases, one terminology service only holds a small number of vocabularies, and it may be impossible to have a terminology service including all the KOS required in a library portal. For this reason, the aim of this research is to develop a framework for establishing a middleware service between different terminologies to improve interoperability between different controlled vocabularies.

In this research, nine terminology experts were interviewed to collect appropriate knowledge to support the development of a theoretical framework. Based on the

developed theoretical framework, a simplified software-based prototype system was developed to demonstrate the methods used in the theoretical framework. Subsequently, six information consultants were invited to evaluate the prototype system, and provide feedback to improve the framework.

The major findings showed that it is important to establish a middleware service to integrate different terminology resources with a locally-used controlled vocabulary for facilitating subject cross-browsing. In order to establish such a middleware service, the specific findings focus on:

1. Selecting a switch language to establish semantic mappings between different terminology resources and services.
2. Applying local vocabularies within a library as subject cross-browsing structures to navigate users.
3. Developing a knowledge base to technically cross-access distributed terminology resources using different formats, protocols, and query languages.
4. Establishing machine-to-machine interactions between a federated search service and the middleware framework, and then enabling the federated search service to use mapped terms to cross-search the distributed resources indexed by these terms. This should acquire item-level metadata results from distributed resources.
5. Applying or developing query expansion algorithms to semantically expand mapped terms to return a range of terms considered semantically close.
6. Developing user-friendly subject browsing interfaces to encourage end-users to put their own intelligence towards selecting the most appropriate terms for their subject searching and discovery.

A comprehensive combination of users' intelligence, computerised intelligence, and intelligence provided by mapping staff was considered to be paramount for the development of such a middleware system.

Content List

List of figures

List of tables

Chapter one: Introduction and Background

1.1 Research context and significance

The number of networked information resources that coexist in today's information environment is increasing. These resources include electronic journals, A&I services, e-print archives, conference paper collections, electronic books, digital images, etc. Each of these resources has its own distinct user interface, provides user-tailored functionality, and is indexed according to particular indexing languages. Users who wish access to some of these resources may have to interact with different user interfaces, learn to switch their mental models between these interfaces, and familiarise themselves with the terminology used by different resources. Faced with this situation, a solution called "portal" was introduced in the early and mid-1990s, which "provides access to a range of heterogeneous network services, local and remote, structured and unstructured" (Miller 2002). In other words, a portal serves as a bridge between a range of digital collections and a digital library environment. It is worth noting that in most cases, a portal system does not create its own content, but link its users to different information resources.

Several types of portals have been developed for different purposes and have been tailored to different communities (Powell 2003, and Davies 2007). Subject-based portals are usually developed by a third party, and are intended to aggregate content from heterogeneous information resources within a particular subject area (Borgan 2003). They usually present end users with a subject-specific view of available resources (Guy 2005). Library portal products are implemented as "applications which allow one-stop-shop access/search and discovery via a unified single-point interface to organised heterogeneous resources and enabling services to pre-defined communities (users)" (ELAG 2002). A refinement to this idea is an institutional portal, which focuses on developing a framework for integrating information resources useful to the people working in an institution (Boye 2006). At the time of writing, the trend is to integrate the

local and remote resources within a portal (Lewis 2008, Davies 2007, Joyce *et al.* 2008, and Guy 2005). For example, Ex Libris' portal service called "Primo" can integrate institutional repositories, local e-learning repositories, remote subject gateways, and remote commercial databases into a single access point (Sadeh 2007 and Lewis 2008). Other library portal products, such as ddWiz, SirSi Rooms, MetaLib, WebFeat, etc., have also been widely-applied in different library services.

A number of important technologies are needed to deploy a fully-fledge portal service (Koch 2000, Cox and Yeates 2003, Ramsden 2003, Boss 2003 and Butters 2003). These include cross-searching technologies, tools for resource discovery, harvesting, alerting, authentication/authorisation, open linking, profiling of users, etc. Among these features, federated search service, which refers to a process in which a user inputs a query to cross-search a number of information resources, has been widely highlighted in the literature (Sadeh 2004, Sadeh 2004b, Sadeh 2007, Ramsden 2003, Mah and Stranack 2005, Fang 2004, Davies 2007, and Koch 2000). Another important feature, which is usually neglected, is a subject cross-browsing interface offering a knowledge navigation tree to link subject metadata across different information resources (Koch 2000 and Joyce *et al.* 2008). This kind of subject cross-browsing is particularly helpful for novice users who are not familiar with a particular subject terminology language, as it enables novice users to be navigated by subject hierarchies to find relevant information.

Most current portal services do not offer this kind of subject cross-browsing service, because of the heterogeneity between different controlled vocabularies used by these information services. It is widely-agreed that different controlled vocabularies differ in their subject areas, levels of specificity of concepts, degrees of pre/post-coordination, semantic relationships, and the use of languages (Doerr 2001, Lancaster and Smith 1983, and BS8723-4). The main effort to improve the semantic interoperability between different controlled vocabularies, and deploy such subject cross-browsing is the EU Renardus Project. This project developed subject cross-browsing by creating mappings between different classification schemes used by participant EU quality control subject gateways (Heery *et al.* 2001), and focused on exploring the approaches to improving the

semantic interoperability between a number of subject gateways/portals. However, further research related to the methods to enable cross-browsing by subject in the context of library portal products, such as Ex Libris Primo, SirSi Rooms, and WebFeat, were not discussed in the Renardus Project.

Other projects, such as JISC HILT Project, and AHRC SRAR Project, point to developing terminology services to improve the semantic interoperability between different controlled vocabularies. These projects were intended to hold a number of controlled vocabularies in databases, develop programmatic Machine-to-Machine (M2M) interfaces that allow other services to manipulate their terminology data in M2M ways during the searching/browsing process. Mappings between different controlled vocabularies were partly established in these projects, and were stored within their databases. Some of the terminology services facilitate subject cross-browsing (McCulloch *et al.* 2005, and Tudhope *et al.* 2006). On the other hand, it may be impossible to have a terminology service that includes all the controlled vocabularies required in a library portal. For example, the HILT terminology service, as the largest terminology service in the UK, only includes 13 controlled vocabularies. A number of widely-used vocabularies are not covered. In addition, many institutions use non-standard in-house vocabularies (HILT Phase 2 2003, Koch 1997). An in-house vocabulary usually represents a long-term commitment of a particular community. In this context, it is important to consider methods to combine a variety of terminology services and local terminology resources to support subject cross-browsing within a local library portal.

In order to combine various terminology services and local terminology resources in comprehensive ways, it is important to consider numerous factors that challenge the interoperability between different terminology services (Arms 2002, Tudhope *et al.* 2006, and Patel *et al.* 2005). One of the most important factors is the semantic diversity between different KOS from different terminology services, and the ways to solve this problem point to the construction of semantic mappings between different vocabularies (Coates *et al.* 1978, McCulloch 2004, Chan and Zeng 2004, and Heery *et al.* 2001). Other factors are multiple data formats (e.g. SKOS, Zthes XML Schema, MARC, etc), multiple

access protocols used to query against different terminology services (e.g. Z39.50, SRW/U, SPARQL, etc), multiple metadata schemes to describe the general terminological information about various vocabularies and mappings (Dublin Core Metadata Scheme, self-defined schemes, etc), and multiple database systems used for the development of terminology services (e.g. relational databases, RDF triple store, file systems, etc). It is important to explore methods to solve the incompatibility problems for all these factors, and then develop the subject cross-browsing services for library portal systems.

1.2 Research scope

Before stating the aim and objectives, it is important to clarify a number of concepts in this research, and declare the scope of this research.

Firstly, the research focuses on providing subject cross-browsing services in the context of library portal systems. The term "library portal" describes a service integrating local and remote resources to satisfy the particular information needs of University library users. Ex Libris Primo and SirSi Room are examples of library portal products. Although an increased emphasis on multi-media, images, cultural heritage objects is noted in recent research into digital libraries (Dempsey 2000), this research will not cover these complicated information resources. The term "information resources" in this thesis mainly refers to textual information.

Secondly, the term "knowledge organisation system" (KOS) is widely-used in this research. The term is applied based on Hodge's (2000) definition, which "encompass all types of schemes for organising information and promoting knowledge management", such as classification schemes, subject headings, taxonomies, authority lists, thesauri, ontologies, etc.

Thirdly, different KOS were developed in many different languages. This research focuses only on improving the interoperability between different English vocabularies.

Interoperability issues related to multilingual vocabulary access are outside the scope of this research.

Fourthly, it is true that there are a variety of information resources on the Internet, such as those for entertainment, sports, politics, etc. This research focuses only on providing methods to retrieve scholarly information for UK further and higher education.

Fifthly, this research uses the term "terminology resources" to describe a variety of KOS-related resources that are published in relevant digital formats. These include:

- Terminology services, such as OCLC Terminology Service, HILT Terminology Service, etc.—which were developed as shared services that allow other services to access their terminological data;
- Controlled vocabularies, which are used to index important collections, and were represented in encoding formats, and published on the Web;
- Mapping sets between different controlled vocabularies which are represented in encoding formats, and published on the Web;
- Local vocabularies which are used by a library portal system for local subject indexing and cataloguing.

1.3 Aim and objectives

This research aims to develop a framework for establishing a middleware between different terminology services to improve interoperability between different controlled vocabularies. The framework also aims to provide a mechanism to achieve interoperability between non-standard local vocabularies and various shared terminology services. In this context, a library portal could use this middleware combining various terminology resources to offer their subject cross-browsing services. The specific aim and objectives are listed as follows:

Aim: Develop a framework to facilitate subject cross-browsing service for library portal products.

Objective 1:

To understand the basic principles involved in developing controlled vocabulary-based information retrieval systems for library portal products, and basic methods to improve semantic interoperability between different controlled vocabularies.

- Objective 1.1: To investigate methods to develop library portal products, such as cross-searching technologies, data conversion, etc, and understand how the various terminology resources could be embedded into the library portal;
- Objective 1.2: To investigate how various controlled vocabularies support information retrieval systems in both traditional and innovative ways;
- Objective 1.3: To review a variety of widely-used controlled vocabularies, and understand their subject areas, levels of specificity of concepts, degrees of pre/post-coordination, semantic relationships, the use of languages, representation formats, services using them, etc;
- Objective 1.4: To investigate methods to improve the interoperability between different controlled vocabularies.

Objective 2:

To explore appropriate methods to combine a variety of terminology resources, and integrate the combined terminologies into library portal services.

- Objective 2.1: To investigate a number of terminology service projects, and explore how they work within a library portal system to facilitate subject cross-browsing;
- Objective 2.2: To identify the appropriate technologies and standards that support the exchange of terminological information between different terminology resources;
- Objective 2.3: To define a semantic network to exchange terminological information between different terminology services and local controlled vocabularies;

- Objective 2.4: To define a workflow to distribute mapping work to different participants for the development of this framework;
- Objective 2.5: To identify a number of relevant user scenarios to facilitate user-friendly subject cross-browsing.

Objective 3:

To formulate a theoretical framework to facilitate subject cross-browsing service for library portal products;

A number of methods were selected for the development of the framework, which refer to the approaches to mapping different controlled vocabularies, the appropriate technologies, the user scenarios that the framework support, the mapping workflow used to distribute the mapping work to different participants, etc. These methods were finally used to develop a theoretical framework.

Objective 4:

To develop a simplified programmatic prototype system of the middleware framework to demonstrate the methods used in the theoretical framework.

Based on the formulated theoretical framework, a simplified programmatic prototype system was developed.

Objective 5:

To evaluate the prototype system for improving the effectiveness and usability of the framework, and then improve the developed framework.

1.4 Structure of the thesis

According to the aim and objectives identified, the thesis is organised as follows:

Chapter Two outlines different methods to organise, discover, and search scholarly information; it points out the heterogeneity between different KOS, and relevant methods to improve semantic interoperability between different KOS, and finally it presents a set of research questions.

Chapter Three introduces the research methodology applied.

Chapter Four describes two types of findings for forming the basis to propose a theoretical framework in Chapter Five. The first type of findings is based on an investigation of different features within a variety of identified KOS. The second type of findings is based on interviewing nine experts to gain a basic understanding of how to develop subject cross-browsing services for library portals.

Chapter Five introduces the rationale and principles used to develop the theoretical framework for facilitating subject cross-browsing

Chapter Six introduces the specific principles for the development of a simplified prototype system based on the developed theoretical framework. This prototype system forms a basis for the evaluation work.

Chapter Seven discusses the findings from six evaluation tests provided by experts.

Chapter Eight discusses the answers to the research questions presented in Chapter Two.

Chapter Nine offers conclusions in response to the aim and objectives, and presents potential areas for further development.

Chapter Two: Literature Review

2.1 Google search engine and Meta-search

In today's information environment, full-text search engines, such as Google, have become more and more popular. These search engines use web crawlers to automatically index every word in web documents, and various ranking algorithms to return results in order of decreasing relevance (Rowland 1998). This approach reduces human indexing effort and cost, and improves recall of information retrieval. Most users feel free to use the search engine to type the query and conduct their searches. In many cases, a full-text search for information may retrieve thousands of results. Among these, however, there may be a huge number of unwanted and unexpected items, because of the ambiguous meanings of a given subject search term. Rowley (2001) identifies this problem as the loss of precision.

In another case, Dawson (2004) and Tudhope and Binding (2008) emphasised the ambiguity problem in free keyword searching. For example, it was realised that different people may use different words to describe the same concept, and the same word may ambiguously be used to represent different concepts. For this reason, users have to struggle with many ambiguous homographs within a free keyword search system, and in order to access an information object, the users have to successively re-formulate their queries until they find high-quality information (Soergel *et al.* 2004).

Research has indicated that most full-text search engines are not able to index the whole range of today's Internet, and there are a lot of hidden resources that search engines cannot reach (NUA Internet Surveys, 1998, Chen 2006). In 2001, Google indexed less then 1% of the total number of digital information resources on the Internet (Bergman 2001). A number of people have identified different types of hidden digital resources on the Internet, which may not be indexed (Chen 2006, pp.413-427, Sadeh 2006, and Jacso 2005).

With this consideration in mind, two issues arise. The first is related to the ambiguity of the users' given search terms, and loss of precision. It is important to develop query disambiguation functions to let users clarify the subject context of a given term. For this reason, a number of experts emphasise that controlled vocabularies, such as synonym rings, authority files, classification schemes, taxonomies and thesauri, can play an important role in aiding information retrieval, improving searching precision, and supporting term disambiguation (Gnoli 2004, Svenonius 2001 and HILT III 2007). In other words, the traditional principle of library subject access, which refers to encoding applied controlled vocabulary terms in subject-related metadata elements in a metadata record, might be a solution to the ambiguity problems. In addition, Kline (2002) indicates that these two approaches (controlled vocabulary and free keyword search) need to be combined depending on different contexts. In the literature, the importance of using controlled vocabulary to complement free keyword search is widely emphasised. (Kline 2002, Green 2000, Rowley 2001, Shiri and Revie 2000, pp.273-280 and Soergel 2004).

The second issue is related to the selection of various information resources tailored to an organisation. It is important for a particular organisation to clarify what information resources should be useful for its users. Some information resources covered by some full-text search engines may not be useful for the particular organisation, but at the same time some useful resources may not covered by engines. The current effort to address this issue points to the development of library portal products, which offer the capability of integrating different information resources tailored to the users' information requirements. A library portal product "allows one-stop-shop access and discovery via a single-point interface to organised heterogeneous resources and enabling services to a pre-defined community (users)" (Davies 2007, p.642). Federated search is a basic function in most library portal products. With this approach, a federated search engine is employed to map an end-user's subject term to query against the search engines adopted by different information providers, and then the federated search service will receive the results from each information provider, convert the heterogeneous metadata into a consistent format, and display them to the end-users in ranked order (Buckland 1999). Although critics point to the issues of complexity and interface problems in library portal systems (Haya

et al., 2007), there is more discussion in the literature about the advantages of using federated search over the use of full-text search (Chen 2006, Mah and Stranack 2005, Fang 2004, Jacso 2005, Sadeh 2006, Sadeh 2004, and Koch 2000). The advantages include:

1. Different information resources can be integrated and indexed according to subject areas, data providers, types of resources, languages, etc. The users can discover important information resources based on various navigate-able categories. In other words, this kind of product is not only a cross-searching tool, but also an information resource discovery tool.
2. Local information, such as local OPACs, institutional repositories, learning objects, etc., can be integrated and cross-searchable in a library portal system.
3. Selecting high-quality information resources is the responsibility of the library staff, which reduces the effort of the users.
4. Most of the federated search products use ranking algorithms to rank the results in chronological order, which is different from Google's. To some extent, these types of algorithms is more tailored to local requirements.
5. Based on the use of library portal software, each department in a University could develop its own subject portal tailored to their special subject needs, and each user could personalise the portal to suit his/her subject requirements.
6. Most library portal systems can easily offer a powerful mechanism to provide direct linking to the full-text content or to the citation.
7. It is much easier to implement Boolean searching in federated search engines than in full-text search engines.

The ways to facilitate cross-searching of different information resources focus on establishing mappings between different metadata elements from different information resources (Chan and Zeng 2006). Thus, next section will focus on discussing metadata and related concepts.

2.1.1 Metadata and metadata schemes

In the literature, the term "metadata" refers to, "structured data about an object that supports functions associated with the designated object" (Greenberg 2003, p.1876). It is a way to describe information items. A metadata scheme refers to a set of metadata elements for some specific purposes (Greenberg 2005, p.18). Today, different metadata schemes, such as the Dublin Core Metadata Scheme, Learning Object Metadata Scheme, Encoded Archival Description, etc, have been developed with the specific requirements of different communities. Sicilia (2006) summarised the roles that a metadata scheme can play. These may include "facilitating retrieval, stating intellectual property rights, claiming authorship, driving the behaviour of online information systems or easing the combination of information resources" (Sicilia 2006, p.213).

One of the most widely-used metadata schemes is the Dublin Core Metadata Initiative (DCMI), which began in 1995 to develop a commitment for resource discovery on the web (dublincore.org). The DCMI, as a broad, interdisciplinary agreement, has been widely applied to support information discovery (Howarth 2003). Fifteen elements included in this scheme. These are: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights. This metadata scheme is a multidisciplinary scheme developed to facilitate resource discovery and support interoperability and data exchange among communities. As a result, DCMI's environmental domain includes the larger information resource description community (Weibel, 1995).

Generally speaking, DCMI metadata plays an important role in both collection level description and item level description. When using DCMI metadata to describe items, it is important to use different DCMI elements to "tag the main characteristics of individual documents to maintain essential linkages between documents and the functional/transactional context of their creation so that every record can be allocated a unique reference, and the registers contained subject terms and metadata relating to both provenance and related documentation" (Tough and Moss 2003, p.24). Similarly, Hunter and Guy (2004, p.168) also point out that item metadata should be able to provide HTTP

hyperlinks to full text documents contained in the archives. For example, in UKOLN's case, the metadata records point directly to the original full-text pages of e-journal papers. When using DCMI scheme to describe collections, Hunter and Guy (2004, p.168) argue, "making collection level details of the resources available in the form of a metadata description, ..., create a way of making entry-level data about collections available to researchers worldwide".

In some specialised domains, however, DCMI is not detailed enough to adequately represent resources (Howarth 2003). Thus, it is important to develop some other metadata schemes to describe different types of information objects.

Today, many other metadata schemes have been designated to address specific problems or needs in describing objects. Some of these schemes are described below.

Learning object metadata: The IEEE (Institute of Electrical and Electronics Engineers) Learning Technology Standards Committee (LTSC) has developed a standard for Learning Object Metadata (LOM). The LOM includes a set of elements to describe learning objects, and support learning objects discovery. Some particular characteristics in this metadata scheme are employed to describe a learning object. These characteristics are classified in "life cycle, meta-metadata, educational, technical, educational, rights, relation, annotation, and classification categories" (IEEE Learning Technology Standards Committee, 2002, p.5).

Machine Readable Cataloguing Record: One of the most important metadata formats is the MAchine Readable Cataloguing Record (MARC) linked with AACR2 (Anglo-American Cataloguing Rules), which is the most used format for library records. The MARC records contained in library catalogues are examples of metadata. "MARC has several fields for the different attributes of a bibliographic entry such as title, author, et al.". (Baeza-Yates and Berthier 1999, p.143) Because MARC is a very complex and detailed metadata system, it is not suitable for untrained electronic document creators or publishers to create descriptive metadata. They would have to formulate the metadata

according to another scheme, and then encode it using the proper MARC fields, subfields, and indicator (Lazinger and Hunter 2001).

Federal Geographic Data Committee: This standard "aims to provide a common set of terminology and definitions for metadata about digital geospatial data. The FGDC standard is a content standard; it does not dictate layout or an encoding scheme" (Taylor 1999, p.93). In addition, the FGDC is a complex metadata standard with over 300 data elements, arranged hierarchically. The main categories of metadata are "identification information, data quality information, spatial data organization information, spatial reference information, entity and attribute, distribution information and metadata reference" (Taylor 1999, p.94).

In addition, many information professionals focus on exploring a range of front-end searching services, such as browsing in categories that navigate people through a hierarchical tree, and keyword search based on inputting the keywords in some search engines. However, as Walker (2001, pp.127-133) argues, "most front-end searching services do not provide for the important back-end integration whereby a user, having chosen a specific resource in which to start their research, can link to other resources and services". For example, when a user finds a useful citation, and wants to access the full-text content of this article, an information system may not be able to provide the user with the relevant link to the full text of this article. Because there might be different copies for the same electronic article within different collections, and a user within a particular organisation may be able to access to only one copy of this article, it is necessary to render appropriate copies to the correct users. As a result, Van de Sompel (1999) highlights the importance of another metadata application called OpenURL. The basic idea is that via an OpenURL, users can be directed to appropriate copies of an electronic article (Walker 2001, pp.127-133). There are main two parts in an OpenURL (Powell and Apps 2001). The first component is a basic URL, which is used to guide the users to appropriate copies in the users' context. The selection of a basic URL for an OpenURL depends on the users' and organisational preferences with the consideration of the agreements with the information providers. For example, it is possible that different

copies of an article are stored in two or more different databases, and a particular organisation may just subscribe to one of these databases, which can be called "Database A". In this case, if an OpenURL for this article is established within this organisation, it is important to select the URL of Database A as the first component of this OpenURL. The second component of an OpenURL is called a "Query" that includes a range of metadata descriptions that make up the citation for an electronic article (Powell and Apps 2001).

Using these different metadata standards to describe different information objects causes the problem of incompatibility between different metadata schemes across the domains (Hunter 2001). Heery (2002) highlights the importance of making different metadata schemes interpretable. In this context, the term "metadata interoperability" refers to "a fundamental requirement for access to information within networked knowledge management systems" (Hunter 2001). The next section focuses on interoperability issues and metadata crosswalks.

2.1.2 Metadata interoperability

Chan and Zeng (2006a, and 2006b) summarised a number of ways to achieve or improve metadata interoperability among different metadata schemes and applications. These different methods can be described as follows:

1. Derivation: one metadata scheme can be developed based on the principle and structure of an existing one (Chan and Zeng 2006a).
2. Application profile: an application profile can be defined by combining a selected range of metadata elements from different metadata schemes for some application-specific purposes (Heery and Petel 2004). For example, in order to make the Renardus portal service cross-search different EU portals, the Renardus Application Profile combined a set of metadata elements from different metadata schemes, each having different namespaces: Renardus Metadata Element Set (rmes), Renardus Metadata Element Set Qualifiers (rmesq), Dublin Core Metadata Element Set, version

1.1 (dc 1.1), Dublin Core Metadata Element Set Qualifiers (dcterms), and DCMI Type Vocabulary (dcmitype).

3. Crosswalk: "A crosswalk is a specification for mapping one metadata standard to another" (St. Pierre and LaPlant 1998). Crosswalks refer to the ability to make different metadata standards interoperable.

In some cases, people create a metadata crosswalk by defining some kind of agreement between two metadata repositories (Heery and Wagner 2002). Hunter (2001) points out the limitations of this one-to-one metadata crosswalk, and mentions, "this approach does not scale to the many metadata vocabularies that will develop". Peig *et al.* (2001) presented another way to make many-to-many metadata crosswalk. The basic idea is to develop or apply a common scheme as a switch scheme, and a wide variety of communities can implement this common scheme, or map their metadata schemes into it. Chan and Zeng (2006a) point out that, in most cases, a crosswalk just focuses on mapping based on metadata specification, but does not consider conversion between the allowable data values of different metadata elements from different metadata schemes. They emphasise that metadata conversion and exchange should be considered at the record level. In order to make different metadata interoperable at the record level, Reese (2006) highlights the importance of designing a data conversion programme, and storing the converted metadata in a central knowledge base.

4. Metadata Registry: A metadata registry refers to an application that provides services based on information about 'metadata terms' and about related resources (Johnston 2005). It should provide an index of terms so that via a metadata registry, users can gain a good understanding of different metadata elements and their definitions and usages from different metadata schemes in distributed information environments (Heery and Wagner 2002).

The CORES registry (h ttp://www.cores-eu.net/registry/) is a good example, currently listing 40 metadata schemes, and supporting searching and browsing by metadata scheme

developer, maintenance agency, element sets, elements, encoding schemes, application profiles, and element usages. In addition, a metadata scheme registry can not only gather different metadata elements of different metadata standards, but also gather a wide range of other metadata-related materials, such as different application profiles, controlled vocabularies, and even ontologies to describe complex information objects. It greatly increases the reusability of different met adata applications. Based on accessing to a metadata registry, an implementer of a specific project, who wants to develop his/her own metadata service, can study how to re-use different well-established metadata elements, and adapt them for his/her local application.

5. Data reuse and integration: This refers to describing information objects by using different elements from different metadata schemes or application profiles (Chan and Zeng 2006b). The Resource Description Framework (RDF) provides a basic platform for integrating different metadata schemes to describe web resources (Heery and Patel 2004). In this sense, it is easy to implement RDF corresponding to an application profile to describe some information objects. Using the RDF platform, the Figure 2.1 shows an example of an instance metadata record corresponding to the BIBLINK application profile. In this example, a combination of metadata elements from DC Metadata Scheme and BIBLINK metadata scheme is used to describe a particular information object.

Note: RDF will be explained in detail in Section 2.5.1.

```
<?xml version="1.0" ?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dc="http://purl.org/dc/elements/1.0/"
 xmlns:bc="http://www.schemas-
 forum.org/registry/schemas/BIBLINK/1.0/bc-ap#">
<rdf:Description about="urn:isbn:0-89887-113-1">
<dc:title>Patrologia Latina Database</dc:title>
<dc:creator>Jacques Paul Migne</dc:creator>
<dc:date>1993</dc:date>
<dc:language>la</dc:language>
<dc:format>CD-ROM</dc:format>
<bc:extent>2 computer laser optical disks; 4 3/4 in</bc:extent>
 <bc:systemRequirements>Multimedia PC 486x or higher, 8mb memory, CD-
 ROM drive, sound card, SVGA 256-colour monitor, Windows 95 or Windows
 3.1</bc:systemRequirements>
<dc:subject rdf:value="Christian literature, Early" bc:subjectScheme="LCSH" />
 <dc:identifier rdf:value="isbn:0-89887-113-1" bc:identifierScheme="URN" />
<bc:placePublication>Cambridge</bc:placePublication>
<dc:publisher>Chadwyck-Healey</dc:publisher>
</rdf:Description>
</rdf:RDF>
```

Figure 2.1: A metadata record corresponding to the BIBLINK application profile (adapted from http://www.schemas-forum.org/registry/schemas/biblink/bc-ap-eg1).

6. Aggregation: This refers to employing a central knowledge base to gather metadata records from different repositories using different metadata standards, converting heterogeneous metadata records into a consistent form, and then developing a range of enhancement services to enrich these metadata records.

For example, ePrint UK project (which aims to build a framework for a general OAI service provider in the UK), noted the heterogeneity of subject metadata harvested from different eprint archives. The project staff adopted different enhancement services to enrich the metadata records harvested. It was demonstrated that these metadata enhancement services could improve the interoperability between the metadata records harvested. These metadata enhancement services include OCLC's subject classification service to automatically create subject metadata, a name authority service to make the formats of different authority data consistent, and a citation analysis service (Martin 2003).

Based on applying methods to improve interoperability between different metadata standards, there are a large number of library portal products widely-used in University libraries, such as WebFeat, MetaLib, SirSi SinlgeSearch, etc. Federated search has been integrated within most of these products enabling library users to search multiple online resources simultaneously from one single uniform user interface, and gain converted consistent metadata records.

However, the retrieved metadata records from a federated search engine are often indexed with different vocabularies and browsed by different subject structures. The development of library portal federated search services was greatly impeded by the heterogeneity of different KOS used by different information resources (McCulloch 2004, and Neuroth and Koch 2001). Based on using established mappings between different metadata schemes, it is also impossible to offer a subject cross-browsing interface in a library portal system. Without a subject cross-browsing service, users, who do not know what term they should use to search, have to jump to each of the identified resources, interact with each of the distinct user interfaces, and browse a range of different subject hierarchies to find relevant terms. It is important to consider methods to develop a subject cross-browsing interface within a library portal interface.

2.2 Theories of organising knowledge

As mentioned in 2.1, a variety of KOS are being used by different information resources. Before considering the ways to improve the interoperability between them, it is important to investigate and analyse various theories to construct these KOS, and the basic syntaxes and semantics of these KOS. Based on differing syntax and semantics of existing KOS,

Zeng and Salaba (2005) presented a typology of different types of KOS, see Figure 2.2.

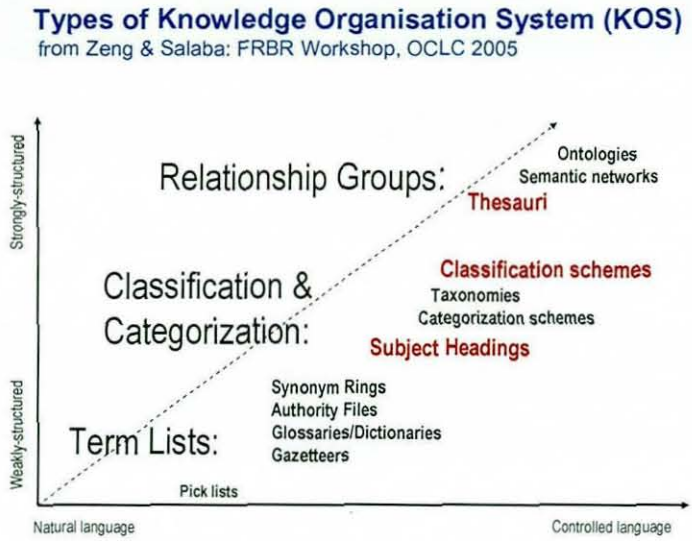


Figure 2.2: Types of KOS (taken from Zeng and Salaba, 2005)

2.2.1 Classification schemes

A huge variety of KOS have been previously developed. The earliest attempt to organise knowledge might be that of the Greek philosopher Aristotle and his followers who represented topics as "places where a rhetorician might look for suggestions treating his theme" (The Unabridged Oxford English Dictionary, 1989). They considered this type of activity as "ontology". Subsequently, the second significant person in the field of organising knowledge was Francis Bacon (Bacon 1640). He applied a hierarchical tree structure to organise and represent different domains of human science knowledge (Welty 1999, pp.155-182). Based on Bacon's work, Diderot and D'Alembert in the mid eighteenth century developed a more elaborate and sophisticated hierarchical structure. This was the first encyclopaedia. Considering the work of Diderot and D'Alembert, other people (e.g. Thomas Jefferson) began to attempt to apply the hierarchies to arrange their library collections. In other words, the Diderot and D'Alembert's theory formed a basis for the development of library classifications. Library classification had become an art of its own by the Nineteenth century. By that time, libraries had an understanding that there was a requirement for better arrangement of the library holdings. One of the earliest

library classifications was the Library of Congress Classification (LCC), which is derived from 18 broad categories similar to the Diderot and D'Alembert's system (Taylor 1999, p.303). Since then, a number of theories for the development of different library classification systems have been produced.

2.2.1.1 The definition of classification

A subject classification is a system "grouping together of similar or related things and the separation of dissimilar or unrelated things and the arrangement of the resulting groups in a logical and helpful sequence" (BS8723-Part3, p.3). In a classification scheme, concepts are represented by classes. Each class has a notation usually represented by a short string of alphabetical and/or numerical characters and/or other symbols (BS8723-Part3, p.6). In a classification system, a notation should be a unified identifier. A notation is used to provide formal and logical access to the shelves in a physical library. In a digital library, a classification scheme can be used to group digital objects under a pre-defined class, and support a subject-based knowledge navigator directing library users to find relevant items. Apart from the notation, each class has a caption to represent the human-readable meaning of this class. A caption, in a digital library, could be used as a subject entry that matches against the users' keyword query. Notes in a classification are usually helpful sources of information to further explain the classes. For example, some classifications provide "scope notes" to give examples of the kinds of topics classed in the number, and a history note to link to the information about the history of the use and meaning of a given classification number.

2.2.1.2 Semantics of a classification

Traditionally, the classes in a classification scheme were set up in a hierarchical manner. The process of setting classes up is called "enumeration" (BS8723-Part4). Apart from the hierarchical relationship between different classes, most classifications provide cross reference from one class to another class. Different relationships could be built in a cross reference. For example, Figure 2.3, which is a MARC record for classification data shows that a Dewey Decimal Classification (DDC) concept "621.252 Pumps" is a broader topic of the concept "621.69 Pumps. Pneumatic pumps" by using the "see

reference". This kind of cross-reference is helpful for connecting different related classes, but is not displayed in the hierarchies.

084 0#$addc$c21 [Dewey Decimal Classification]

153 ##$a621.252$hTechnology (Applied sciences)$hEngineering and allied operations$hApplied physics$kFluid-power technologies$hHydraulic-power technology$hPumps and accumulators$jPumps

553 0#wjga621.69$hTechnology (Applied sciences)$hEngineering and allied operations$hApplied physics$hBlowers, fans, pumps$kPumps$jPumps. Pneumatic pumps$thydraulic pumps

Note: In this figure, under the MARC field 553, the subfield "0#$wjg" means that there is a controlled subfield ($w) that include a "see reference" (j) element. In this "see reference" (j) element, there is a classification number "621.69" representing a broader topic (g) of the concept "621.252".

Figure 2.3: A MARC record to represent a DDC concept

2.2.1.3 Library of Congress Classification

Through long-term continuous revisions, the newest LCC 6th Edition is an essentially enumerative classification system built on 21 classes. Notation for the LCC uses letters mixed with the numerals to represent a concept. In this scheme, there are a large number of common subdivisions and form divisions listed under different classes or sub-classes, which allow local classifiers to increase the specificity of this scheme (Foskett 1996, p.324). However, critics point out that LCC has a lack of notational synthesis capability, which leads to the voluminous nature of this scheme (Chan 1999, p.27). Furthermore, because most of the individual schedules in LCC were developed separately by different subject groups, and "revisions are made independently within special classes and subclasses" (Koch 1999), different main classes may greatly differ in their subject arrangement, and their devices for indicating the subdivisions.

2.2.1.4 Dewey Decimal Classification (DDC)

Another significant contribution to the development of library classification is the Dewey Decimal Classification Scheme (DDC). Melvil Dewey developed a way to specify the subject of books by using the decimal principle. The scheme organises all knowledge into ten main classes from 000 to 999. In his work, subjects were specified more precisely by extending digits for particular subjects. In DDC, a number of standard subdivisions have been developed to "represent recurring non-primary characteristics of a subject and non-topical characteristics that are related to the document itself instead of its main subject" (Scott 2005, p.33). These standard subdivisions are related to the theory, dictionaries, organisations, eductions, etc. of a given subject, and applicable to any class number. The standard subdivisions could be further divided into a more specialised level. In addition, some particular subdivisions were developed tailored to different subject areas in DDC. For example, the "Subdivisions for Arts, for Individual Literature, for Specific Literary Forms" are just used within base numbers specified in 808-890 (Scott 2005, p.85). In the version of DDC 22nd, for example, there are six tables offering subdivisions to expand various parts of DDC.

Compared with the LCC, DDC offered the capability of notational synthesis. In DDC, it is possible to handle complex multi-dimensional concepts found in works by combining notations into a concept (Satija 2007, p.70). For example, 026 (library) and 780 (music) can be synthesised into a compound concept 026.78 (music library). However, Scott (2005, p.33) points out that the capability of notational synthesis in DDC is still very limited. Because of DDC's linear structure and relatively long notation numbers for specific concepts, it is difficult to combine two specific concepts into a compound concept. Gnoli (2005) noted that the combination of DDC concepts only happens in very occasional cases.

2.2.1.5 Universal Decimal Classification (UDC)

With the increase in the number of multi-disciplinary documents, the importance of combining different classes is realised (Brown 1914, Ryward 1975, and Beghtol 1998). As Brown (1939, p.18) noted, "knowledge could be combined and recombined in in-numerable ways".

In parallel with Brown, Otlet and LaFontaine proposed an initiative to develop a universal index to knowledge, "to which different people all over the world would contribute". The development of this universal index occurred between 1854 and 1943 (Foskett 1996, p.281). Based on their idea, the UDC has been developed as an international classification scheme for all areas of knowledge. In the UDC, new concepts are allowed to be created through notational synthesis as needed. The UDC is derived from DDC, but expanded DDC standard subdivisions to about twelve applicable auxiliaries. Because the authors of the UDC realised that the long notation numbers in DDC are unwieldy, they decided to split the UDC notation number by a dot after every third digit unless there was no other indicator appearing in the middle. A number of symbols are used as the facet indicators for the auxiliaries. Considering the lack of capability of combining specific notations into a concept in DDC, the authors of the UDC use colon signs or plus signs to combine two concepts into one (Taylor 2000, p.300). For example, two complex concepts (621.384.634 and 621.318.3) can be synthesised into one concept (621.384.634:621.318.3 Synchrotrons-electromagnets).

One of the main criticisms of the UDC is the complexity of the structure of the scheme (Koch 1999) and high overhead imposed by the work of combining concepts.

2.2.1.6 Faceted classifications

Before introducing the theory of fully faceted classifications, it is worth noting that a number of "non-faceted" classifications incorporate the features of faceted classification (Taylor 1999, p.274). For example, the latest DDC not only provides facets for its "standard subdivisions" and geographic areas to all its classes, but also for individual literature and languages, for racial, ethnic, national groups, for persons, etc (Scott 2005).

However, unlike those traditional enumerative classifications that have rigid linear structures, a fully faceted classification is more flexible, enabling it to accommodate composite subjects, or to handle the relationships between different concepts within it. UDC is recognised as a partial faceted classification, in which a variety of symbols are

used to represent the facets and simple colon/plus signs are used to synthesise the concepts. Based on the idea of synthesising the concepts in the UDC, Perrault (1969) further analysed subject relationships during the synthesis of the classes. These relationships are listed in Table 2.1.

Table 2.1: Relationships used to synthesise the conepts in a faceted classification (Taken from Broughton and Slavic 2006, pp.509-533 with a little revision)

Type of relationship	Example
Addition	Information and education
Range	The field of subjects spanning from education to religion
Coordination	Religion and education in coordinate reciprocal
Comparison phase	Comparison between religion and education
Influence phase	The influence of religion on library management
Bias Phase	Library management for education purpose
Exposition phase	Philosophy as viewed by religion
Sub-grouping	Library management as part of religion

The basic theory of faceted classification is based on Ranganathan's analytico-synthetic view of classification, in which "the process of classification take place when the document is analysed in order to discover its concepts and a notation is then synthesised to express those concepts" (Beghtol 2004). Five basic facets were identified in his Colon Classification, which are personality, matter, energy, space, and time. Based on the five facets, a number of generalised rules were set up and guide the local classifiers to establish the concepts for composite subjects.

The UK Classification Research Group found that the five facets in Ranganathan's classification were too broad, consequently they expanded them to thirteen facets, which are listed below:

Form—time – space – agent – by-product – product – patient (i.e. system operated on) – operation – process – material – property – part – kind – thing

Figure 2.4: Thirteen facets listed in Bliss Classification Edition 2

These 13 facets are used in the development of all classes in Bliss Classification Edition 2 (BC2). These thirteen facets are applied to analyse different subjects with various degrees of complexity and establish compound concepts for these subjects. BC2 uses capital letters and digits to represent the notations. The digits in the notation are used to indicate the common facets. The notations are retroactive, which facilitates the notation synthesis for compound concepts including different facets.

Broughton (2001, pp.67-102) highlighted that facets were always used to contain mutually exclusive groupings of subject terms. In a faceted classification, it is important to consider how to define a logical and consistent order of facets and the order of concepts within each facet. An important concept called "citation order" needs to be considered. This refers to the "order in which preferred-terms or notations are combined in a pre-coordinate indexing system or a classification scheme to form strings representing compound concepts" (Clarke *et al.* 2005). In the schedule of a faceted classification, the sequence of facets should start with the most abstract facet (e.g. form, time) and end up with a tangible facet (e.g. kind, thing), as shown in Figure 2.4. As BS8723-Part3 noted, however, "the citation order of facets in a compound concept should be the reverse of the order in which facets are enumerated in the schedule". For this reason, in BC2, the citation order should be based on the inversed order of the sequence shown in Figure 2.4. "thing - kind - part - property - material - process - operation - system operated on - product - by-product - agent - space - time – form" is the citation order of BC2. Figure 2.5 is an example of a compound concept that is analysed into

different facets, in which concepts within different facets were highlighted by different colours.

Field studies of kinship in hunter-gatherer societies: KSX (agent) J PG (product) 9V (operation)

K9V Field studies (operation)
KPG Kinship (product)
KSX J Hunter-gatherer societies (agent)

Figure 2.5: A compound concept that is analysed into different facets

2.2.1.7 Subject-specific classifications

Apart from various general classifications, a number of subject-specific classification schemes have been developed for different subject needs (Koch 1997). Compared with universal classifications, these subject-specific classifications provide well-defined and up-to-date structure and terminology for their particular usage.

These subject-specific classifications were developed by a variety of independent communities using different classification theories, therefore, they greatly differ in their subject areas, semantic structures, degrees of coordination, etc. This leads to the heterogeneity between these classifications, and between these classifications and some universal classifications.

2.2.2 Subject headings

As mentioned in Section 2.2.1.1, a classification scheme offers a library a logical arrangement of objects based on the subject areas, form of treatment, or the format of the object. The verbal subject information in a classification is mainly based on the captions and scope note, which is insufficient. Many captions cannot provide the real meaning of a class. For example, the caption of DDC concept "004.0151" is "mathematical principle", but the full meaning of this concept should be "mathematical principle in computer

different facets, in which concepts within different facets were highlighted by different colours.

Field studies of kinship in hunter-gatherer societies: KSX (agent) J PG (product) 9V (operation)

K9V Field studies (operation)
KPG Kinship (product)
KSX J Hunter-gatherer societies (agent)

Figure 2.5: A compound concept that is analysed into different facets

2.2.1.7 Subject-specific classifications

Apart from various general classifications, a number of subject-specific classification schemes have been developed for different subject needs (Koch 1997). Compared with universal classifications, these subject-specific classifications provide well-defined and up-to-date structure and terminology for their particular usage.

These subject-specific classifications were developed by a variety of independent communities using different classification theories, therefore, they greatly differ in their subject areas, semantic structures, degrees of coordination, etc. This leads to the heterogeneity between these classifications, and between these classifications and some universal classifications.

2.2.2 Subject headings

As mentioned in Section 2.2.1.1, a classification scheme offers a library a logical arrangement of objects based on the subject areas, form of treatment, or the format of the object. The verbal subject information in a classification is mainly based on the captions and scope note, which is insufficient. Many captions cannot provide the real meaning of a class. For example, the caption of DDC concept "004.0151" is "mathematical principle", but the full meaning of this concept should be "mathematical principle in computer

science". For this reason, other verbal subject access methods should be developed to add another dimensions to the subject classifications used.

Subject headings, defined as a "subject concept term or phrase found in a subject heading list and used in catalogue records" (Taylor 1999, p.251), is an approach to offer verbal subject access capability to a library catalogue. In this sense, a metadata record in a catalogue can be indexed according to both a classification and a subject heading list. For example, a book called "Using XML: a how to do it manual for librarians" is indexed according to DDC (005.72 "Data preparation and representation, record formats") and Library of Congress Subject Headings (three headings, including XML, metadata, and machine-readable bibliographic data). A subject librarian may assign more than one subject heading to an item. This improves the subject accessibility of the item.

2.2.2.1 Library of Congress Subject Heading

One widely-used subject heading system is the Library of Congress Subject Heading (LCSH), which consists of subject terms with references through all subject areas. It is very extensible and adaptable when new subjects are developed (O'Neill and Chan 2003). In LCSH, three categories of headings are used to express the concepts:

- Topical headings are used to represent a discrete, or identifiable concept, such as things, processes, activities, etc;
- Form headings are applied to indicate what a work is, such as film, poetry, books, etc.;
- A name heading is used to indicate the specific name that a work is about. These name headings can be a personal name, a geographical name, or the name of an organisation.

The topical headings are mainly used to facilitate subject access for the purpose of information retrieval. These topical headings are established in different ways, ranging from a single noun to compound phrases. Basically, six syntactical types are used to establish LCSH topical headings, which include (Taylor 1999, p.355):

1. A single word (e.g., "Metadata");

2. A noun preceded by an adjective (e.g., "machine-readable bibliographic data");
3. A noun preceded by another noun used like an adjective (e.g., "catalogue of rare books");
4. A noun connected with another by a preposition (e.g., "user interface in computer system");
5. A noun connected to another by *and* (e.g., "catalogue and index");
6. A phrase or sentence (e.g., "show driving of horse-drawn vehicles").

Like most classification schemes, a number of pre-defined subdivisions are used to further expand the LCSH concepts. These subdivisions include topical and chronological subdivisions specific to particular headings, geographic subdivisions, form subdivisions, free-floating subdivisions, and subdivisions under pattern.

In addition, a number of semantic relationships between different topical headings are defined. The basic relationships, such as narrower, broader, and related, could be represented by a variety of references in the LCSH. However, the basic structure of LCSH is alphabetical. It is difficult for users to be navigated by this kind of alphabetical structure. Blocks (2004, p.9) emphasised that LCSH only contains very shallow hierarchies, which are not suitable for subject browsing. In most subject headings system, many terms have no broader terms, narrower terms or related terms at all.

The original LCSH is a traditional pre-coordinated system, in which controlled terms are pre-combined from several concepts before indexing (Blocks 2004, p.iii). Chan (2000) emphasises the complicated syntax and application rules of LCSH. This makes it difficult for LCSH to be applied outside of the OPAC environment, in particular by current Web search engines. In order to overcome the inherent limitations of LCSH, people are incorporating faceted features into LCSH (O'Neill and Chan 2003).

In a faceted system, terms are divided into different facets. Each of the facets represents one aspect of a subject. As Tudhope et al. (2002) indicate, faceted systems do not apply a

large number of subject headings to describe the subject concepts, but they do group terms depending on a range of basic defined facets to represent subject matter. One related project is FAST Project (Faceted Application of Subject Terminology), underway at OCLC (Dean 2003, pp.337-352). This project has re-designed the LCSH into six different facets. The six facets are topical, geographic, form, period, personal and corporate names. In this case, the rigid hierarchical structure of LCSH has been converted to a range of facets, which makes the new vocabulary more powerful when representing subject content of an information resource.

2.2.3 Thesauri

As mentioned in Section 2.2.2.1, in a pre-coordinated system, the cataloguers have to combine a number of terms in a compound concept. This type of combination, to some extent, is awkward in a digital environment. It is difficult for both computer programmes and human searchers to understand long pre-coordinated concepts including a number of subject terms. With the increase in the number of online databases, it is useful that the coordination of terms could become the responsibility of the searchers instead of the subject cataloguers (Foskett 1996, p.359). Thus, the term "post-coordinated indexing", in which controlled terms are combined to form concepts at the time of searching, is widely-accepted in the development of online bibliographic databases. In this kind of systems, instead of using composite headings, single terms are widely-used to represent the important concepts. End-users can combine some single-term concepts to search documents covering complex subject areas. Nowadays, there are a number of large bibliographic databases, in which the items might cover thousands of complex subjects, thus, post-coordinated methods are extremely useful to support searching large databases (NISO Z39.19 2005, p.37), because various subject terms could be easily combined at the time of both indexing and searching. It is not necessary for the cataloguers to spend a lot of time and effort in pre-adjusting a huge number of subject headings in advance. Boolean operators could be used for the combination.

With this in mind, a number of thesauri have been developed as post-coordinated systems to support information retrieval. According to Aitchison and Bawden (2000, p.1), "a

thesaurus is a vocabulary of a controlled indexing language, formally organised so that 'a *priori'* relationships between concepts are made explicit, to be used in information retrieval systems, ranging from the card catalogue to the Internet". In other words, "A thesaurus is a particular type of controlled vocabulary that represents a formalized description containing a finite set of terms and relations between terms, and frequently also contains information on how the terms are to be applied" (Delphi Group 2002, p.15). There are some key points in devising thesauri, and they are:

- Preferred terms;
- Non-preferred terms;
- Semantic relations: These refer to the semantic relationships between different thesaurus concepts;
- How to apply terms: This refers to the ways to choose appropriate terms, combine the terms into a concept when necessary, select appropriate linguistic forms (singular/plural noun, verb, etc) to represent identified terms, and give the definition of different terms (Aitchison 2000, p.2)

Because of using single terms to represent concepts in most cases (NISO Z39.19 2005, p.40), thesauri are usually applied in particular domains such as law, medicine, and agriculture fields.

2.2.3.1 Semantic relationships

Three semantic relationships are widely-accepted in thesaurus construction. These are equivalence, hierarchy, and related-term relationships (Svenonius 2000, p.156, Broughton 2006, p.117, Aitchison 2000, BS8723-Part 2, and NISO Z39.19 2005). These relationships are applied to link terms with similar or related meanings. In a thesaurus-based bibliographic database system, by using the established semantic relationships, a subject query could be expanded to improve recall of information retrieval. As Block (2004) argues, "the semantic relationships in a thesaurus provide a large potential for information searching and retrieval". A number of information retrieval systems have been developed based on the use of thesauri with relevant algorithms.

However, Soergel (2004) noted that the three widely-used types of relationships between the terms in most of thesauri are still poorly defined and not differentiated enough to support query inference and expansion, and term disambiguation. In this context, different methods are applied to enrich the thesaurus relationships. For example, in the latest thesauri construction standards (BS8723-Part 2 and NISO Z39.19), the relationships are further specified as shown in Table 2.2.

Relationship Type	Example
Equivalency	
Synonymy	UN / United Nations
Lexical variants	pediatrics / paediatrics
Near synonymy	sea water / salt water smoothness / roughness
Hierarchy	
Generic or IsA	birds / parrots
Instance or IsA	sea / Mediterranean Sea
Whole / Part	brain / brain stem
Associative	
Cause / Effect	accident / injury
Process / Agent	velocity measurement / speedometer
Process / Counter-agent	fire / flame retardant
Action / Product	writing / publication
Action / Property	communication / communication skills
Action / Target	teaching / student
Concept or Object / Property	steel alloy / corrosion resistance
Concept or Object/ Origins	water / well
Concept or Object / Measurement Unit or Mechanism	chronometer / minute
Raw material / Product	grapes / wine
Discipline or Field / Object or Practitioner	neonatology / infant

Table 2.2: The thesauri relationships (taken from NISO Z39.19)

Soergel (2004) considered the ways to re-design the traditional representational vocabulary. He designed a range of ontological relationships and rules between different thesauri terms instead of the traditionally defined BT, NT, and RT relationship in order to promote unambiguous query expansion. For example, he determined that *cow* NT *cow milk* should become *cow <hasComponent> cow milk.* In addition, Broughton and Slavic (2006, p.49) investigated the theories to apply faceted analysis into the constructions of

thesauri, in which a faceted classification scheme could form a basis to further create faceted thesaurus, and listed a number of examples to create faceted thesauri.

One of the most typical examples using faceted approach is the Art and Architecture Thesaurus (AAT). This thesaurus is designed to provide terminology for indexing and retrieval of art information. The conceptual scope includes art, architecture, decorative arts, archaeology, material culture, and archival materials. Within this thesaurus, seven facets (associated concepts facet, physical attribute facet, styles and periods facet, agents facet, activities facet, materials facet, and objects facet) were applied, different preferred terms are arrayed in these facets, and therefore, this thesaurus was divided into thirty three hierarchies grouping into these facets.

2.2.4 Ontologies

As mentioned in Section 2.2.3, the structures of traditional KOS, such as subject headings, thesauri, classification, etc, lack rich semantics. They cannot represent the real semantics between different information objects. It is possible to develop richer KOS in terms of semantics and syntaxes. In the field of computing science, professionals are exploring a novel way to develop a rich network of detailed relations that facilitates more powerful concept expansion and semantic reasoning, and even infers new knowledge. This field is called ontology. Generally, an ontology is developed for artificial intelligence modelling and inferencing purposes (Tudhope *et al.* 2006).

The term "ontology" was derived from philosophy. Generally, Koepsell (2000) regards ontology as "the study of categories". Today, ontologies have been become a multi-disciplinary field, and extended to a broader scope, including computing science, library science, and knowledge management (Sowa 2008). A number of important definitions of ontologies have been summarised by Perez and Benjamins (1999) and Ding (2001). They emphasise that most definitions are derived from Gruber's (1993).

> "An ontology is a formal, explicit specification of a shared conceptualization. '*Conceptualization*' refers to an abstract model of phenomena in the world by having identified the relevant concepts of those phenomena. '*Explicit*' means that the type of

> concepts used, and the constraints on their use, are explicitly defined. '*Formal*' refers to the fact that the ontology should be machine readable. '*Shared*' reflects that ontologies should capture consensual knowledge accepted by the communities" (Gruber 1993, p.200).

An ontology can be represented by a range of classes, relations, axioms, instances and functions (Qin and Paling 2001). Compared with other library classification solutions, ontology has the following values summarised by Qin and Paling (2001):

1. "Higher levels of conception of descriptive vocabulary;
2. Deeper semantics for class/subclass and cross-class relationships;
3. Ability to express such concepts and relationships in a description language; and
4. Reusability and "share-ability" of the ontological constructs in heterogeneous systems" (Qin and Paling 2001).

Furthermore, all library KOS focus on describing information subjects and their relations, and are not capable of representing any information object (e.g., a full bibliographic metadata record). In order to describe different library information objects, different metadata schemes are employed. In this sense, an ontology in a digital library is referred to as a combination of subject classifications and metadata sets (Garshol 2005).

In semantic web communities, ontologies are seen as a central component in managing the complexity of the semantic web, and "facilitating knowledge sharing and re-use between agents, be they human or artificial" (Fensel 2001). With this in mind, a spectrum of innovative features and enhanced functionalities have been mentioned in the literature. Warren and Alsmeyer (2005) implied that an ontology is a rich framework for describing information objects and subjects. Typically, an ontology can be used to describe people, organisations, industries, their various relationships and so on. In this context, different types of ontologies have been identified by a number of researchers from different points of view. In the literature, five types of ontologies are commonly agreed. They are described as follows:

1. An up-level ontology refers to a common vocabulary including the basic concepts,

such as things, space, events, time, behaviour, etc, and the relations between them (Wu 2005, Gomez-Perez and Benjamins 1999, and Ding and Funsel 2001).

2. A domain ontology is a domain-specific vocabulary that encompasses the concepts in a given domain (such as medical, agriculture, computer science, etc) and their relationships (Gomez-Perez and Benjamins 1999, and Guarino 1997).
3. A Core ontology is essentially the upper ontology for broad application domains. Different from the upper-level ontology, a core ontology usually analyse the common conceptualisations for certain problem domains behind metadata structure to facilitate data exchanging, mediation, and merging (Doerr *et al.* 2003).
4. A task ontology is based on establishing the relationships between different concepts across domains in order to dynamically solve some problems in a given task (Gomez-Perez and Benjamins 1999, and Guarino 1997).
5. An application ontology includes a selected set of concepts and their relations for modelling a specific application (Gomez-Perez and Benjamins 1999, and Guarino 1997).

Considering the common nature of different ontologies, Fensel et al. (2003, pp.1-6) note that inside an ontology, inference engines can be adopted to reason over instances and classes of an ontology or over different database schemes in an ontology. Thus, it is easier to use inference engines within an ontology to facilitate concept reasoning and expansion.

Because of the extensibility of ontologies, they can be considered as middleware to support interoperability between different systems, or between systems and end-users in networked information environments. For example, Davies *et al.* (2005, pp.68-74) imply that an ontology applied to describe information resources can also be extended to describe the users' information requirements, and support personalisation services. This can be based on expanding the pre-defined ontological classes and properties of an ontology. Another important feature of ontologies mentioned by Davies *et al.* (2005) is that they can offer a user-friendly visualisation and be directly translated into browsable formats. Many ontology tools reproduce the ontology structure in a visual formalism.

2.3 KOS in modern information retrieval systems

Different theories to develop a variety of KOS have been described in Section 2.2. Various KOS have been integrated into modern information retrieval systems to improve both recall and precision. Thus, this section will introduce a number of functions provided by KOS.

Svenonius (2000, p.147) recognised two basic issues when using a KOS in an information system. These refer to referential semantics and relational semantics. Referential semantics are related to the approaches to disambiguating a term having different meanings, and relational semantics are about the methods to establish the relationships between different terms having similar or related meanings. With these considerations in mind, two basic KOS-based subject services are highlighted below.

2.3.1 Term disambiguation

The first KOS-based subject service is related to "term disambiguation" to reduce the ambiguity of a given term and improve the precision of information retrieval. Handling homonyms and polysemes is the most difficult part during a subject search (Tudhope *et al.* 2006). For example, a search using the subject term "mercury" could produce metadata results relevant to the planet, chemistry, and Greek apotheosis. It is necessary to provide users with some further terminological information clarifying this kind of fuzzy terms. Miles (2006) highlighted the distinction between the "concepts" and "terms" in a KOS. As he states, "A concept is an abstract idea or notion; a unit of thought" that should have a clear definition, but a term is just a simple linguistic unit that may have ambiguous meaning. For this reason, it is important to consider how to use the most appropriate terminological information accurately presenting a concept. Svenonius (2000, p.148) described a number of methods to present concepts, which help end-users disambiguate subject terms. These methods are listed as follows:

1. Domain specification: subject information related to the subject coverage of a KOS might form a basis to help users disambiguate the subject terms. For example, in a computer science KOS, the users would know that the term

"weeding" is related to the activity of discarding irrelevant data in a computer system.

2. Qualifier: This method is to add another term in a bracket at the end of the ambiguous subject term. For example, adding (computer collection development) to the term "weeding" could help users understand the specific meaning of a given term;
3. Hierarchy: In this approach, a hierarchical tree where the given term is located is displayed to the users. This enables the user to know the contextual information of this given subject term. For example, the hierarchy information "Dewey 22 > Computer science, information & general works > Library & information sciences > Operations of libraries, archives, information centers > Acquisitions and collection development > Collection development > Weeding" is added to the term "weeding". In this example, through showing users the hierarchy information, users can understand that the term "weeding" is in the field of computer collection development instead of being in the field of agriculture;
4. Notes: A note about the definition of a term is given to the users. For example, the note "any plant that crowds out cultivated plants" is added to the term "weeds";
5. Other conceptual terms with relationships to a given term: In this method, a number of terms with the semantic relationships could be displayed to help users know what the real meaning of a given term is. For example, Figure 2.6 shows a full thesaurus record to describe a concept "cataloguing".

Descriptor: Cataloguing

Used For

- Bibliographic description
- Centralized cataloguing
- Cooperative cataloguing
- Descriptive cataloguing

Broader term

- Information processing

Narrower terms

- Bibliographic control

Related terms

- Bibliography compilation
- Catalogues
- Cataloguing rules
- Documentary analysis
- Documentation
- Information/library cooperation

Figure 2.6: A thesaurus record (Source: http://www.ukat.org.uk/thesaurus/term.php?i=6140)

In different types of KOS, the methods to display the concepts are different. For example, in a classification, where the caption only has partial meaning about the real class, and the relationships are not represented very well, it is more reasonable to combine scope notes with the caption and notation to present the given class.

2.3.2 Query expansion

The second KOS-based subject service is related to the "subject term query expansion" to improve the recall of information retrieval. In many cases, a subject query term initially provided by an end-user might be inadequate to reflect the user's subject requirement. There is a desire for the user to find more related concepts for re-formulating the subject search term. The effort to identify relevant "terms to complement the initially chosen term is an instantiation of query expansion" (Efthimiadis 1996). In the literature, it is widely-agreed that KOS-based semantic relationships play an important role in

facilitating query expansion (Binding and Tudhope 2004, Efthimiadis 1996, Greenberg 2001, and Blocks 2004).

Three types of query expansions (manual, automatic, and interactive) based on various degrees of user interaction are identified in the literature (Greenberg 2001, and Efthimiadis 1996). Manual query expansion is completely based on the users' own intellectual effort, possibly with the help of local library staff. Automatic query expansion depends on developing well-constructed computing algorithms to find all the semantic-relevant terms around a given subject term, and using all these relevant terms to re-conduct the search to find the results. Interactive query expansion is a combination of automatic and manual query expansion methods. After automatic query expansion, the end-user is often required to select other preferred terms, classification notations, narrower terms, broader terms, related terms, etc., from a list for the query re-formulation.

Greenberg (2001) points out the importance of defining the optimal query expansion processing methods, and identifying how to apply the most appropriate query expansion methods in different use scenarios. She found that "synonyms and partial synonyms (SYNs) and narrower terms (NTs) are generally good candidates for automatic query expansion and that related terms (RTs) are *better* candidates for interactive QE" (Greenberg 2001). Based on her finding, Tudhope and Binding (2004) developed a thesaurus-based semantic closeness algorithm for expanding from a given term over the thesaurus-based network to produce a neighbourhood of subject terms semantically close for retrieval purposes. In their work, a parameter is given to each expanded concept, which informs the user how relevant the expanded term and the given term are.

2.3.3 Subject browsing

Another important subject service based on the use of KOS is subject browsing. This focuses on using the intuitive subject tree and familiar terminology from the user's perspective to facilitate subject navigation along conceptual dimensions (Marchionini 1995). This kind of subject service is particularly helpful for novice users who are not familiar with a particular subject terminology language. The benefits of offering a deep

classification structure for subject browsing is widely-emphasised as a desirable complement to a search engine type service in the literature (Koch 1999, Koch 2000, Taylor 2000, p.273, Efthimiadis 1996, and Foskett 1996, p.172). Subject browsing is "more heuristic, interactive, data-driven, and opportunistic" than normal subject-based searching (Cunliffe *et al.* 1997). Today, a number of visualisation tools have been developed to facilitate subject cross-browsing, such as Aduna ClusterMap, Brian, etc (Wester 2007, and Veltman 2004).

In addition, Cunliffe *et al.* (1997) pointed out the limitation of this kind of manual subject browsing, which requires greater interaction efforts from the users. Based on measures of semantic closeness between terms and reasoning over the relationships in a KOS, they proposed a framework to combine query expansion technologies with subject browsing. This is called "subject navigation via query". Traditional query-based retrieval required the users to plan the analytic queries, and a traditional browsing interface requires greater interaction efforts from the users. For this reason, subject navigation via query is a system that can provide a harmonious transition between the query and navigation. In modern information retrieval systems, these different KOS-based subject services should be combined and complement one another to make a smooth balance between the user interaction and computing automation.

2.4 Semantic interoperability between different KOS

Because different communities have adopted different subject classification systems at different levels of semantics, it is impossible to apply or develop a universal KOS to all information collections (Wakes and Nicholson 2001). It is of growing importance to develop relevant methods to improve the compatibility between different KOS so that users are able to conduct their cross-searching and browsing by subject successfully. McCulloch (2004, pp.297-300) notes that this is a very urgent question that needs to be answered as soon as possible. In order to develop an optimal solution to the interoperability problem between different KOS, Cordeiro (2003) points out four basic features that a successful solution should have. These are wide-sharability, adaptability, extensibility and reusability. These four features are basic criteria in every stage of the

development of KOS-based services. To ensure the long-term involvement in improving interoperability, Swan (2004) suggests the development of collaborative approaches to providing subject access between different relevant stakeholders (data providers, service providers, software developers). Chan and Zeng (2002, 2004) investigated the existing theories to improve interoperability between different controlled vocabularies. They (Chan and Zeng 2004) identified the main methods to facilitate interoperability between knowledge organisation systems. All these methods are discussed below:

2.4.1 Concept mapping

In order to develop a unified platform to interoperate two different terminology data sources, and build a bridge connecting heterogeneous terminology resources, a lot of effort has been spent in developing different terminology mapping methods to improve interoperability between KOS. The basic idea of concept mapping consists of identifying different mapping relationships between concepts from different knowledge organisation systems, and establishing the equivalence between two or more concepts (Chan and Zeng 2004). However, it was realised that a number of other factors challenging the mapping work need to be considered before conducting the mapping work. These include the structural models for mapping (BS8723-Part4), the direction of the mapping (BS8723-Part4), the types of mapping relationships (Miles 2004, Chaplan 1995, Vizine-Goetz *et al.* 2004 and Koch *et al.* 2001), the mapping logic (Doerr 2001), collaboration in mapping work between different participants (Koch *et al.* 2001a), the ways in which compound concepts are handled (BS8723-Part4), and the top-level metadata schemes to describe different KOS (Zeng 2008).

2.4.1.1 Structural models for mapping

Chan and Zeng (2003) identified two types of structural models for mapping. One method is direct mapping. This refers to mapping between two controlled vocabularies. In this sense, different direct mapping sets could be constructed between different KOS, which could form a basis offering mappings to different services (see Figure 2.7). The mappings established in this method are usually high-quality because it is relatively

simple to find the relevancy of terms from only two vocabularies. However, Gysenns *et al.* (1994, pp.572-586) indicates that it is impossible to implement this approach in large and distributed information environments, because of the labour intensiveness of this approach.

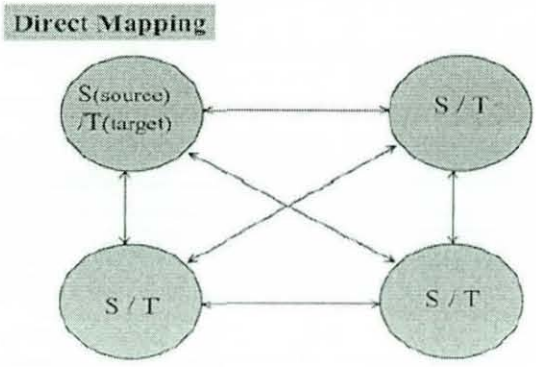


Figure 2.7: Direct mapping (Chan and Zeng 2006)

Another approach called "switch language" has been employed in many terminology mapping projects (Shiri *et al.* 2004, and Neuroth and Koch 2001). The basic idea is derived from the development of Broad System of Ordering (BSO) Initiative, which serves as a universally acceptable switch language that covers all scientific and technical information at the broad level required (Coates *et al.* 1978). In 1971, 111 countries and 62 international organisations achieved a consensus to map their KOS into the BSO (Foskett 1996, p.303). This system is built based on the use of analytico-synthetic classification theory so that within BSO different concepts can be easily synthesised by using the defined facets. However, this scheme is based on using non-hierarchical notations. The BSO structure is not a good basis to provide subject browsing service. Thus, a linear sequence of classes in BSO was further designated (Dahlberg 1980).

Another widely-applied switch language is (DDC) Dewey Decimal Classification (Vizine-Goetz *et al.* 2004). In many projects, DDC was used as a mediator for exchanging terminological information among equivalent terms in different vocabularies. Koch (2001) considered the advantages of DDC over other switch languages, and referred to its online availability, global usage, suitability and functionality, frequency and character of updates, and OCLC's development efforts.

There are some terminology-mapping projects, which try to map different knowledge organisation systems into an upper level ontology. For example, Medical Subject Headings have been mapped into the OpenCyc ontology knowledge base (Reed and Lenat 2002). Also, Alonso and Barriocanal (2006) offered a framework, in which they published different knowledge organisation systems in semantic web enabled form, established inter-mapping relationships between different knowledge organisation systems, and finally mapped all concepts and relationships into the OpenCyc Ontology. In the medical area, the Unified Medical Language System (UMLS) merges concepts from about fifty medical controlled vocabularies into a medical switch language. This metathesaurus also can link to each of its original sources. One of the competitive advantages is that it can create multiple external metathesaural views, allowing different communities of practice to build their own topical maps of the world from pieces of other thesauri and vocabularies (Johnson 2004). It thereby increases reusability.

Mili (1998, pp.204-220) implies that one problem of terminology mapping may be the lack of exact equivalence and varying granularity in different schemes. It is important to identify a variety of mapping relationships before establishing the actual mappings.

2.4.1.2 Mapping relationships

Chaplan (1995, pp.39-61) points out the complexity of identifying mapping relationships. In order to develop an effective path to interoperability, she identified at least 19 different types of relationships between terms in different subject schemes. These 19 relationships were further simplified to develop specific applications (Miles 2004, Doerr 2004, BS8723-Part4, and Vizine-Goetz *et al.* 2004). Broad, narrow, exact, and related relationships are the four basic types of mapping relationships widely-mentioned in the literature. Besides these relationships, Deorr (2004) highlights the importance of using Boolean operators in the process of establishing mapping between a compound concept and two or more post-coordinated concepts. For example, a concept "cat and dog" could be mapped against a combination of two concepts "feline" and "canidae". In this case, the Boolean operator "and" is used to connect these two separate concepts. Vizine-Goetz *et al.* (2004) highlighted the importance of establishing inter-thesaurus mapping by finding

a non-preferred term in one KOS that is equivalent to a preferred term in another KOS, and vice versa.

2.4.1.3 Mapping directions and logics

According to BS8723-Part4, the directions of mapping play a crucial role in supporting different subject services. Mapping KOS A to KOS B could produce different results from mapping KOS B to KOS A. Take Figure 2.8 as an example. In this figure, the mapping from DDC concept 004.6076 to the relevant the UK Archival Thesaurus (UKAT) concepts (computer network an examination) is different from the mapping from the UKAT concepts (computer network an examination) to DDC concept 004.6076.

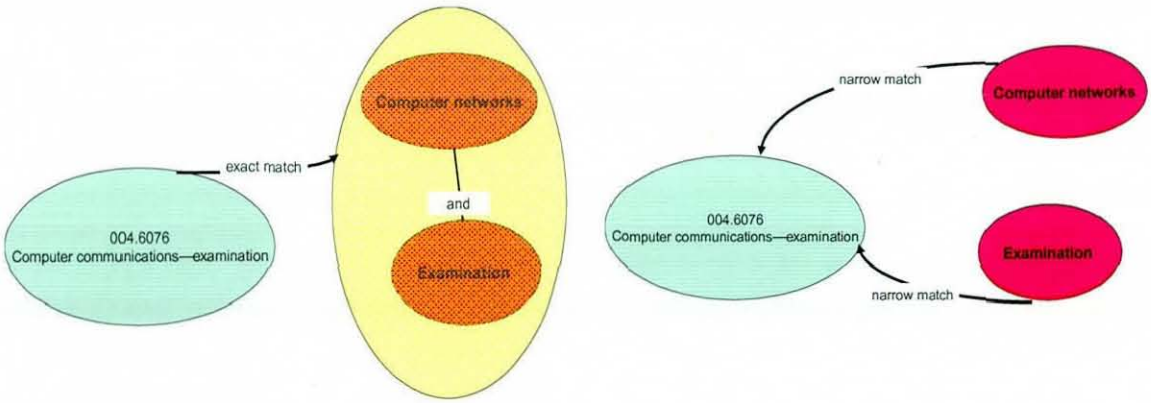


Figure 2.8: Two mappings established by using different mapping directions

2.4.1.4 Co-occurrence mapping

The principle of co-occurrence mappings is based on establishing terminology mapping via subject metadata. Because a metadata record could be indexed according to more than one KOS (Hodge 2000), this approach is based on creating mappings from subject terms from different schemes in the same subject metadata or catalog record (Vizine-Goetz, *et al.* 2004). This method is shown in Figure 2.9. If used effectively, these metadata records can provide a large pool of loosely mapped terms from various controlled vocabularies as well as from freetext. This is a quite cost-effective method. Such sets may be used in mapping between vocabularies or directly for retrieval. However, because of its loose

nature, it is impossible to provide a single subject browsing structure integrating a range of vocabularies. A range of union catalogues, which are indexed by more than one subject vocabulary, could form a basis to discovery co-occurrence mappings.

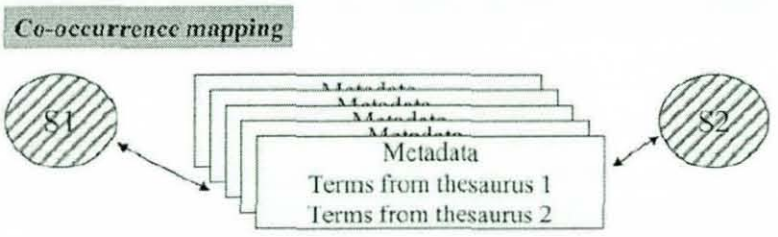


Figure 2.9: Co-occurrence mapping (Chan and Zeng 2004)

It is possible to establish different sets of mappings only from KOS A to KOS B based on different mapping logics. For example, in HILT Phase VI, three different sets of mappings are being constructed for different purposes. High-level mappings are being established. Based on the high level mappings, a user can begin to interact with the top three levels of DDC concepts to find the mapped concepts from MeSH vocabulary, and then the user can jump to the hierarchy of MeSH, and be navigated through the hierarchy of MeSH to find relevant information. Also, deep mappings are being constructed between DDC and MeSH, which can reduce the user-interaction. In parallel, the co-occurrence mappings can be established between DDC and MeSH to improve the recall of information searches.

2.4.2 KOS registry

As mentioned in Section 2.4.1.4, based on the development of metadata elements to characterise different KOS and their mapping sets, a service called "KOS registry" was highlighted (Tudhope and Golub 2008, and Zeng 2008), and a variety of important metadata elements to describe the characteristics of various KOS were recommended for the development of a KOS registry service (Tudhope and Golub 2008, and Zeng 2008). This concept "KOS registry" was defined as "a shared service that lists, describes, and

points to sets of vocabularies, and can hold vocabulary information: member terms, concepts and relationships, provide terminology services, for both human inspection and m2m access" (Golub and Tudhope 2008a). When a subject service needs relevant terminology data to support its subject searching and browsing, it is possible for this subject service to access to a KOS registry, review the metadata records to describe the characteristics of different KOS, and see if there are some appropriate well-developed KOS that are suitable for its subject requirements. If so, instead of developing a new KOS, it is possible to re-use some established KOS for the development of the subject service. In other words, a KOS registry can "facilitate sharing, reusing, and collaboration of different vocabularies" and improve the interoperability between different KOS (Zeng 2008).

In addition, Golud and Tudhope (2008) outlined different architectures for the development of KOS registry services. Importantly, they pointed out that a KOS registry can link to different vocabularies through accessing to various terminology resources in machine-to-machine fashions.

2.4.3 A centralised OAI-PMH metadata repository using a single KOS

The Open Archives Initiative (OAI) was developed as an access protocol to harvest metadata records from e-print archives, and put them into a centralised database. Two important concepts need to be noted in the OAI: data provider and service provider. A data provider is basically an OAI archive that offers its metadata to service providers, and then a service provider, as a third party, harvests the metadata provided by a range of data providers, and provides value-added services to the end-users.

In this OAI model, the metadata from different repositories can be harvested into a central database (also called a service provider), and the service provider can enhance the subject metadata by using a single subject classification system (Cliff 2001). At the centre of this work is the OAI Metadata Harvesting Protocol. In fact, not many metadata repositories provide high-quality subject metadata based on the use of controlled vocabularies. In some cases, metadata repositories are just capable of providing the

simplest service for searching metadata, and there are no controlled vocabularies used in these repositories. In this context, it is possible that a general service provider (SP) can assign subject terms to metadata records. Some terminology services, such as an automatic subject classification service can be used for this purpose (Day 2003).

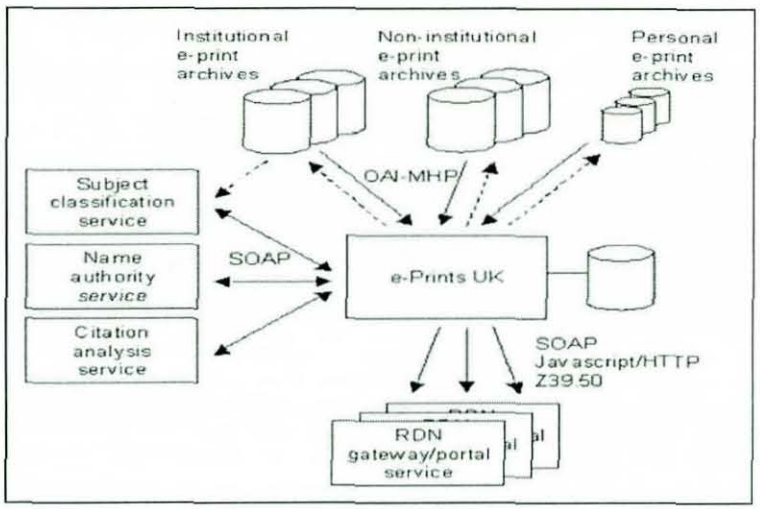


Figure 2.10: OAI-based approach (Taken from Day 2003)

As shown in Figure 2.10, some shared automatic subject indexing software should be used by an OAI service provider to enhance the subject metadata records harvested (Day 2003). Most of these automatic classification systems are based on "string-to-string matching of controlled vocabulary terms and text in documents to be classified", stemming the words, and removing the stop-words in the selected text (Golub 2006). Term weighting, computational linguistics methods, and text extractions are widely-used technologies in different automatic classification systems (Taylor 1999, p.410).

OCLC's Scorpion Project is a project to develop an automatic subject classification system. Hickey (2000, pp.49-56) discusses that the basic theory of this system that is to help librarians automatically create subject metadata for any web-based electronic resources. This system uses a DDC database to connect and analyse information resources, generate candidate DDC codes for subject metadata, after which the librarians will select the useful DDC codes from the candidate DDC codes for local applications. Meanwhile, a selected number of LCSHs have been mapped into this DDC system. Thus, alternatively, librarians can combine some subject headings with the selected DDC codes.

Because new DDC codes frequently are generated with the increase of the number of new terminologies in different domains, the system offers some facilities to automatically update all linked metadata records.

Based on using this kind of automatic subject indexing service, it is possible for an OAI service provider to harvest metadata records from different data providers, and use a single controlled vocabulary based on an automatic subject indexing service to automatically index all the metadata records harvested. This not only improves the effectiveness and efficiency of creating the subject metadata, but also enhances the interoperability between metadata records from different repositories.

2.4.4 Linking through a KOS access protocol

This is a technical approach. In this approach, an integrated environment is created, in which different KOS are located, and a request could query against these different KOS, and get the terminological results from them (Zeng and Chan 2003). This approach is quite similar to most meta-search agents, which search against a number of metadata repositories via an agreed access protocol. In this illustration, each KOS can be treated as a single metadata repository, and a subject term input by an end-user can be a query to cross-search against these KOS-based metadata repositories. The subject queries form a basis to establish temporary mappings.

A number of standards related to access protocols, query language and exchange formats have been developed for this purpose, and these standards are applied in different stages of manipulating the terminology services. These standards will be specifically introduced in the next section.

2.5 KOS and the semantic web

2.5.1 A brief introduction to the semantic web

Berners-Lee *et al.. (2001)* describe the Semantic Web as: "...an extension of the current web in which information is given well-defined meaning, better enabling computers and

people to work in cooperation". In other words, the semantic web provides enhanced information access based on the exploitation of machine-readable metadata, and forms a common platform that allows distributed data to be shared and reused across different applications, information resources and community boundaries. Different layers of the semantic web were proposed to manage different types of complex information. Figure 2.11 shows the layers for the development of semantic web.

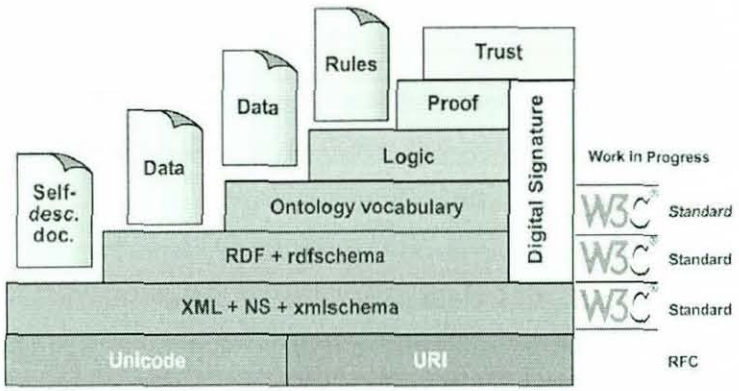


Figure 2.11: Semantic web layered cake (Berners-Lee 2003)

The lowest layer of this architecture is Unicode and URI (Uniform Resource Identifier). The World Wide Web Consortium (W3C) proposed to develop a global network, in which each information resource should be identified by a unique identifier. Thus, each URI plays an important role of identifying and locating a single information resource in this global network. Matthews (2005) and Miles (2005) emphasises the importance of using URI to identify concepts in the network. From the perspective of publishing a controlled vocabulary on the web, Miles and Pérez-Agüera (2005) suggested a URI be assigned to a concept in the vocabulary.

XML and its related standards, such as Namespaces, and Schemas, form a common means for structuring data on the web, but it is difficult to represent complex semantic relationships using XML schemas. Today, a number of metadata standards have been defined using an XML schema, such as MARC21, Dublin Core metadata schema, etc.

The basic layer of the semantic web is standardised by the Resource Description Framework (RDF), which is based on existing standards of XML, URI, and Unicode (Berners-Lee 2003, pp. xi-xiv). Unlike an XML schema, whose structure is tree-based, RDF data is based on a "subject-predicate-object" structure, which allow complex semantics to be represented. "RDF Schema is a simple type modelling language for describing classes of resources and properties between them in the basic RDF model. It provides a simple reasoning framework for inferring types of resources."(Matthew 2005)

Central to the vision of the semantic web are ontologies. Ontologies can play a crucial role in managing the complexity of the semantic web. They are responsible for "facilitating knowledge sharing and re-use between agents, be they human or artificial" (Fensel 2001). Vickery (1997) pointed out the disagreements in developing ontologies across different communities, and emphasised the importance of developing conceptual ontologies in the field of library and information management systems. Arroyo *et al.* (2005) regard knowledge organisation systems as lightweight ontologies, even though they do not have some of the important characteristics of a real formal ontology, such as modality, transitivity, the use of description logics, etc.

A logic and proof layer has been proposed to develop the logics between different ontological elements, and then provide a reasoning service based on the structure of an ontology (Matthew 2005). Thus, inference agents could be employed to conduct logical deductions among different ontological elements.

The trust layer is the final layer of the semantic web. This is proposed to provide assurances of different information resources on the semantic web, and refers to a range of the standardisation activities (Berners-Lee 2003, pp. xi-xiv).

In the semantic web, it is important to convert information into appropriate semantic web-enabled formats, and develop applications to manipulate the encoded data. For this purpose, a number of knowledge representation formats have been developed to encode different types of KOS, and exchange terminological information between different

systems. These formats vary from simple XML-based formats to complicated ontological language, and will be introduced in the next section.

2.5.2 Semantic web-enabled formats (XML, RDF, XTM)

Before introducing the formats to represent a KOS, it is important to look into the semantic web-enabled formats for encoding general information, and then consider the ways to use these formats to represent KOS. In order to exchange information between two systems, information needs to be marked up by markup language. In this case, information can be understood by both computer and humans. As such, XML may be assumed to be a useful language for describing structured information. To some extent, it does provide "support for explicit structural, cardinality and datatyping constraints" (Hunter and Lagoze 2001). However, although XML permits the use of tags to describe objects, and XML schema is capable of defining constraints on the syntax of an XML file (Klein, *et al.* 2003), XML schema provides little support for modelling the semantic relationships between different knowledge elements, in particular in a dynamic information environment.

To overcome these limitations, the Resource Description Framework (RDF) was designed. This is an extremely simple modelling language, based on the use of "subject, predicate, object" triples (e.g. "Snoopy"—is a—dog) (Warren & Alsmeyer 2005, pp.196-205). With this in mind, three object types are defined in RDF: Resource, Property, and Statement (Klein 2003, p.115). Palmer (2001) summarises the advantages of the use of RDF over XML. He highlights unambiguity and decentralisation in the use of RDF, which makes the information created by RDF more understandable for computers, which maximises knowledge sharing. In this case, it is desirable to use RDF to create mappings between distributed RDF resources, and develop a programmatic interface above the RDF mapping data to access and query and distributed resources. For example, by developing an application profile, a range of elements from different schemes can be combined in a RDF document, and an application could be developed above the application profile to cross-search different resources using different schemes.

It is true that RDF permits data modelling, but RDF does not enable us to define any semantics (Warren & Alsmeyer 2005, pp.196-205). Broekstra (2003) found that it is difficult to adopt RDF to declare complex metadata elements. Cross *et al.* note that in a rich knowledge-based environment, RDF is inappropriate for modeling some semantic relationships between different concepts.

For this reason, RDF Schema (RDFS) was designed to support a simple datatyping model for RDF (Palmer 2001). RDFS is used to define vocabularies to model and specify RDF data, in which semantic relations can be defined between different RDF data (Brickley and Guha 2004). RDFS extends RDF by giving external specified semantics to specific resources: rdfs:subclassOf, rdfs:class (Klein 2003, p.117). In other words, RDFS has a few more properties that people and applications can make use of: rdfs:subClassOf and rdfs:subPropertyOf, and some more advanced features allowing people to create properties and classes with ranges and domains for properties. These allow expression that indicates that one class or property is a sub class or sub property of another. In this sense, RDFS combined with RDF can be used to describe simple ontologies or KOS.

RDFS is a very limited language for mainly describing generalization-hierarchical ontologies. It is impossible to use RDFS to define complex logic between different ontological elements for facilitating semantic reasoning services. The semantics of RDFS are defined only in textual descriptions (Klein 2003, p.119). Thus, a more descriptive language, OWL, was developed to provide an exact mapping to expressive description logic.

The W3C website summarises the important features of OWL compared with other ontology languages, as follows,

> "OWL adds more vocabulary for describing properties and classes: among others, relations between classes (e.g. disjointness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties

(e.g. symmetry), and enumerated classes" (McGuinness and van Harmelen 2004).

With this in mind, Palmer (2001) highlights that OWL provides more facilities for modelling data meaning and semantics than XML, RDF, and RDFS. For instance, RDFS can support a generalisation-hierarchical ontology structure, but OWL can provide a richer way to establish a multi-dimension ontological structure, and express and type semantic information.

However, considering the ontology languages developed by W3C, it is obvious that most of these languages do not deal with context at all, assuming that everyone is living in the same open Semantic Web universe. Developing an ontology language that can deal with context becomes crucial. The complexity of the use of OWL is emphasized by Vatant (2003). Also, as Russell and Day (2001) argues, "RDF, RDFS and OWL have been widely criticised for being overly complex, thereby involving high development overheads" (Vatant 2003). For instance, Palmer (2001) admits that publishing RDF data is a tedious process. Thus, another ontology language called XML Topic Map (XTM) has been developed to describe ontologies. It is the subject of international standard ISO/IEC 13250. "XTM allows the formulation of the four key concepts of the XTM model: topics, associations, roles, and occurrences" (Bater 2004). A topic can be very abstract and complicated. In theory, everything could be defined as a topic, and given a URI. The real information resources relevant to the topic can be specified by the <occurrence> element. The relationships between different topics can be defined by the <association> element. In an <association> element, there are two basic child elements. "The <instanceOf> element can specify the class to which this association belongs, and the <member> element can specify all topics that play a given role in an association" (Pepper and Moore 2001). The child element <rolespec> of <member> is used to specify the role played by a member in an association. In Figure 2.12, a "cat" topic can be related to an "animal" topic by the "superclass-subclass" relationship in the <association> element. In this way, we can define a topic "Snoopy" as an instance of a topic "dog" by the relationship "class-instance".

```
<association>
    <instanceOf>
      <subjectIndicatorRef
        xlink:href="http://www.topicmaps.org/xtm/1.0/core.xtm#superclass-
subclass"/>
    </instanceOf>
    <member>
      <roleSpec>
        <subjectIndicatorRef
          xlink:href="http://www.topicmaps.org/xtm/1.0/core.xtm#superclass"/>
      </roleSpec>
      <topicRef xlink:href="#animal"/>
    </member>
    <member>
      <roleSpec>
        <subjectIndicatorRef
          xlink:href="http://www.topicmaps.org/xtm/1.0/core.xtm#subclass"/>
      </roleSpec>
      <topicRef xlink:href="#cat"/>
    </member>
  </association>
<topic id="cat">
    <subjectIdentity>
      <subjectIndicatorRef
        xlink:href="http://www.zoologypark.org/animal.xtm#0001"/>
      <subjectIndicatorRef
        xlink:href="http:/
/www.biologyrepository.net/animal/cats.html"/>
</subjectIdentity>
</topic>
<topic id="animal">
    ...
  </topic>
```

Figure 2.12: Using XTM to describe the relationship between two topics

Topic maps provide a mechanism to declare a range of topics and then a range of other elements (e.g. documents, events, and persons) can be related to the topic (Trippe 2001). Unlike formal ontologies relying on logic, Topic Maps provide no provision for checking logical consistency at any level, not even simple taxonomy integrity (Vatant 2003). In other words, it is difficult to use description logic to define the relations between different elements within a topic, and develop formal logical reasoning service based on the structure of a topic map.

In fact, the Topic Map specification is deliberately intended to support any kind of inconsistency the authors would like to assert. Thus, it is possible to build one topic maps

to provide ready access to resources in any information architecture. The topic maps can contain a number of hierarchies semantically cross-linked to produce a rich, three-dimensional navigational network (Bater 2004, p.133). With this in mind, topic maps can be a tool for creating more complex semantics among related resources with higher reusability. Figure 2.13 gives an example of a topic map.

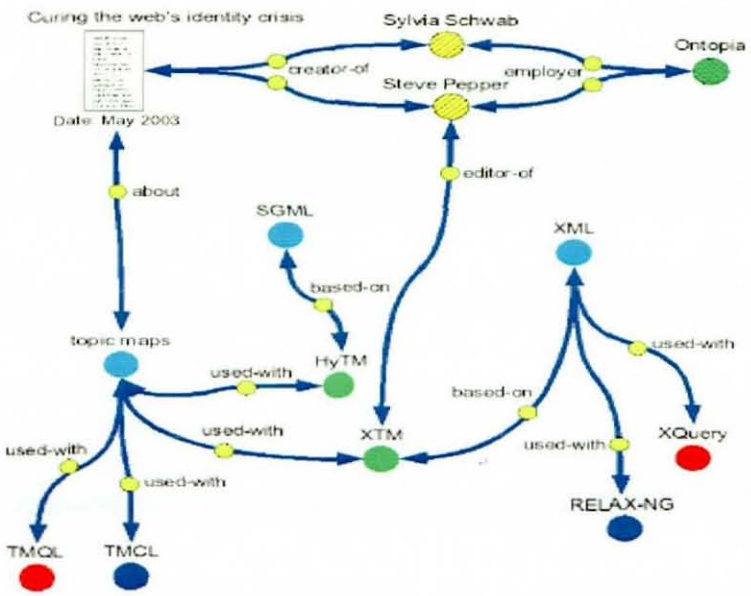


Figure 2.13: An example of a topic map (Taken for Garshol 2003)

In this figure, the example merges the article 'curing the web's identity crisis' into an established networked knowledge structure. Thus, Garshol (2003) notes that ontologists can combine the different metadata that describe information objects and different subject classifications that describe information subjects, into a single topic map.

Comparing XTM and RDF, topic maps start by building a semantic framework from topics and their associations, then 'attach' resources to that framework at appropriate points. They are aimed at humans. In contrast, RDF starts with the resources themselves, then applies 'identities' and 'types' to indicate semantic relationships. It is intended for computer manipulation (Bater 2004).

Nonetheless, XTM lacks consistency in a wide context, because at this moment there is no standard schema language for topic maps. This particularity has attracted serious

critics from the Formal Logic community for obvious theoretical reasons. Due to the lack of standard schema language, as Garshol (2003) argues, "There is substantial risk that the information professionals creating the ontologies may use the vocabulary incorrectly".

To some extent, different communities need different encoding formats for different requirements. This has led to the development of diverse and sometimes competing standards such as XTM and OWL. This has hampered homogeneous development of standards. More essentially, it cannot be said which ontology language is better than the others, what wise ontology developers should do is make them work together, and complement each other to address the specific context they face. As Garshol (2003) states, "it does appear that it is possible to live with both RDF and topic maps, by merging or converting data back and forth between the different representations using simple, declarative, vocabulary-specific mappings". In another case, based on the analysis of the advantages, disadvantages, and functional overlap of the two schemas XMLS and RDFS, Hunter and Lagoze (2001) point out the possibility of combining RDF Schema and XML Schema. They found that RDF Schemas provide more support for semantic modelling but overlook the specification of local data usage constraints, such as structural, cardinality and datatyping constraints. Compared with RDF Schema, XML Schemas provide more support for explicit structural, cardinality and datatyping constraints. Thus, they formulated an approach that clarifies the demarcation of responsibility between RDF and XML Schema, in which the RDF Schemas are adopted in description of semantic definitions, but XML Schemas are used in defining the local usage constraints (Hunter and Lagoze 2001). In this approach, the two schemas can complement each other.

2.5.2 Encoding formats and standards to represent KOS (XML, RDF, XTM)

In Section 2.5.1, a number of formats to make data available in semantic web-enabled ways have been described. These formats have also been further defined to represent various KOS data. This section will introduce four widely-used formats for representing KOS data.

2.5.2.1 MARC21 XML for authority data and for classification data

MAchine Readable Cataloguing Version-21 format (MARC21) linked with AACR2 (Anglo-American Cataloguing Rules) is the widely-used standard to encode the data in the library world. Five types of data can be represented by MARC21 format. These include bibliographic data, authority data, holding data, community information data, and classification data (Taylor 2003, p.72). MARC21 for authority data could be used to represent subject headings data, and thesaurus data, and MARC21 for classification data could be used to represent classification data.

A MARC21 record is a collection of fields. A field may consist of different subfields to further describe the information objects. Each field is identified by a three-digit code, and each subfield is represented by a character code. Different fields and subfields are logically organised to describe different attributes of entries for an information object, such as classification number (153 $a), caption (153 $j), preferred term (155 $a), etc (Baeza-Yates 1999, p.143).

MARC21 has been defined by an XML schema, which greatly facilitates the interchange between different information systems. Today, a number of classification schemes and subject headings have been established in the MARC format, such as DDC, LCC, LCSH, UDC, NLM, etc. Different information systems can reuse existing MARC data for KOS in their own contexts. It is worth noting that MARC21 is a very detailed standard for describing the concepts in a KOS. In Figure 2.14, for example, the concept "HF1

commerce" in the full version of the LCC is described in MARC21 XML. By using different tags with different subfields codes, MARC21 XML can record concepts of a classification scheme in detailed ways. Also, different concepts in a classification scheme can be linked to each other by a number of special tags, such as tag 453 (invalid number tracing) and 553 (valid number tracing). Other tags (e.g. 680 scope note, 684 auxiliary instruction note) can help provide notes to further explain the concepts. The hierarchical relationships within a classification schemes can be represented by the subfield codes $h, and $j in the field 553. In addition, field 750 is used to represent the association with mapped terms in other KOS.

```
<record>
<leader>00834nw 2200193n 4500</leader>
<controlfield tag="001">CF 91000008</controlfield>
<controlfield tag="003">DLC</controlfield>
<controlfield tag="005">19960528091722.0</controlfield>
<controlfield tag="008">910215acaaaaaa</controlfield>
<datafield tag="010" ind1="" ind2="">
<subfield code="a">CF 91000008</subfield>
</datafield>
<datafield tag="040" ind1="" ind2="">
<subfield code="a">DLC</subfield>
<subfield code="c">DLC</subfield>
</datafield>
<datafield tag="084" ind1="0" ind2="">
<subfield code="a">lcc</subfield>
</datafield>
<datafield tag="153" ind1="" ind2="">
<subfield code="a">HF1</subfield>
<subfield code="c">HF6182</subfield>
<subfield code="j">Commerce</subfield>
</datafield>
<datafield tag="453" ind1="0" ind2="">
<subfield code="w">j</subfield>
<subfield code="a">HT684.22</subfield>
<subfield code="h">Communities. Classes. Races</subfield>
<subfield code="h">Classes</subfield>
<subfield code="h">Classes arising from occupation</subfield>
<subfield code="h">Middle class</subfield>
<subfield code="j">Commercial</subfield>
</datafield>
<datafield tag="553" ind1="0" ind2="">
<subfield code="w">l</subfield>
<subfield code="a">HE1</subfield>
<subfield code="c">HE9900</subfield>
<subfield code="j">Transportation and communications</subfield>
<subfield code="t">Commerce</subfield>
</datafield>
<datafield tag="553" ind1="0" ind2="">
<subfield code="w">l</subfield>
<subfield code="a">HE561</subfield>
<subfield code="c">HE971</subfield>
<subfield code="h">Transportation and communications</subfield>
<subfield code="h">Water transportation</subfield>
<subfield code="j">Shipping</subfield>
<subfield code="t">Commerce</subfield>
</datafield>
<datafield tag="553" ind1="0" ind2="">
<subfield code="w">l</subfield>
<subfield code="a">GT6010</subfield>
<subfield code="c">GT6060</subfield>
<subfield code="h">Manners and customs (General)</subfield>
<subfield code="h">Customs relative to special classes. By occupation</subfield>
<subfield code="j">Commercial occupations</subfield>
<subfield code="t">Commerce</subfield>
</datafield>
</record>
```

Figure 2.14: MARC21 for classification data (source: http://www.loc.gov/standards/marcxml/xml/clasmrc.xml)

It is noted that MARC21 for classification data can represent the classification data very accurately, which might be one distinct advantage of MARC21 over other encoding formats. In addition, MARC21 for authority data can be used to represent the thesauri data and subject headings data. In this sense, the field 155 is used to encode preferred

terms in a thesauri, the field 455 is used to encode non-preferred terms, and the field 555 is for representing broader terms, narrower terms, and related terms of a given thesaurus concept.

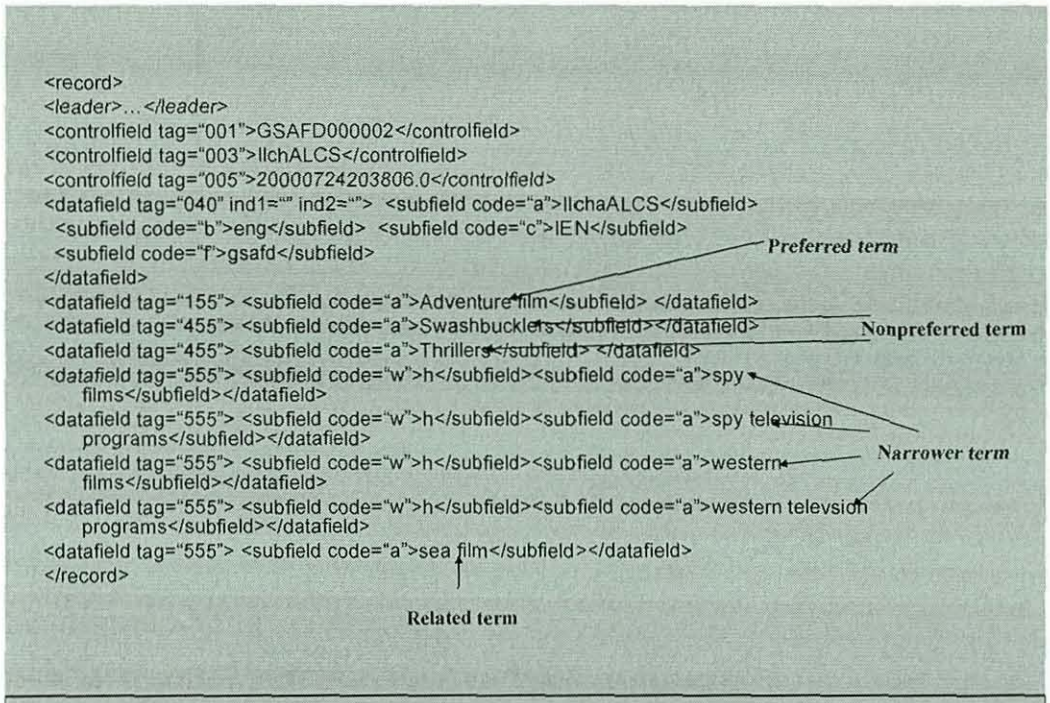


Figure 2.14a: A MARC21 record for authority data

(Source: Si 2007)

However, MARC21 is limited to record simple KOS, such as classification, thesaurus, subject headings, etc., and it cannot represent the ontological relationships for a complex thesaurus, such as whole-part, an instance of, etc. It is impossible to use MARC21 to encode complicated ontologies, and mappings between them. In addition, because MARC21 is only widely-used in the library community, and using this standard to encode knowledge requires a very specific skill, it is not appropriate to use this standard to encode KOS in non-library environments.

2.5.2.2 Zthes XML schema

Zthes is an abstract model for representing thesaurus terms. This model is defined by a XML schema based on a Z39.50 profile for thesaurus navigation (Zthes 2006). In this sense, terminological data can be exchanged from one database to another by using this

defined XML Schema. In this model, a number of thesaurus-based relationships, such as RT, NT, BT, UF, and USE, are made available. Another element defined in this schema is linguistic equivalent relation, which may improve the recall of information retrieval. Also, a term can be distinguished by different term types, e.g. preferred term, non-descriptor, and node label. Figure 2.15 is an example of using this schema to represent a thesaurus term.

```
<?xml version="1.0" encoding="utf-8" ?>
  <Zthes>
  <term>
  <termId>1</termId>
  <termName>Brachiosauridae</termName>
  <termType>PT</termType>
  <termNote>Defined by Wilson and Sereno (1998) as the clade of all organisms more
closely related to _Brachiosaurus_ than to _Saltasaurus_.</termNote>
  <postings>
  <fieldName>title</fieldName>
  <hitCount>23</hitCount>
  </postings>
  <relation>
  <relationType>BT</relationType>
  <termId>2</termId>
  <termName>Titanosauriformes</termName>
  <termType>PT</termType>
  </relation>
  <relation>
  <relationType>NT</relationType>
  <termId>3</termId>
  <termName>Brachiosaurus</termName>
  <termType>PT</termType>
  </relation>
  <relation>
  <relationType>NT</relationType>
  <termId>4</termId>
  <termName>Sauroposeidon</termName>
  <termType>PT</termType>
  </relation>
  </term>
  </Zthes>
```

Figure 2.15: Zthes XML format to represent a term (Source: http://zthes.z3950.org/schema/dino.xml)

Because each terminological record in this schema is based on a term, when putting these records into real applications, it may be necessary to develop algorithms to group relevant terms into a concept by using the defined relationships between terms. Importantly, this XML schema does not define complicated relations between different terms, such as an instance of, whole-part, inter-mapping to another KOS, etc. When representing a complicated thesaurus, it is important to extend this XML schema to define these extra relations in this XML schema. However, this schema is just used to represent thesaurus data, and is not suitable for encoding classifications, and ontologies.

2.5.2.3 Topic Map

In the XML Topic Map community, there have been a number of published subject identifiers within various defined XTM vocabularies to facilitate encoding various KOS, such as faceted classification, thesauri, ontologies, classifications, etc., into the XTM format (Techquila 2003). A number of ontological relationships have been defined in a thesaurus-based XTM vocabulary, such as part-whole, etc. Because of the flexibility of XTM, it would not be difficult to extend the established XTM vocabulary to describe more complicated semantic relationships. Figure 2.16 is an example of how XTM can be used to represent two thesaurus concepts and their relationships.

```
<topic id="0001">
<xtm:instanceOf>
 <xtm:subjectIndicatorRef
xlink:href="http://www.techquila.com/psi/thesaurus/#concept" />
 </xtm:instanceOf>
 <subjectIdentity>
     <resourceRef
xlink:href=http://www.zoologypark.org/animals.xtm#cats />
 </subjectIdentity>
 <baseName>
     <baseNameString>cats</baseNameString>
   <variant>
      <variantName>
        <resourceData>felines</resourceData>
      </variantName>
   </variant>
 </baseName>
</topic>
```

```
<topic id="0012">
<xtm:instanceOf>
 <xtm:subjectIndicatorRef
xlink:href="http://www.techquila.com/psi/thesaurus/#concept" />
 </xtm:instanceOf>
 <subjectIdentity>
     <resourceRef
xlink:href=http://www.zoologypark.org/animals.xtm#mammals />
 </subjectIdentity>
 <baseName>
     <baseNameString>mammals</baseNameString>
</baseName>
</topic>
```

```
<association>
  <instanceOf>
   <subjectIndicatorRef
    xlink:href="http://www.techquila.com/psi/thesaurus/thesaurus.xtm#broader-narrower"/>
  </instanceOf>
  <member>
   <roleSpec>
    <subjectIndicatorRef
     xlink:href=" http://www.techquila.com/psi/thesaurus/thesaurus.xtm#broader"/>
   </roleSpec>
   <topicRef xlink:href="#0012"/>
  </member>
  <member>
   <roleSpec>
    <subjectIndicatorRef
     xlink:href=" http://www.techquila.com/psi/thesaurus/thesaurus.xtm#narrower "/>
   </roleSpec>
   <topicRef xlink:href="#0001"/>
  </member>
 </association>
```

Figure 2.16: XTM to represent thesaurus data (Source: Si 2007)

2.5.2.4 Simple Knowledge Organisation System (SKOS)

One of the most important initiatives in the SWAD-Europe Project is to develop a semantic web-enabled format called SKOS (Simple Knowledge Organisation System) to represent different types of knowledge organisation systems (Miles 2004). SKOS is derived from the full version of RDF(S) and a little OWL. In the SKOS Project, three parts are included. These are SKOS-Core, SKOS-mapping, and SKOS-extension. SKOS-Core is a RDF vocabulary to represent and share different types of knowledge organisation systems on the web. Compared with the term-based Zthes XML Schema, SKOS-Core is a concept-based knowledge representation language. In other words, different terms that have the same meaning can be encoded within a single concept that is identified by a URI. See Figure 2.17. In this case, the terms "freedom" and "liberty" are grouped into a single concept whose identifier is a URI ("http://www.socialsciencepark.org/thesaurus/concept/a092"). Also, different thesaurus relationships can be represented within a concept by using the elements <skos:broader>, <skos:narrower>, and <skos:related>. Thus, different concepts can be linked to each other. In a concept, one <skos:prefLabel> element and several <altLabel> elements can be treated as a synonyms ring to fully express a concept, and different lexical variants of a term can be added into <altLabel> elements.

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#">

<skos:Concept rdf:about= "http://www.socialsciencepark.org/thesaurus/concept/a092">
<skos:prefLabel>freedom</skos:prefLabel>
 <skos:altLabel>liberty </skos:altLabel>
<skos:scopeNote>the rights to control one's own right</skos:scopeNote>
<skos:broader rdf:resource="http://www.socialsciencepark.org/thesaurus/concept/a045"/>
<skos:narrower rdf:resource="http://www.socialsciencepark.org/thesaurus/concept/a0945"/>
<skos:narrower rdf:resource= "http://www.socialsciencepark.org/thesaurus/concept/a0946"/>
<skos:narrower rdf:resource= "http://www.socialsciencepark.org/thesaurus/concept/a097"/>
<skos:related rdf:resource= "http://www.socialsciencepark.org/thesaurus/concept/b056"/>
<skos:inScheme rdf:resource= "http://www.socialsciencepark.org/thesaurus"/>
</skos:Concept>
</rdf:RDF>
```

Figure 2.17: Using SKOS to represent a thesaurus concept (Source: http://www.socialsciencepark.org/thesaurus/concept/a092)

Another important feature of SKOS-Core is its great extensibility derived from RDF data structures. Because a lot of thesauri and subject classification schemes have their own particular structures, some representation formats, such as Zthes XML Schema, MARC21, etc, may not be able to represent these KOSs properly. In this situation, it is desirable to extend SKOS-Core to describe particular subject structures, and even an ontology. In this context, combining SKOS with OWL is very powerful for describing complex terminology data. Figure 2.18 shows creating a <scientificLabel> element that is derived from <altLabel> element.

```
<rdf:Property rdf:ID="scientificLabel">
  <rdfs:label xml:lang="en">scientific label</rdfs:label>
  <rdfs:subPropertyOf rdf:resource="&skos;altlabel"/>
  <skos:definition xml:lang="en">A scientific label for a concept.</skos:definition>
</rdf:protpery>
```

Figure 2.18: Create a new property by using RDFS for SKOS

Because every concept in SKOS format has a URI as its identifier, this provides a basis to create the mappings between different KOSs. In this context, SKOS-Mapping offers a RDF vocabulary to create mapping. There are four basic mapping relationships defined in SKOS-Mapping vocabulary, which are exactMatch, relatedMatch, broadMatch, and narrowMatch. Also, in order to deal with more complicated mapping, the Boolean operators (and, or, not) have been defined in SKOS-Mapping. Due to the extensibility of RDF itself, it is possible to define more mapping relationships for the SKOS-mapping vocabulary. In the SKOS-extension, a number of thesaurus-based ontological relationships are defined, such as broader generic, narrow generic, broader partitive, narrow partitive, related part of, etc. However, SKOS is originally designed to encode the thesaurus data rather than classifications. When using SKOS to represent classification data, it is important to adapt SKOS, and designate the new properties and classes for the important elements in a classification scheme, such as cross-reference, auxiliary instruction note, etc. Based on the ways to encode different semantic relations, OCLC Research (2008) compared three types of encoding formats, which include Zthes XML, MARC21, and SKOS. The findings are listed in Table 2.3.

Label	MARC21	SKOS	Zthes
PT	1XX and 008/09=a,d,f	skos:prefLabel	termType=PT and termName
PT(LE)		skos:prefLabel and xml:lang	relationType=LE and termName
NPT	4XX	skos:altLabel	relationType=UF and termName
NPT(E)	4XX$w/0=a	sub-property skos:altLabel	relationType=X-NPT(E) and termName
NPT(A)	4XX$w/0=d	sub-property skos:altLabel	relationType=X-NPT(A) and termName
BT	5XX$w=g and 5XX$0	skos:broader	relationType=BT and termId
BTG		sub-property skos:broader	relationType=X-BTG and termId
BTI		sub-property skos:broader	relationType=X-BTI and termId
BTP		sub-property skos:broader	relationType=X-BTP and termId
NL	1XX and 008/09=e	[proposed]	termType=NL and termName
NT	5XX$w=g and 5XX$0	skos:narrower	relationType=NT and termId
NTG		sub-property skos:narrower	relationType=X-NTG and termId
NTI		sub-property skos:narrower	relationType=X-NTI and termId
NTP		sub-property skos:narrower	relationType=X-NTP and termId
RT	5XX$w≠g,h and 5XX$0	skos:related	relationType=RT and termId
SN	680	skos:scopeNote	termNote
HN	688	sub-property skos:note	termNote label=HN
MT	7XX and 7XX$0	smap:exactMatch	relationType=X-MT and sourceDb and termId
MT(B)		smap:broadMatch	relationType=X-MTB and sourceDb and termId
MT(N)		smap:narrowMatch	relationType=X-MTN and sourceDb and termId
CN	052,053,065,072,083,087	dct:subject	relationType=X-CN and termId

Table 2.3: Representation format comparison (Source: http://tspilot.oclc.org/resources/overview.pdf)

By assessing these formats, several points can be concluded as follows:

1. The XML-based formats are limited when representing some more complicated thesauri or ontologies and mappings between them, and therefore RDF-based or XTM-based formats are more appropriate to encode complicated vocabularies;
2. It is impractical to use only one representation format to encode all the controlled vocabularies, because different controlled vocabularies have their own structures and syntax, and each of these formats is designed to represent one type of KOS. More importantly, different representation formats can be converted into each other depending on the specific requirements. Van Assem *et al.* (2006) and Summers *et al.* (2008) provided a number of guidelines to convert KOS into SKOS format, and emphasised the importance of extending SKOS to capture specialised features within different KOS.
3. In the KOS community, there is a continuous argument about whether to apply a term-based or concept-based representation formats to encode the KOSs. Most term-based encoding formats are designed to represent thesauri whose basic description element is based on terms, and concept-based formats may be appropriate to represent classification data. However, from the end-users' points of view, they may prefer to use different KOSs as knowledge navigators. This

places emphasis on grouping relevant terms into a concept and representing the concept to the users. Thus, it is important to develop a variety of algorithms and applications to encode KOSs in both term-based and concept-based ways.

A wide range of access protocols and APIs have been developed to support access to encoded data. It is important to look into these different access protocols and APIs, and analyse how these could work with the exchange formats in comprehensive ways.

2.5.3 Access protocols and APIs to access the KOS

The advent of different exchange formats, such as XML, RDF, OWL, etc., forms a basis for systems in distributed information environments to exchange information. It is important to apply standardised protocols or APIs for developing a programmatic interface to query and access the distributed data. As Tudhope mentioned, "Protocols for retrieving vocabulary data are closely linked to representation formats" (Tudhope *et al.* 2006). The semantic web uses different protocol layers to make different web services communicate (Blois *et al.* 2007). Two basic protocols are widely-used to exchange XML-based messages between different applications. One is Simple Object Access Protocol (SOAP), and the other is Representational State Transfer Protocol (REST). In the REST, it is assumed that each information resource can be referenced to using a global identifier (a URI). A RESTful web service processes a URI-based query from a client side to the server. Different parts in the URL string represent a range of input parameters for the server. When the server receives the query, the server should parse and process the string, and return the wrapped results in XML format. The http protocol is the only transport protocol that can be used in a RESTful web service. A SOAP service is similar to a RESTful services, but an XML SOAP envelop should be pre-defined to encode the SOAP request. In this case, each operation on the SOAP server would have an URI. A number of other transport protocols could work with the SOAP web services. In other words, the REST protocol focuses on dealing with the information objects, but the SOAP protocol pays more attention to handling the procedure.

2.5.3.1 Z39.50/SRW/U

Based on the use of these two basic protocols, a number of access protocols have been developed to allow programmatic access to the terminology. The Z39.50 model was an

access protocol designed to provide a basic mechanism for conducting the queries and results for different databases in a distributed information environment (Miller 1999). It is based on a client/server model. In this model, a client is typically a cross-searching service where a user can conduct queries to a range of databases in a network, and a server is a database in the network waiting for the distributed user's query. At the beginning, the protocol originated from library world, but now it has been extended to a wide range of communities, such as e-business, museum, e-government, and so on (Needleman 2000, pp.158-165). Similar to the way application profiles are used in metadata crosswalk, in Z39.50, different profiles have been defined for different purposes. Needleman (2000, pp.158-165) provides three most important applications of the Z39.50 profile:

1. Define the functionality supported;
2. The minimum requirement of searching attribute and attribute combinations;
3. The required record syntaxes.

In addition, a number of interesting profiles have been developed in a variety of contexts. As mentioned in Section 2.5.2.2, one of these profiles is *Zthes,* which "describes an abstract model for representing and searching thesauri using Z39.50" (Taylor 2003, p.25).

Two other protocols, Search/Retrieval Web Service and Search/Retrieval URL service, are derived from Z39.50 to support "people and HTTP agents to query internet databases more seamless" (Morgan 2004), and facilitate the URL-based access mechanisms . SRW is based on the use of SOAP, but SRU is a RESTful protocol. Compared with the Z39.50, these two protocols can provide the returned information in more structured formats, and support web services-based applications. For example, as part of the HILT Project (High Level Thesaurus Project), a machine-to-machine interface was developed to take the form of a SRW plus SOAP-based application interface (toolkit) that will return SKOS data (HILT Phase III).

Common Query Language (CQL) is a powerful Boolean query language developed to support SRW/U. Based on the use of CQL, different functions could be further defined to effectively access and query the KOS data.

2.5.3.2 RDF triple store database, SKOS API and SPARQL protocol

Because of the rich semantic structure in RDF data, it is not suitable for most relational database management system to directly store and manage RDF data. In this situation, a number of applications, called RDF triple storage systems, have been developed to systematically access and store the RDF data/SKOS data, such as JENA, Sesame, etc. In a system, a storage container for RDF, which is usually called a RDF repository, is developed to store the data. A RDF repository can be a relational database or file system (Laborda and Conrad 2006). A number of application layers above the RDF repository could be designed to facilitate the storage of the RDF data in the repository, and expose the well-defined RDF data to a programmatic interface. Figure 2.19 is a diagram to represent the basic technical architecture of a framework for storing the RDF data, called Sesame (Sesame 2008).

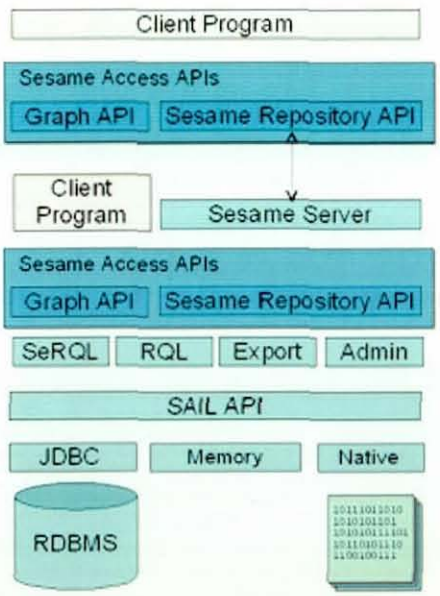


Figure 2.19: Sesame RDF framework (Sesame 2008)

On the top of this architecture, a number of programmatic interfaces that might be based on the use different protocols, query languages and relevant APIs are designed to access and manipulate the RDF data that is stored in the RDF repository, and retrieve the results in commonly-accepted formats, such as XML, JSon, etc.

One of the APIs to manipulate SKOS RDF data is SKOS API. SKOS API was designed to define a range of functions to programmatically access and query SKOS-based terminological data. In this API, several functions could request concepts by given URIs, labels, or relevant expressions, and get terminological information about the concept (SKOS API 2004). Some functions are responsible for requesting a number of semantic relationships for a given KOS, and other functions are responsible for returning the concepts semantically linked by a specific relationship or other linked concepts. "It is possible also to get a set of concepts connected by a relation up to a given path length" (Tudhope and Binding 2005, p.242). Thus, this API could facilitate the in-depth query expansion.

One of the protocols to access and query RDF data is the SPARQL protocol and query language. This is a standard way to "convey SPARQL query to a SPARQL query processing service and return results to the entity that requested them" (Clark 2005). Because the RDF data is based on the use of "subject, verb, object" triples, the SPARQL is designed to conform to this pattern to effectively access the remote RDF storage server. As Chebotko *et al.* (2006) mentions, "SPARQL allows the specification of triple and graph patterns to be matched over RDF data resources". The SPARQL query language seems similar to the SQL, which "uses a simple syntax to specify variables and triplet templates for retrieving information from an RDF store" (Van der ham *et al.* 2006). Figure 2.20 is an example of using the SPARQL to query SKOS data and get non-preferred terms about the concept "Medical transplantations".

```
"PREFIX skos: <http://www.w3.org/2004/02/skos/core#>"+
    "SELECT ?name  " +
    "where{{?concept1 skos:prefLabel \"Medical transplantations\"}
            {?concept1 skos:altLabel ?name }}"
```

Figure 2.20: A simple example of using SPARQL to query SKOS data

Based on this pattern, the query could be further extended to explore more semantic-relevant concepts that are related to the concept "Medical transplantations" by using the SPARQL. Figure 2.21 show a more complicated query to spread out from the given term

"Medical transplantations" over a KOS-based network to produce a range of sibling concepts of the term "Medical transplantations".

```
"PREFIX skos: <http://www.w3.org/2004/02/skos/core#>"+
    "SELECT ?siblingnames  " +
    "where{{?concept1 skos:prefLabel \"Medical transplantations\"}
            {?concept1 skos:broad ?conceptparent }
             {?conceptparent skos:narrow ?conceptsibling }
             {?conceptsibling skos:prefLabel ?siblingnames }        }"
```

Figure 2.21: An example of using SPARQL to query SKOS data

Obviously, SPARQL is a very flexible and extensible query language, and fully supports semantic query expansion within a KOS. One limitation of using SPARQL is noted by Damljanovic *et al.* (2008). This is because SPARQL is mainly used to query an RDF database. SPARQL is oriented towards searching the RDF graph, not towards searching text-based data. This limitation can be traced to the lack of ability in RDF to define local usage constraints. In order to solve this problem, some text-based interfaces need to be developed to work with the SPARQL query language (Damljanovic *et al.* 2004). Patel *et al.* (2005) indicated that a comprehensive combination of various protocols and query languages to access and query a KOS resource would be a trend in future. For example, in order to query a SKOS database, it is preferable to use the SPARQL query language to achieve the query expansion functionality, but use some text-based query interface to conduct text-based searches for getting relevant conceptual terms.

2.6 Terminology service applications

Different theories and technologies to develop a terminology service enhancing the interoperability between different KOS were introduced in the previous sections in this chapter. In this research, a terminology service refers to a web service holding a number of KOS and mappings between them, which can be accessed by a number of other services in machine-to-machine fashions. Thus, this section will briefly introduce a number of established terminology services, describe the developmental principles, and

indicate the common features among these services, and the issues existing in these terminology services.

2.6.1 Renardus Project

The Renardus Project aimed to build a collaborative framework that integrated a range of European subject gateways covering different subject areas, such as Agriculture, Engineering, Earth Sciences, Mathematics, History, Literature, Social Sciences etc, and provide a single access point to these gateways (Heery *et al.* 2001). Basically, these subject gateways are using heterogeneous classification schemes and different metadata elements to organise their information. Two important services provided by the Renardus portal were subject cross-searching and cross-browsing. In order to search heterogeneously indexed resources simultaneously, DCMI (Dublin Core Metadata Initiative) was defined as a common scheme that other metadata schemes are mapped to. However, because other important metadata elements are being used by different subject gateways, a simple DC metadata set cannot support the sophisticated subject cross-searching. In this case, a Renardus metadata application profile was formulated based on adding a number of metadata qualifier to Dublin Core metadata (Neuroth and Koch 2001a).

Furthermore, in order to enable cross-browsing by subject between the resources of the participating gateways, different classification systems were mapped to a DDC, as a central mapping spine (Koch *et al.* 2001). In this situation, end-users can be navigated through a DDC hierarchical tree to find relevant information. In this project, the mappings work between DDC and other classifications were distributed to staff working at different subject gateways. A distributed mapping tool called CarmenX was developed to facilitate the distributed mapping work. No encoding format was used in this project to share vocabularies and mappings with other communities, and other information services cannot access Renardus service in machine-to-machine fasions. For this reason, Renardus service cannot be treated as a terminology service, but it forms a theoretical basis to develop terminology services.

One common issue in the Renardus Project was noted widely in the literature. As Mai (2003, pp.3-12) states, "a general classification system fails to represent the documents at a level of specificity that is required or desired by the users". Thus, in the Renardus Project, when a user find a mapped term from the Renardus subject cross-browsing interface, the user has to leave the Renardus interface, and jump to another subject browsing interface used by a participant subject gateway. Another issue mentioned by Day (2001) is the difficulty in identifying the degree of equivalence between subject terms in different participating gateways. In Renardus, only five types of equivalence relationships were defined: fully equivalence, narrower equivalence, broader equivalence, minor overlap, and major overlap (Neuroth and Koch 2001). In some cases, mapping workers may misunderstand the meanings of these mapping relationships. Therefore, some inconsistency has occurred in the distributed mapping work.

2.6.2 OCLC terminology service

OCLC, as the owner of DDC, has developed a DDC-based terminology service. In this terminology service, a number of other vocabularies have been mapped to DDC, such as LCSH, MeSH, DCMI Type Vocabulary, Newspaper Genre Term Guide, GSAFD, ERIC Thesaurus, etc (Vizine-Goetz *et al.* 2004). Table 2.4 lists a number of vocabularies that are held in the OCLC terminology service database (OCLC Research 2008). The occurrence mappings were established between DDC and a number of vocabularies (e.g., LCSH) via the metadata records in the WorldCat Catalogue. Also, direct mappings were constructed between several vocabularies.

From	To							
Vocabulary	DDC	ERIC	GSAFD	LCC	LCSH	LCSHac	MeSH	NLMC
DDC (Dewey Decimal Classification)				Direct	Direct & Co-occur	Direct & Co-occur	Direct	Direct
ERIC Thesaurus					Direct			
GSAFD (Genre terms for fiction)					Direct	Direct		
LCC (Library of Congress Classification)	Direct							
LCSH (LC Subject Headings)	Direct & Co-occur	Direct	Direct	Co-occur			Direct	
LCSHac (LC Children's Headings)	Direct & Co-occur							
MeSH (Medical Subject Headings)	Direct				Direct			
NLMC (National Library of Medicine Classification)	Direct							

Table 2.4: The mappings within OCLC TS (Vizine-Goetz *et al.* 2004)

All the terminological data within OCLC Terminology Service can be wrapped in MARC21 for authority data by OCLC SOAP server, and be exposed to the OCLC SRW server. In this context, different SRW clients could be set up by different services to query against the OCLC terminology SRW server, and retrieve the MARC21-based terminological data. So far, the OCLC terminology service client has been set up within a number of applications. One of OCLC Terminology Service applications is Microsoft Office 2003. Machines with Office 2003 include a SRW client to query the OCLC terminology service, and retrieve the terminology data.

2.6.3 HILT Terminology Service

Beginning with an investigation of the situation with regard to cross-searching and browsing by subject across a range of communities, services and resource types, the HILT Project aims to develop a JISC IE-based terminology service providing terminological information to different JISC services. In a similar manner to the Renardus project, because a number of different KOS are being used by different JISC services, HILT Project also employed a DDC spine as a mediator for exchanging terminological information between different vocabularies (Wake and Nicholson 2001).

Different KOS and their mappings to DDC were stored in a centralised HILT database. The vocabularies stored in HILT database consist of DDC, AAT, CAB, GCMD, HASSET, IPSV, JACS, LCSH, MeSH, NMR, SCAS, SPEIR, and UNESCO. In order to ensure the usefulness of the mappings between these vocabularies, two different strategies have been used.

In the first approach, HASSET, IPSV, and UNESCO have been mapped to the top three levels of DDC. This is known as high level mapping. Here, a user can begin to interact with the top three levels of DDC concepts to find the mapped concepts from other vocabularies (HASSET, IPSV, and UNESCO), and then the user can go into the hierarchies of other vocabularies (HASSET, IPSV, and UNESCO), and navigate through those hierarchies to find relevant information. Although top level mapping has some limitations due to the lack of 'depth' in the mapping, it is relatively quick to perform. The second approach is to establish deep mappings. In this approach, mappings occur at deeper levels. This is a more labour-intensive approach and the mapping itself is a time consuming process. However, this approach requires less user-interaction to be built into any resulting service as user navigation of hierarchies would be unnecessary.

It is important to develop HILT as a machine to machine (M2M) service so that other JISC services could easily access and query the HILT terminology database. For this reason, a toolkit based on the use of SRW protocol and SOAP protocol was deployed to enable the HILT to be used by third party services. The toolkit consists of an SRW client used to query the HILT SRW server, instructions for making it interact with the HILT, and illustrative users interface routines (which could be customised by local JISC information services) enabling the client to exploit the HILT facilities, and access terminological information. In order to improve the effectiveness of retrieving the data from the HILT database, a SOAP-based HILT request and response handler was developed to manipulate the data within the terminology database, and encode returned terminological results in SKOS format. CQL, as a Boolean text query language, is used to query the HILT terminology server. Figure 2.22 is a diagram to show the basic architecture of the HILT service.

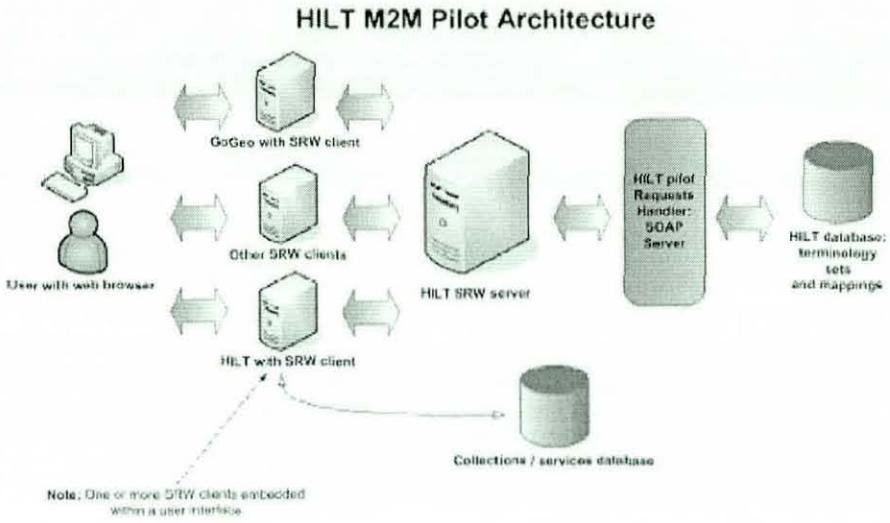


Figure 2.22: HILT M2M architecture (HILT Phase III)

2.6.4 STAR Terminology Service

The Semantic Technologies for Archaeological Resources Project aims to "develop new methods for linking digital archive databases, vocabularies and the associated grey literature, exploiting the potential of a high level, core ontology and natural language processing techniques" (Tudhope *et al.* 2008). In this project, six archaeology thesauri were converted into SKOS format, and stored in a RDF-triple store database application. These thesauri are:

- English Heritage Archaeological Sciences Thesaurus
- English Heritage Evidence Thesaurus
- English Heritage Building Materials Thesaurus
- English Heritage Monument Type Thesaurus
- MDA Object Type Thesaurus
- Alexandria Digital Library Feature Type Thesaurus

A SOAP server has been developed in this web service and the SKOS API with extensions has been used above the SOAP server to provide eight identified functions. These eight functions are listed in Table 2.5 (Binding *et al.* 2007). Among these functions, one of the most important functions is the query expansion function based on traversal of

the semantic relationships in a KOS and measures of semantic closeness of terms. Because the SKOS API and relevant RDF query languages are not strong when processing textual data, a "GetKeywordMatch" function "uses the MySQL full text (Boolean) search, allowing various operators to be used/combined as necessary" (Binding 2008). Currently, there is no mapping created between these archaeological KOS, and the queries form a basis to establish on-the-fly mappings.

GetTopmostConcepts(uri : String) : Concept[] Get the Concepts at the 'top' of the hierarchies for the specified ConceptScheme - returns an array of Concept.
GetConceptScheme(uri : String) : ConceptScheme Get the specified ConceptScheme. Returns a single ConceptScheme object.
GetConceptSchemes() : ConceptScheme[] Get all ConceptSchemes supported by the service. Returns an array of ConceptScheme objects.
GetConcept(uri : String) : Concept[] Get the specified Concept. Returns a single Concept object.
GetAllConceptRelatives(uri :String) : ConceptRelative[] Get all Concepts directly related to the specified Concept. Returns an array of ConceptRelative objects.
GetConceptsMatchingKeyword(keyword : String, includeNPT : Boolean, matchType : MatchTypeEnum) : Concept[] Find lexical term matches across all supported ConceptSchemes. There is an option to include non-preferred terms in the search, and the type of match to look for (exact match, starts with, or contains). Returns an array of Concept objects.
ExpandConceptSimple(uri : String) : ConceptRelative[] A simple Concept expansion function, using internally fixed relationship traversal cost parameters. Returns an array of ConceptRelative objects.
ExpandConcept(uri : String, costBT : Double, costNT : Double, costRT : Double) : ConceptRelative[] A more configurable Concept expansion function, allows specification of traversal costs for the core relationship types. Returns an array of ConceptRelative objects.

Table 2.5: the identified functions in STAR Project (Source: http://reswin1.isd.glam.ac.uk/STAR/SKOS_WS_EH/SKOS_WS.asmx)

In this project, because different metadata schemes are being used to describe different cultural objects from distributed archaeological datasets, and it is important to improve the interoperability between these metadata schemes used by these archaeological datasets, it was decided to map all the cultural objects to an extended version of CIDOC CRM (which is called CIDOC for the English Heritage, CIDOC-EH). The data in different archaeological databases was extracted to RDF format, and the mappings

between the CIDOC-EH and different metadata elements were established manually, and represented by different RDF relationships (Tudhope and Binding 2008). This research project also considered the ways to establish the links between thesaurus concepts and CIDOC-EH data items that the concepts represent. As a result, a specific RDF relationship called "EHP10.is_represented_by" was designed to represent the links. It is worth noting that this "EHP10.is_represented_by" relationship is very flexible, and can indicate different semantic meanings. Figure 2.23 is an example of using the "EHP10.is_represented_by" relationship to establish the link between a thesaurus concept and an information item. In some particular cases, it is possible to use more specific relationships to replace this "EHP10.is_represented_by" relationship.

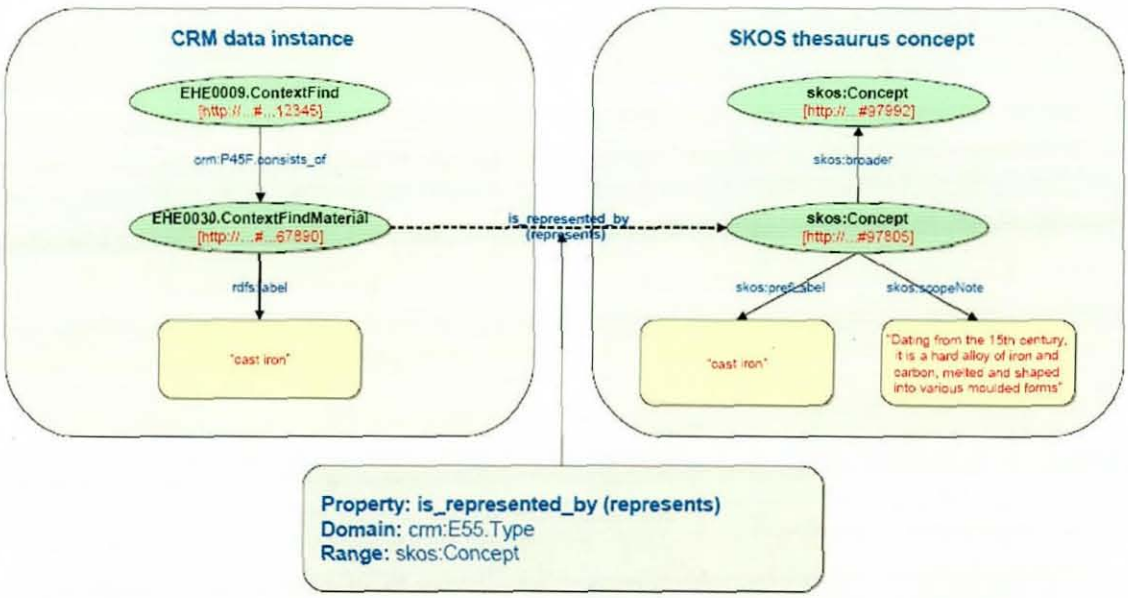


Figure 2.23: A link established between a thesaurus concept and a CRM data item (Tudhope *et al.* 2008)

Furthermore, this project investigated the possibility of establishing the mappings between the CIDOC-EH and different archaeological thesauri. It was found that, in most cases, SKOS representation can provide a cost-effective solution for many annotation, search and browsing-oriented applications, and it is not necessary to develop first order logic supporting automatic inferencing for many subject access services.

2.7 Research questions

Through reviewing relevant methods to develop terminology service-based projects, a number of important issues can be summarised as a basis to develop a fully-fledged terminology service. These issues are listed as follows:

Firstly, it is important to identify a number of KOS relevant to a given context, and understand the basic semantics and syntaxes of different KOS. Different KOS are used to describe different metadata records in different databases, and these KOS differ greatly in their subject areas, degree of coordination, level of granularity, the ways to handle compound concepts, semantic relationships, etc. These semantic and syntactic differences lead to various decisions on designing the ways to query different KOS. For example, it is necessary for a terminology service developer to know that the concept "022.3 buildings" is about the library buildings.

Secondly, based on the understanding of various KOS, it is important to develop appropriate ways to improve the semantic interoperability between different KOS. There is a requirement to exchange terminological information between heterogeneous KOS. As mentioned in Section 2.4, there are a number of methods to enhance the interoperability between different KOS, most of which point to the construction of mappings between different KOS. However, establishing the mappings is a complicated and tedious process. A number of issues need to be considered before creating the mappings, which might include identifying a structural model for mappings, clarifying the elements to be mapped, dealing with compound concepts, defining the mapping relationships and logics, distributing the mapping work to different participants, etc.

Thirdly, when the semantic and syntactic heterogeneous problems have been solved, there are also a number of other factors that challenge interoperability between different KOS. These factors include:

- Multiple formats, such as SKOS, MARC21, Zthes XML, XTM, etc;
- Multiple access protocols, such as SRW, SRU, Z39.50, SPARQL, SKOS API, etc;

- Multiple metadata elements to describe the mappings, such as provenance (source), method (intellectual, co-occurrence, other automatic, etc), perhaps a quality indicator, etc;
- Multiple information systems where different KOS are located, such as MySQL database, Oracle database, Sesame RDF triple store application, etc.

Thus, it is important to design an appropriate technical architecture to provide other services with appropriate terminological data. Two basic philosophies are used to design the technical architecture. One philosophy is based on developing a centralised database on the web creating or collecting different KOS data and mappings, and a number of functions tailored to the designed use scenarios could be developed above the centralised database. Different services could use the functions to query the centralised database, and gain the terminological information. In another approach, it is assumed that different terminological data has been established and published on the web, and a third part could create a programmatic interface to access these data resources.

Fourthly, when other services integrate a terminology service, it is also important to design a number of applicable use scenarios to facilitate the users' interaction with the terminology service. Within these use scenarios, it would be necessary to incorporate some innovative features, such as query expansion algorithms, precise term disambiguation, and full text query. These features could enhance the functionality of a terminology service.

In order to address these issues, a number of research questions are formulated for further research. These are stated as follows:

1. What is the most appropriate approach to improving semantic interoperability between different KOS used by different collections? Is the identified approach to improving semantic interoperability between different KOS appropriate and sufficient to offer a subject cross-browsing service?
2. Which KOS structure is the most appropriate to facilitate subject cross-browsing services?

3. Who should create the mappings for the service? How should the mappings be created?
4. What technologies are suitable for the development of this service?
5. What other services could be integrated with this subject cross-browsing service?
6. What functionality could this subject cross-browsing service offer?

Chapter Three: Research Methodology

This research began by outlining the semantic interoperability problems between different KOS used by different information services. A basic literature review were conducted to clarify the problems, and identify some approaches to solving them. Because most terminology mapping projects are currently in the pilot stage, there is a lack of developmental effort into terminology mapping research in the real world. Hence, the findings of literature review cannot be directly translated into a real system implementation. The following issues related to system implementation cannot be covered by the findings from literature review:

1. The selection of mapping methods: different mapping methods were reviewed in the literature review chapter. However, it is necessary to compare their advantages and disadvantages, and select an appropriate one in the context of library portal systems;
2. The selection of encoding formats: a number of encoding formats to represent terminological information co-exist today. It is important to ascertain which approach is most suitable for represent mappings between concepts from different KOS;
3. Network infrastructure: because the mappings will be established in distributed information environments, identifying a network infrastructure to manage the mapping and distributed KOS is still needed;
4. User scenarios: although some established mappings already exist, a question still remains as to how to maximise the use of mapping for users;

Thus, further research needs to be conducted to address the issues above. First, it is necessary to define research in the field of information science. Goldhor (1972, p.7) emphasises research as an inquiry process that has clearly defined parameters and has an aim. Hernon (1991, p.3) splits the research process into three categories:

1. Discovery or creating of knowledge, or theory building;
2. Testing, confirmation, revision, refutation of knowledge and theory;
3. Investigation of a problem for local decision.

He identified five different activities in any research, which are listed as follows:

1. "Reflective inquiry (identification of a problem, conducting a literature search to place the problem in proper perspective, and formulation of a logical or theoretical framework, objectives, and hypotheses/research questions);
2. Adoption of appropriate procedures;
3. The collection of data;
4. Data analysis; and
5. Presentation of findings and recommendations for future study" (Hernon 1991, p.4).

Glaser and Strauss (1967, p.64) described these research activities as a recursive configuration, which means that the process certainly moves forward, but there is also movement in the opposite direction as succeeding stages uncover data or suggest ideas that revise approaches decided upon or conclusions drawn in earlier stages. Figure 3.1 presents this basic research framework.

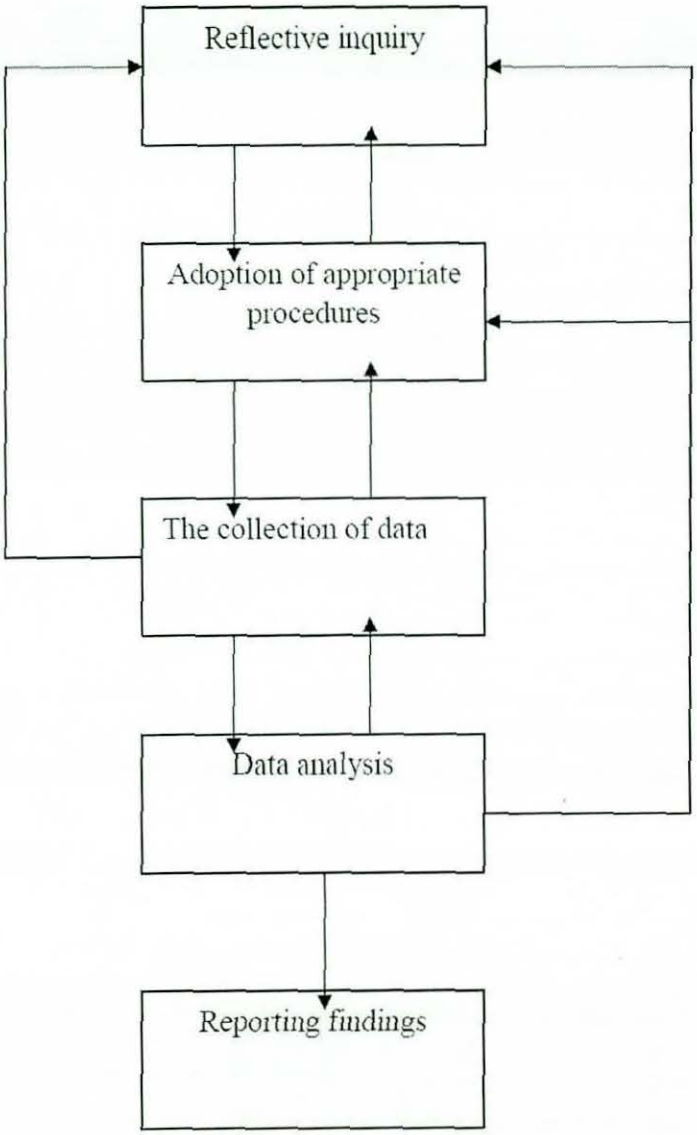


Figure 3.1: The recursive research process

This framework is very general. It can be applied in any research regardless of what the inquirer's worldview is, or what the research methods are used. However, in a specific research activity, it is crucial to consider the nature of research, and select appropriate methods to conduct the research, and map the research into defined research paradigms.

3.1 Research methods

3.1.1 Qualitative research

A paradigm "represents a worldview that defines, for its holder, the nature of the world, the individual's place in it, and the range of possible relationships to that world and its part" (Guba and Lincoln 1994, p.105). In order to select an appropriate paradigm for conducting the research, three research aspects must be noted. These are listed as follows:

- Ontology refers to the study that describes the nature of reality;
- Epistemology is the study that explores the nature of knowledge. In addition, the epistemology refers to the nature of the relationship between the inquirer and the studied topic. It is important to investigate the inquirer and the studied topic they influence each other;
- Methodology is the theory of the research process which "include the assumptions and values that serve as a rationale for research and the standards or criteria that researcher uses for interpreting data and reaching conclusion" (Bailey, 1982, p.26).

Two main paradigms are widely cited in the literature, which include positivist/post-positivist and interpretive/constuctivist. The positivist/post-positivist paradigm considers the world as a single reality, and everything in the world is measurable. A hypothesis, in most cases, is proposed by mathematical formulas expressing functional relations between different variables, and a range of quantitative research activities would be conducted to verify or falsify this hypothesis. This paradigm is widely-accepted in a number of scientific subject areas, such as physics, chemistry, etc. In this paradigm, the inquirer studies the research topic without influencing it or being influenced by it. The findings/knowledge based on the use of the positivist/post-positivist paradigm is objective, apprehendable, and easy to be described in mathematical forms. Quantitative research methods are widely-used in this paradigm to "measure, standardize, and replicate a wide range of observable events" (Struebert Speziale & Carpenter, 2003).

Compared with the positivist/post-positivist paradigm, the interpretive/constuctivist paradigm assumes that there are multiple socially constructed realities in a natural world,

and "knowledge is socially constructed by people active in the research process, and that research is a product of the values of researchers conducting it" (Mertens, 1998). The interpretive/constuctivist paradigm "permits a flexible understanding of complex and evolving social constructs" (Gorman and Clayton 1997, p.31). In this paradigm, inquirer and research topic are inter-linked, and the findings are created during the investigation. Qualitative research methods are largely applied. Gorman and Clayton define qualitative research as:

> "A process of enquiry that draws data from the context in which events occur, in an attempt to describe these occurrences, as a means of determining the process in which events are embedded and the perspectives of those participating in the events, using induction to derive possible explanation based on observed phenomena" (Gorman and Clayton 1997, p.23).

Although the findings based on qualitative research are in-depth and rich (Grover and Glazier 1985, p.247), critics of this paradigm note that "the results from qualitative research may be challenged as invalid by those outside the field of social science" (Audience Research Basics [n.d.]), and biases may occur in the information obtained. Coombes (2001, pp.30) considered these qualitative methods as subjective. Gorman and Clayton (1997, p.29) and Vishnevsky (2008) conclude different roles are played by both qualitative and quantitative research. They notes, "the quantitative researcher is more likely to be predictive, beginning with theory and then collecting; the qualitative researcher is more likely to be interpretive, tending to begin with evidence and then building theory" (Gorman and Clayton 1997, p.29).

It is widely-cited in the literature that the joint use of quantitative and qualitative research methods was effective in many research cases to yield greater insights than just one single research methodology (Robson 1993, p.69 and Mangan *et al.* 2004, p.569). Considering the nature of this research, however, it is difficult to employ quantitative research methods when creating a middleware framework between different KOS used by different information resources. This is because:

1. Ontological issues: This research starts with realising the problems of heterogeneity between different KOS used by different information resources, and aims to develop a framework to overcome this problem. The research realities might change when introducing different theories and technologies. There is no truth in any absolute sense. There is no observable event that could be measured and standardised. A wide range of factors could influence the research situations in various ways. For example, automated mapping technologies could change research realties, upper-level ontologies may facilitate the integration of different KOS, introducing new complicated vocabularies may increase the difficulty in improving the interoperability, and when more intellectual mapping work has been finished, the situation may change again. In addition, different information professionals are likely to have different views on how to develop relevant solutions to this problem. For example, terminology experts may tend to propose certain defined mapping logics and relationships to solve the problem, database developers may select the most appropriate database application to store the terminological data and develop programmatic access to the data, and human-computer-interface experts may consider the ways to design a more user-friendly interface to facilitate the human-computer interaction. This complexity causes the researcher to seek multiple views on the development of a middleware system between different terminologies;
2. Epistemological issues: In this research, the inquirer and the findings are inter-linked. The researcher needs to analyse different factors influencing the inquiry, and try to build up a research context where different problems could be solved. The findings may therefore be based on establishing the research context during the research process. New factors could be introduced, and irrelevant factors could be deleted at any stage of the research process. Different people have different understandings of these influencing factors. Thus, the findings are heavily based on the inquirer's understandings of "the meanings others have about the world" (Creswell 2007, p.21).
3. Methodological issues: This research focuses on building a theoretical framework using and combining different knowledge elements. The available knowledge has

been identified in the literature review, but it is important to use this to build a theory to combine the identified knowledge factors in comprehensive ways to solve the problems. Because different technologies and approaches are complex objects in different applications, it is difficult to use quantitative methods to measure the problems and solutions in mathematical ways.

In this sense, it seems appropriate to use qualitative methods to collect deep knowledge, and build appropriate theories to propose and develop effective middleware between different terminology resources. In the literature, there are a number of comparisons of qualitative research approaches (Creswell 2007, pp.78-79 & Vishnevsky 2004). These deal with different research complexities, which include narrative research, phenomenology, grounded theory, ethnography, and case study. Most of these approaches focus on identifying appropriate end-users (called participants), collecting data from them, and generating the findings. However, it was found in other projects that most of end-users may have trouble understanding the purpose of a terminology mapping services, and they were just concerned with the metadata results finally listed and ranked in an information retrieval system (HILT Phase II). For this reason, it is more reasonable to collect data from a number of persons with known or demonstrable experience and expertise in the development of this kind of services. The collected views of experts would also form a basis to construct a theoretical framework for the development of a middleware system between different terminologies. The theoretical framework will play following roles:

1. To link different controlled vocabularies in comprehensive ways;
2. To input the linked controlled vocabularies into an application;
3. To designate a conversion and transmission programme to process the controlled vocabularies in different formats;
4. To support a range of subject-based services, which include subject cross-searching, cross-browsing, query expansion, term disambiguation, reasoning, etc.

3.1.2 Design research

In the previous section, qualitative research methods based on collecting data from experts were considered to build theories for a middleware between different terminologies. When the theory is built, it is still important to test the theoretical

framework to see if it could meet certain desired goals. Simon (1996) pointed out a number of gaps between the developed theories (inner environment) and their real implementation. In order to bridge the gaps, it is oft en preferred to develop a real implementation based on the constructed theory, and to find out the inconsistencies between theories and implementations, and potential problems of the theoretical framework (Takeda *et al.* 1990). The gaps can be gradually bridged through the design of the system. This is known as design research (Purao 2002). In other words, with the development of design research, the researcher will re-consider the phenomenon and mould it to an adapted version, and also the characteristics of the system that is being built will also become imbued with this changed version of phenomenon. Purao (2002) call this as a hermeneutical process to "involve two kinds of dialogical exchanges, one, between the researcher and the idea of the artifact, and the other, between the researcher and the perception of the phenomenon". Hevner *et al.* (2004) presented a general design cycle to indicate various creative efforts involved in the design research. This is shown in Figure 3.2.

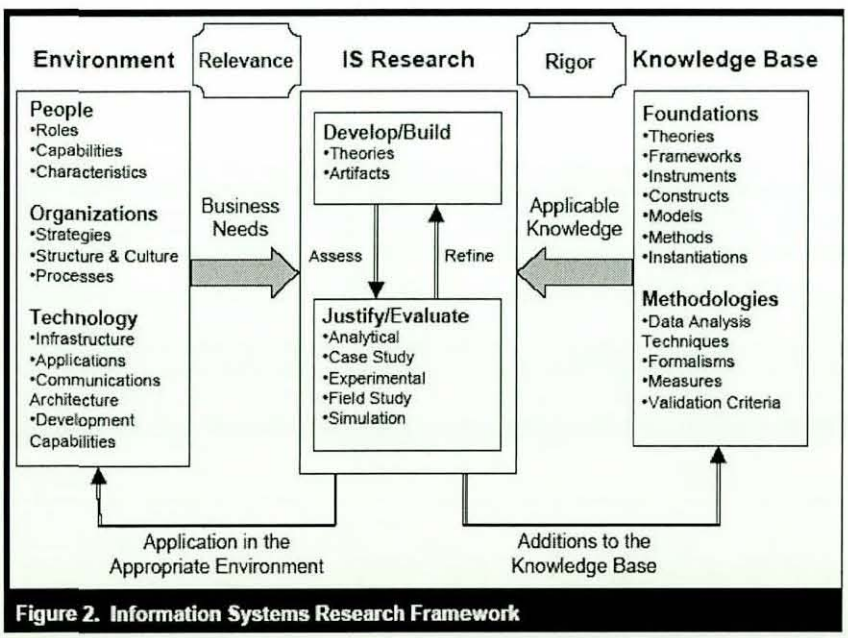


Figure 3.2: Design research (Hevner *et al.* 2004)

Nunamaker and Chen (1990) define design research as a research paradigm called Socio-technologist/Developmentalist and explain how it addresses the valuable contribution of information systems and associate processes to scientific knowledge. In this paradigm, the knowledge building happens through three steps: conceptualisation, formalisation, and development. Figure 3.3 shows basic research process of systems development.

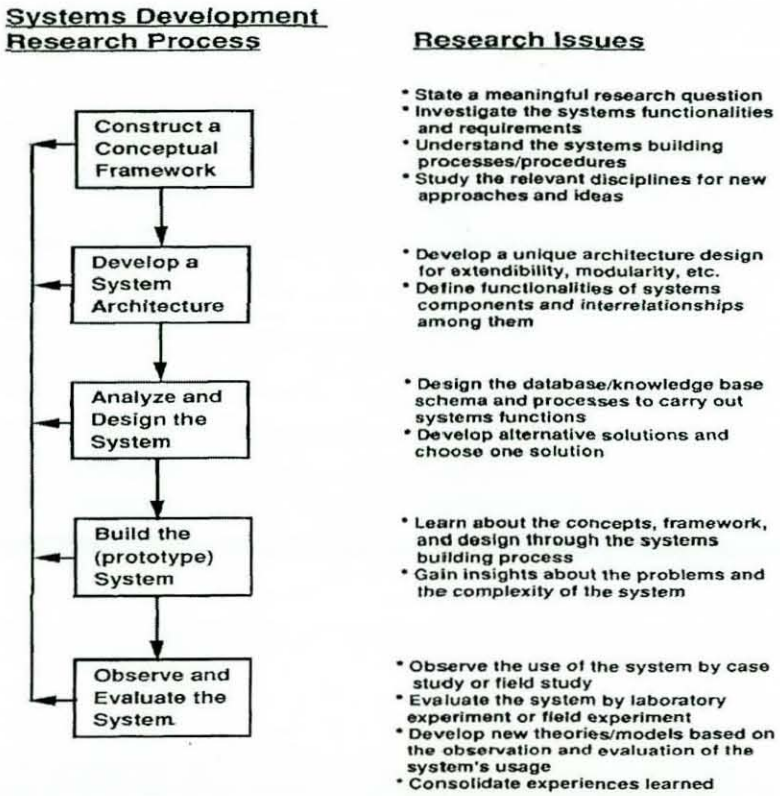


Figure 3.3: The design research paradigm (Nunamaker and Chen 1990)

For example, based on Nunamaker and Chen's (1990) research methods, Vaishnavi *et al.* (1997) developed an operations support system, and found that different repeated processes can generate in-depth understandings.

Considering the nature of this research, two research loops emerge as the recursive research process. In the first loop, the findings of the literature review formed a basis to construct a conceptual framework. There are theories to develop the system architecture in this framework. To some extent, the theories are not mature, because of the lack of the

developmental effort into terminology mapping research in the real world. It is important to gain the suggestions from a number of related experts for the development of an effective system. Hence, these theories were used to formulate a number of interview questions for gaining more in-depth knowledge from the experts. The findings gained from the expert interviews should feed back to re-develop the framework. With this in mind, a research loop emerges as shown in Figure 3.4.

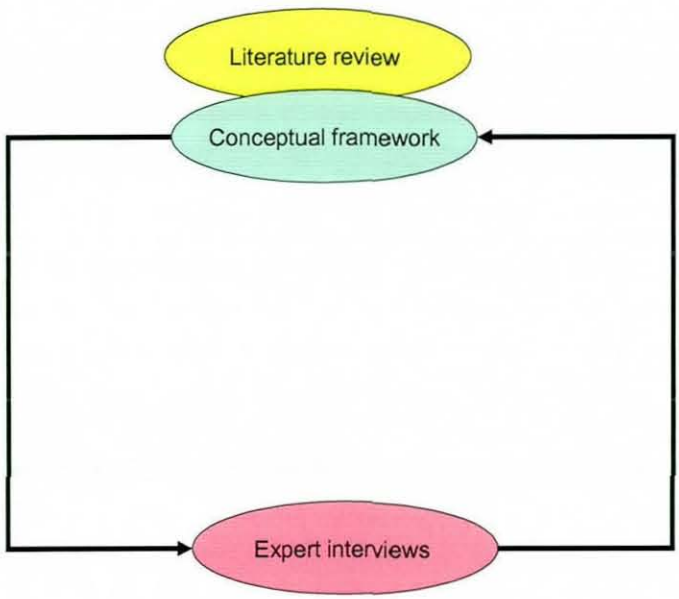


Figure 3.4: The first loop of this research

In the second loop, when the feedback knowledge from the expert interviews had been used to improve the conceptual framework, it was important to use the improved theories to develop a prototype system. This prototype system was applied to assess if the framework can meet certain desired goals, and find out the potential problems within the framework. Hence, a formative evaluation was conducted to test the prototype, and further collect the feedback information from experts for re-improving framework. This loop is displayed in Figure 3.5.

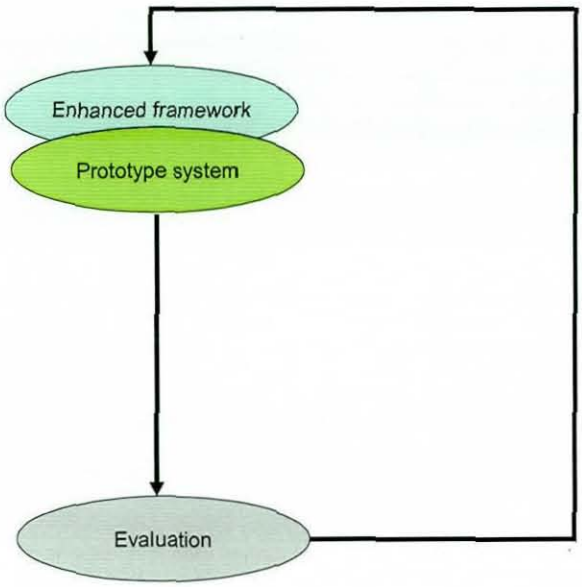


Figure 3.5: The second loop of this research

Considering the prototype system, because there are a huge number of KOS in today's information environments, it is impossible to integrate all these KOS into the designed prototype system. In this research, therefore, it is important to investigate the important features in most of widely-used KOS, and then select some typical KOS to develop a simplified prototype system.

3.1.3 Methods used in this research

Developing a middleware system between different terminologies requires a basic understanding of different KOS being used by different online services. A range of specific data about the characteristics of different KOS is required, such as the number of the concepts within a particular vocabulary, the granularity, the degree of coordination, the subject areas covered, etc. Once the researcher had a basic understanding of different aspects of different vocabularies through the investigation, some typical vocabularies can be selected to develop a simplified prototype. However, due to the rapid expansion of the Web and its changing nature, quantifiable measures of variables of interest are difficult, so a quantitative research methodology is inappropriate. Moreover, the restricted time scale for this research means that it is impractical to choose a large-scale quantitative methodology. In some terminology service-related projects, important survey reports

exploring the amount and quality of different KOS would be helpful. Also, registry services that list the significant vocabularies would be appropriate sources for data collection. Similarly, some professionals' web sites maintain web pages listing all of the online thesauri that they know about that are served by standard protocol thesaurus services. All these provide a basic foundation for conducting a secondary data investigation to understand the landscape of various KOS used by different information providers.

In this context, a joint use of qualitative research, secondary data investigation, and design research methods can yield greater insights than just one single research methodology. Figure 3.6 shows the research methods used in this particular research. The next section will introduce the specific methods to collect the data for creating the conceptual framework.

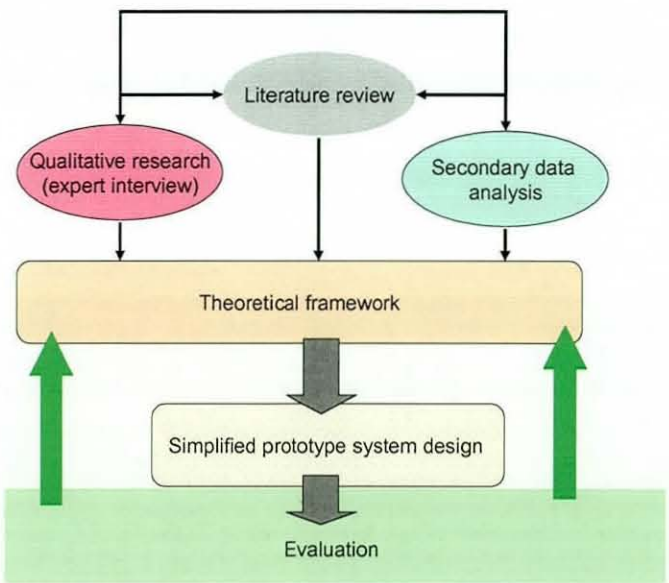


Figure 3.6: Basic research methods used in this research

3.1.4 Limitations of the research methods

Traditionally, design research would undertake several iterations through the research process to gradually improve the prototype system and related theories. However, in this research, only two loops have been undertaken. This is because:

1. The prototype is only a middleware system between different information services. In order to provide the users with the real functions, this prototype system needs to work with a number of other shared information services, such as collection registry, meta-search engine, etc. The end-users may have difficulties in using an immature terminology mapping service without interacting with other services in M2M ways (HILT Phase II). They may not be able to provide useful feedback information to the researcher;
2. This framework is proposed to interact with a number of other shared services in Machine-to-Machine fashion. However, in a real situation, it is impractical to get in touch with all these different services. Many services may not allow this prototype to access them. Also, the time taken to perform the programming to enable the prototype system to interact with other systems in M2M ways is a major barrier. For example, it is very difficult to develop a programmatic interface between the meta-search engine and the middleware system;
3. The main purpose of this framework is to enable the end-users to find item-level metadata results through a subject cross-browsing interface. The lack of relevant metadata records is another barrier of this research for user evaluation. In this research, no metadata repository was used as a resource for the development of the prototype. Thus, this research is only based on two iterations. These two iterations are aimed at collecting different perspectives from different experts to improve the existing theories.

3.2 Data collection methods and findings for investigating different KOS

As mentioned in Section 3.1.3, before interviewing the experts, it was important to investigate different KOS used by different information services, and analyse the characteristics of these KOS. This is intended to help the researcher formulate the most appropriate interview questions and identify appropriate resources for the development of a prototype system.

Koch (1997) investigated the typology of different controlled vocabularies used by different online services. Aitchison and Bawden (2000) explained the important features in the construction of thesauri. Warner (2004 p.179) identified some types of organization systems, including classifications, gazetteers, lexical databases, synonym rings, authority files, taxonomies and thesauri. In addition, some less-traditional knowledge organization systems have been developed representing complex relationships among information objects, such as ontologies and faceted approaches. It is important in this investigation to be based on some basic principles that can classify these different knowledge organisation systems existing on the Internet. This research will be based on a taxonomy of knowledge organisation sources developed by Hodge (2000), and will summarise the important features of different classification systems. In this taxonomy, specific KOS types are grouped into general categories. Figure 3.7 is the basic structure of this taxonomy.

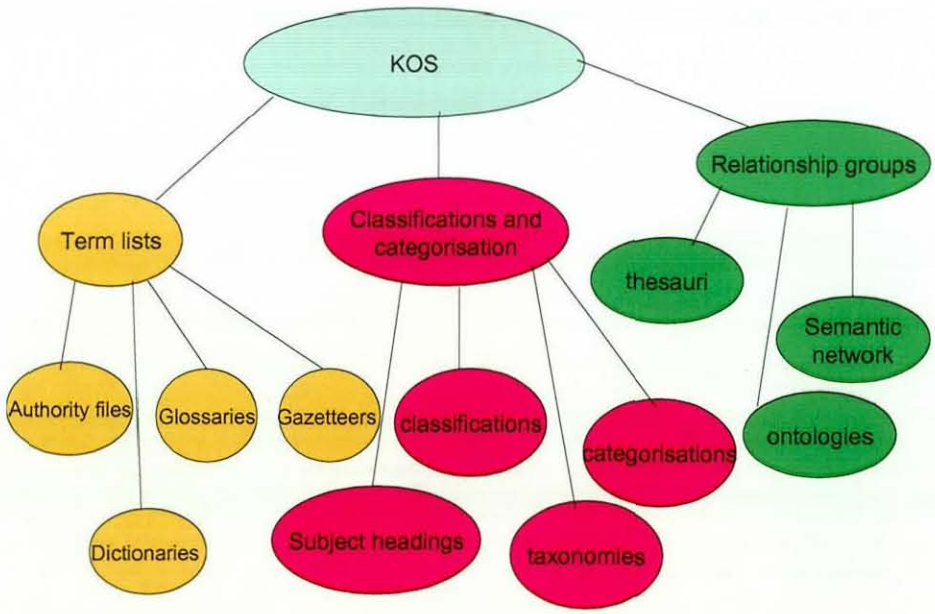


Figure 3.7: Taxonomy of KOS (Derived from Hodge 2000)

A number of resources were fou nd, listing a wide range of knowledge organisation systems used by different online services. These resources provide a basic background for this investigation, and are shown as follows:

1. Controlled vocabularies, thesauri and classification systems available in the WWW. DC Subject http://www.lub.lu.se/metadata/subject-help.html;
2. Services survey result in HILT Project Phase II http://hilt.cdlr.strath.ac.uk/hilt2web/finalreport.htm;
3. SPECTRUM Terminology Bank. http://www.mda.org.uk/spectrum-terminology/termbank.htm;
4. Synapse (now: Factiva) Taxonomy Warehouse http://www.taxonomywarehouse.com/ ;
5. Leonard Will's web site include a number of widely-used vocabularies with the detailed information http://www.willpowerinfo.co.uk/thesbibl.htm#lists;
6. OCLC Terminology Service http://www.oclc.org/terminologies/default.htm;

Through using these resources, 55 English language vocabularies were identified for the investigation. It was important to develop a set of metadata elements for recording the main characteristics of these vocabularies. There are a number of metadata elements that were suggested to record different specific features of a vocabulary (Zeng 2008). These elements might include provenances, creation and revision date, the number of terms included in a vocabulary, flexibility for integration with other KOS, semantic relationships within a given vocabulary, the degree of coordination, granularity, accessibility of a given vocabulary, available formats, copyright and license, etc. However, this particular investigation only aims to explore the possibility of mapping one vocabulary to another. Thus, it is important to collect the descriptive information about the semantic features of different vocabularies, and the usage of the vocabularies by different other information services. The main elements included the subject areas, degree of coordination, type of vocabulary, granularity, the services using the vocabularies, and the number of concepts included. Because this research only focuses on developing the subject cross-browsing services based on English language, non-English vocabularies were excluded in this investigation.

In conclusion, the investigator began by identifying the important KOS through looking into a number of terminology registries, accessing the identified KOS to gain a further understanding, and finally recording the main characteristics of these KOS. Appendix 1 presents the main data from this investigation.

3.3 Data collection methods for creating the framework

A number of technologies, standards, and methods used to develop a middleware system between different terminologies were identified at the end of Chapter 2. These form a basis to define a theoretical framework for the middleware system.

3.3.1 Semi-structured expert interview

As mentioned in Section 3.1.2, applying qualitative research methods to collect the views from various experts is appropriate. In the field of qualitative research methodology, three main methods are widely accepted and applied, observations, interviews, and focus groups (Powell 1986, p.8 and Powell 1991). They will be introduced briefly as follow:

1. Observation studies involve the systematic recording of observable phenomena or behaviour in a natural setting, which provides useful insights into unconscious behaviour and how this might relate to self-perceptions of those involved in an event. It can reflect hidden attitudes or views (Gorman and Clayton 1997, p.44).
2. A focus group is defined as "a selected set of people used to test and evaluate a concept or product" (LevelTen Knowledgebase [n.d.]). Focus group interviews are good at exploring in depth the feelings and beliefs people hold and to "learn how these feelings shape overt behaviour" (Goldman and McDonald 1987, p.7).
3. Interviewing can obtain detailed, in-depth information from subjects who know a great deal about their personal perceptions of events, processes and environments (Gorman and Clayton 1997, p.44).

Observation focuses on working with participants rather than experts. In other words, it is impractical to observe experts' activities when they are developing or programming their terminology services. This approach was therefore impractical during the initial iteration through the design research loop. Focus groups can be used for determining the

perception, feelings, and thinking of gathered experts about the problems, solutions, products, services to gain understanding about a given topic (Krueger and Casey 2000, pp.7-9). Also, the intensive interactions between different group members during the discussion could produce new data and ideas. However, a number of undesirable effects of focus groups have been noted (Evmorfopoulou 2000). In this particular research, the multi-disciplinary nature of developing a middleware system between terminologies requires the data to be collected from a variety of experts. If there are a variety of experts in a focus group, it is possible that some reserved group members, who have great knowledge in the field of developing terminology services, may be reluctant to talk. This may lead to bias. In addition, because experts are the leading researchers in the studied field, and busy doing their research day by day, it is impractical to gather all the leading persons at the same time to answer the questions asked.

With this in mind, face-to-face semi-structured interviews, as a data collection method to explore ideas, probe responses, and look into motives and the feelings, was selected. There are a number of advantages over other methods:

1. A variety of experts could offer different suggestions from different perspectives for the development of the middleware;
2. Face-to-face interviews are very flexible. The experts could be encouraged to have more dialogue with the interviewer, and elicit more in-depth information about their understandings;
3. Because of the intensive interaction between the interviewer and interviewee, it is easier for the interviewer to guide the interviewee to focus on the important questions, and avoid bias.

3.3.2 Heterogeneity sampling

When sampling strategies are described, a key distinction is made between probability and non-probability samples. Probability sampling is generally held to be the most rigorous approach to sampling for statistical research, but is largely inappropriate for qualitative research. In a probability sample, elements in the population are chosen at random and have a known probability of selection (Ritchie and Lewis 2003, p.78). In non-probability sample, units are deliberately selected to reflect particular features of

groups within the sample population (Ritchie and Lewis 2003, p.78). Traditionally, in qualitative research, non-probability sampling is used to select the participants who have experienced the identified process (Crewell 2007, p.63). Based on interacting with these selected participants, the researchers would gain a general explanation of a process or an action. In many cases, theories were discovered through gradually interviewing the participants. This approach is called "grounded theory". In grounded theory research, the guidelines for determining the sample sizes rely on the concept of "saturation", which means a point at which no new findings are found in the data (Guest *et al.* 2006). Crewell (2007, p.78) noted that interviewing 20-60 individuals can basically reach the saturation for most grounded theory research.

In this research, however, the researcher gained most theories for the development of the proposed framework from the literature review, and it is not necessary to find new knowledge elements based on interviewing experts for "saturation". In other words, the main purpose of interviewing experts is not to develop the theories through gradually conducting expert interviews, but to understand how to use the appropriate theories to establish the framework.

Selecting appropriate experts, as non-probability sampling, involves the gathering of a sample of experts with known or demonstrable expertise in some subject areas related to issues, such as terminology mapping, subject cross-searching, semantic web, etc. This research focuses on gathering a broad spectrum of ideas instead of identifying the typical ones (Research Centre for SRM [n.d.]). Heterogeneity sampling, as a particular strategy, was introduced. In this sampling strategy, it is the intention to sample ideas not people (Research Centre for SRM [n.d.]). Considering the nature of this research, heterogeneity sampling is extremely appropriate. Different types of experts, who are involved in different research projects, and have different ideas to develop a terminology services, were sampled. These experts included:

1. Experts responsible for developing and managing library portal products: They have a good understanding of how these work, how a subject cross-searching service is integrated into the library portal, and what subject requirements are

helpful for different end-users. They might be able to indicate the potential needs for the development of the middleware system to facilitate subject cross-browsing.

2. Terminology experts: They have strong knowledge of different controlled vocabularies, and have experience of the development of methods to improve the interoperability between different KOS.
3. Experts in the development of the semantic web: They understand different technologies related to developing the terminology servers, and are familiar with different semantic web-enabled formats to encode terminological data.

Table 3.1 outlines the specific experts who were interviewed with their experience. All the interview data has been analysed to form a basis for developing the middleware system. The detailed interview data analysis is presented in Chapter 4.

Table 3.1: The details of the interviewees

Interviewee	Experience	Type of experts
Interviewee 1: A local librarian responsible for maintaining the library portal.	Ex Libris MetaLib: a project related to developing library portal and federated search	Experts responsible for managing library portal products
Interviewee 2: A local librarian responsible for maintaining the library portal.	SirSi Subject Room: a project related to developing library portal and federated search.	Experts responsible for managing library portal products
Interviewee 3: A research officer working in Ex Libris	Ex Libris MetaLib: a project related to developing library portal and federated search	Experts responsible for developing library portal products
Interviewee 4: A research officer	Renardus Project: it is research project focusing on developing subject cross-browsing services through terminology mapping	• Terminology expert • Experts in the field of developing semantic web • Experts responsible for developing library portal products
Interviewee 5: A research lecturer in the field of information science	HILT Project: a project that aims to develop a terminology service for the JISC Information Environment	• Terminology expert • Experts in the field of developing semantic web
Interviewee 6: A professor in the field of information science	STAR Project: a project that aims to develop a terminology service using query expansion in the field of archaeology	• Terminology expert • Experts in the field of developing semantic web
Interviewee 7: A research officer	SKOS Project: a project that aims to develop a RDF vocabulary for encoding the KOS	• Experts in the field of developing semantic web • Terminology expert
Interviewee 8: An information consultant	BS8723: a project to formulate a new British standard to identify methods to develop vocabularies, and methods to improve the interoperability	• Terminology expert
Interviewee 9: A lecturer in a Department of Information Science	CAIRNS Project: a project that aims to develop a meta-search system to cross-search a number of library catalogues. Also, the HILT terminology service is integrated into CAIRNS.	• Experts responsible for developing library portal products • Experts in the field of developing semantic web

3.3.3 Interview questions and coding

Three categories of experts were identified in Section 3.3.2. They have a complex stock of knowledge about the topic under study. According to Legard (2003, p.148), the aim of an in-depth interview is to reach both breath of coverage across key issues, and depth of coverage within each. Thus, when the interview questions were designed, it was

important to take two key issues very seriously: content mapping and content mining. Content mapping questions were designed to open up the research territory and to identify the dimensions or issues that are relevant to the participant. Content mining questions were designed to explore the detail which lies within each dimension, to access the meaning it holds for the interviewee, and to generate an in-depth understanding from the interviewee point of view (Legard and Keegan 2003, p.148). Probes were another important issue in designing interview questions. These took the form of follow-up questions to elicit more information, description, explanation and so on. For this reason, the interview questions needed to be extensive enough to explore the detail. Table 3.2 shows a number of anticipated issues related to developing the middleware systems.

Table 3.2: Issues related to developing the framework

Key categories	Sub-categories	Notes
Library portal with terminology mapping middleware	Metadata crosswalk	This is related to the methods to improve interoperability between different metadata schemes used by different collections.
	Access protocols * Z39.50 * SRW/U * OAI-PMH * SPARQL	The protocols to access the metadata repositories, and support the cross-searching.
	Subject services requirements: * Subject browsing; * Subject searching; * Collection finder; * Query expansion; * Query disambiguation; * Subject indexing.	This refers to a number of use scenarios, in which the end-users could use the defined subject-related functions to retrieve information in a library portal
	Data conversion and transmission	This refers to translating the query in a meta-search engine into various forms of queries that different participant online databases could recognise and accept, and converting heterogeneous metadata results into a consistent format, and present to the users.
Terminology	Terminology Mapping methods: * Switch language; * One-to-one mapping; * Mapping different KOS to an ontology; * Automatic mapping * Co-occurrence mapping	This refers to the ways to improve the interoperability between different KOS.
	Mapping relationships	This refers to the ways to identify the semantic equivalence between terms from different vocabularies.
	Mapping logics	This refers to defining a number of rules to develop mapping.
	Mapping partnership	It is related to the ways to distribute the mapping work to different communities.
	Mapping tools	Computer-based tools to help the people create the mappings
Semantic web-related issues	Encoding formats: * RDF/SKOS; * MARC21; * Zthes; * XTM; * BS8723-PART5;	The formats are used to exchange the terminological information between different services.
	KOS-related access protocols: * SKOS-API; * SPARQL; * SRW/U; * Z39.50.	The protocols to access the terminology services
	Information architecture: * Centralised terminology mapping service; * Distributed terminology mapping services.	This refers to the ways to construct a database(s) for storing the terminological data.

As well as ensuring that the interview questions addressed these topics, these issues could also be defined as a variety of codes for analysing the interview data. This provides a basic guideline to code the interview data in a systematic way. In addition, it is worth noting that because a key strength of qualitative research is that it can explore some unanticipated issues, and the answers to the research questions are made by different interviewees, and reflect different dimensions, the hierarchy is extensible to cover unanticipated issues and methods. Specific unanticipated issues would be presented in Chapter 4.

In order to make the interviewees clearly understand the aim and objectives of this research, and avoid unnecessary confusion during the interview, it was important to introduce the contextual information and an outline of the proposed theoretical framework to the participants before beginning the interview questions. With all these considerations in mind, a detailed interview question list was designed and can be seen in Appendix 2.

3.4 Evaluation

As mentioned in Section 3.1.3, an evaluation should be conducted to identify the potential problems in the theoretical framework and relevant solutions, and to collect suggestions for improving the theoretical framework. When the simplified prototype system was developed, it was important that the selected approach would solve the identified problems. However, it was important to evaluate the feasibility of these approaches with the intention of finding problems and relevant solutions, and finally adding value to the theories. Before developing an evaluation plan, Patton (1987, p.8) points out the importance of clarifying a number of questions. These questions include:

- Who is the information for and who will use the findings of the evaluation?
- What kinds of information are needed?
- How is the information to be used? For what purposes is the evaluation being done?
- When is the information needed?
- What resources are available to conduct the evaluation?

These questions are addressed below.

3.4.1 Who is the information for? When is the information needed?

Kumar (1990) pointed out two types of evaluations serving the needs of different groups. These are formative evaluation and summative evaluation. Summative evaluation offers information on the systems' efficacy to those users who are going to use the system (Kumar 1990). This kind of evaluations is often conducted after the development is finished. Formative evaluation is conducted during the developmental stage or system enhancement stage, and offers feedback information to help improve the system (Scriven 1967). It is beneficial to those who are involved in the system development process.

Taking a broad view of the scope of the people that can benefit from the research, these people might include a variety of stakeholders: end-users (students, researchers, or internet seekers, etc), software product developers, terminology mapping project staff, etc, but the final framework will serve as a range of technical and practical suggestions for portal system developers when they develop library portal systems. The target audience of this research should be library service developers who can benefit from the findings when they develop terminology mapping systems. For this reason, formative evaluation is more appropriate in this research. Formative evaluation was therefore used to ensure that the goals of the theoretical framework were being achieved, and to gain suggestions on how to improve the system.

Hence, this research was aimed at providing the developers of this kind of terminology middleware with a number of guidelines, to help them develop their related solutions to integrating various terminology resources.

3.4.2 What kinds of information are needed? What is the purpose of the evaluation?

Regarding different aspects of the feasibility of an information system, Scholtz (2001) summarised five types of attributes in usability evaluation. They include learnability, efficiency, memorability, errors, and user satisfaction. He also indicated that "depending on the type of application, one attribute might be more critical then another". Scriven

(1967, p.39) made an important distinction for two types of evaluation methods between system intrinsic features measurement and payoff performance measurement, which are called analytic and empirical evaluation. In this context, analytic evaluation refers to "examining intrinsic features and attempting to make predictions concerning payoff performance, and empirical evaluation refers to attempting to measure payoff performance directly" (Gray and Salzman 1998, p.203). Scriven (1977, p.346) took an example to illustrate the difference between analytic evaluation and empirical evaluation. As he remarks, "If you want to evaluate a tool . . . say an axe, you might study the design of the bit, the weight distribution, the steel alloy used, the grade of hickory in the handle, etc., or you might just study the kind and speed of the cuts it makes in the hands of a good axe-man".

In this research, the research outcome is a complex process, and a number of judgments need to be made during the design. For this reason, an analytical evaluation was chosen to test the intrinsic features of prototype system, find out the problems and relevant solutions, and improve an iterative development process. The analytical evaluation should test the following elements in the system:

1. Usefulness of the mapping established;
2. Viability of the protocols selected to access the KOS, and the encoding formats to represent KOS;
3. Viability of the whole network infrastructure applied to exchange terminological information between distributed KOS;
4. Functionality of the system designated based on the network infrastructure;
5. Potential cultural difficulty in implementing the system designated.

Obviously, it is difficult for an end-user to understand each of these identified elements, and some end-users may even have trouble understanding the purpose of this middleware system for terminologies. Rosson and Carroll (2002) indicated that user-centred evaluation methods are appropriate to be used in most empirical evaluation cases, but that expert-based approach may be more appropriate in analytical evaluations. In this sense, it

was decided to employ experts' knowledge as a basis for identifying and predicting usability problems with the middleware and underlying framework.

3.4.3 What resources are available to conduct the evaluation?

In the literature, a variety of prototypes with different degrees of fidelity have been summarised to help users or developers understand the requirements for systems. In general, paper-based mockups, as the visualisation of developed theories, are used at the early stages of the design process, and computer-based prototype systems are assumed to be effective in testing usability at the later stages. However, new research findings indicate that low-fidelity paper-based prototype can also be applied at the later stages to test different aspects in a system, and vice versa (Bailey 2005, and Sefelin *et al.* 2003). Klee (2000) pointed out different tools to create a prototype system reflecting the theories, which include HTML pages, paper mockups, existing information system, databases, etc.

In this research, one purpose of prototyping is to test if the selected technologies could be effectively used in the middleware system between terminologies. Therefore, a paper-based mockup is inappropriate for this purpose. A computer-based prototype system, which applies most of technologies identified in the theoretical framework, needed to be developed. The prototype is the main resource for the evaluation. In addition, a document describing the principles of the theoretical framework was presented to help the experts further understand the prototype. This document is presented in Appendix 4.

3.4.4 Evaluation methods

As mentioned in Section 3.4.2 and 3.4.3, expert evaluation was chosen based on the use of a computer-based prototype system. There are a number of specific expert-based usability evaluation methods summarised by Nielsen (1993), which include pluralistic walkthroughs, cognitive walkthroughs, heuristic evaluation, etc.

In these methods, "Heuristic evaluation is best used as a design time evaluation technique, because it is easier to fix a lot of the usability problems that arise" (Nielsen 1993, p.25). The basic idea of heuristic evaluation is that a number of expert evaluators use a number

of criteria, independently evaluate a prototype system, and then claim potential usability problems. Heuristic evaluation is best used during the system design stage (Nielsen 1993, p.25). As emphasised in Section 3.4.1, the evaluation in this research needed to be conducted during development. Thus, heuristic evaluation was used as a method to assess the framework. Also, the heuristic evaluation method makes it easier to fix a number of potential usability problems (Nielsen 1993, p.25). Based on the use of heuristic evaluation methods, Nielson (1993, p.47) noted that "around five evaluators usually result in about 75% of the overall usability problems being discovered".

Compared with heuristic evaluation, a cognitive walkthrough emphasises the importance of making the experts interact with the prototype based on a number of given tasks (Wharton *et al.* 1994). In order to fully evaluate different aspects of established theories, a combination of heuristic evaluation and cognitive walkthrough was applied in this research. In this case, five usability experts and one software developer, who are also knowledgeable in the field of terminology services, were selected to test the prototype system. These evaluators are not those who were previously interviewed. The detailed information about these people is presented in Table 3.3.

Table 3.3: The Details of evaluators

Usability evaluator	Qualification
Expert 1: information consultant	• Knowledge of development of ontologies, taxonomies and thesauri and knowledge elicitation. • Familiar with the applications of KOS TO metadata schemes, associative databases, and XML-based technologies such as RDF, XSLT and XTM (Topic Maps).
Expert 2: information consultant	• Developer of BS8723; • Advisor of a number of terminology services, such as HILT.
Expert 3: lecturer in the department of information science	• Knowledge of faceted classification theory; • Joint editor of BLISS Classification.
Expert 4: information consultant	• Developer of BS8723; • Author of a famous government-based thesaurus.
Expert 5: information consultant	• Joint editor of UDC Classification.
Expert 6: Software developer: Programmer	• Developer of BS8723-Part 5; • Experience of converting a famous thesaurus into a well-defined digital format.

In this evaluation, the experts agreed to base the walkthrough on three different roles. They were:

Role 1: Developers of middleware system between different terminologies;

Role 2: End-users;

Role 3: Collaborators of the developers of middleware system between different terminologies, such as the developers of meta-search engines, developers of terminology services, developers of service registries, etc.

Based on these three roles, a number of heuristics for evaluating the middleware were defined. These heuristics include:

Heuristic 1—Semantic extensibility: Through exchanging terminological information between different KOS, a terminology mapping middleware system that supports subject cross-browsing functions should be able to incorporate different kinds of KOS having different semantic structures across different subject areas. Semantic extensibility refers the capability to map different KOS together. New resources with new KOS may be

developed in future. It is important for a terminology mapping middleware system to continuously incorporate the new conceptual resources, and provide appropriate functions to maximise the use of most KOS. One most important part of this evaluation is to test whether the framework can be applied to integrate different KOS in widely distributed information environments. In addition, different decisions have been made to establish conceptual mappings between different KOS. As mentioned above, these KOS differ greatly. A number of factors need to be tested, which include the structural models of mapping, mapping relationships identified, the ways to treat compound concepts, the methods to select the conceptual elements to be mapped.

Heuristic 2—Technical adaptability: Technical adaptability refers to the capability to process different KOS using different formats, access protocols, and being located in different information systems. When a meta-search service cross-searches a wide variety of information resources, the KOS used by these information resources may greatly differ in their subject areas, the degree of pre-coordination/post-coordination, the level of granularity, encoding formats, access protocols, and the use of language. The heterogeneity caused by these differences greatly impedes the effective use of the meta-search service. In this context, this middleware system should be able to provide the capability to harmonise heterogeneity caused by different KOS.

Heuristic 3—User interactivity: This refers to the ability to satisfy the different subject needs of different organisations using this terminology mapping service. A subject browsing interface cannot satisfy all the subject needs across different communities. For example, it was found that many users find it difficult to follow DDC structure to find relevant information. In many cases, people from a community may have been familiar with a taxonomy for a number of years. It is important to re-design the subject browsing structure to satisfy the specific subject needs of each community. In addition, it is hoped that the terminology mapping service can be interoperable with the taxonomy used by a particular organisation, and the users in this organisation can directly interact with their own taxonomy that they are familiar with, but gain the terminological results from

different KOS. In addition, it might be a long process from a user's initial interaction with the browsing interface to getting the mapped term to searching against these collections and then results. It is important to optimise the process in a more user-friendly fashion.

Heuristic 4—Cultural feasibility for the development of this kind of terminology mapping middleware services: In this research, it is proposed to develop a collaborative mapping workflow. A number of participants are involved in the mapping efforts. These people include local librarians, terminology service providers, KOS owners, etc. Thus, it is important to evaluate the feasibility of the proposed mapping workflow, and test whether these identified people are able to make a collaboration of creating the mappings.

Heuristic 5—Technical Feasibility: In the software-based prototype, different protocols, formats of data, and agents to process the data are employed, and all these components have been integrated together to offer the services. Being aware whether these techniques are used in appropriate places and appropriate ways is important. In other words, a feasible technical infrastructure should be able to effectively access different KOS without fatal errors, and provide appropriate subject cross-browsing services to users. Some suggestions are expected to improve the technical feasibility of the prototype and make the service more practical in real digital library environments.

Heuristic 6—Quality of mapping: In the prototype system, based on defined guidance, a range of mappings were manually established. It is important to test whether the established mappings could provide effective subject cross-browsing services. Also, expert evaluators were asked to offer suggestions to improve the quality of the mappings.

Because each heuristic may include a number of relevant factors for the development of the middleware, these heuristics were translated into a number of questions to experts. The experts were asked to read the document describing the principles of the theoretical framework, and then independently walk through the prototype system. Four tasks based on different use scenarios were formulated to make the experts interact with the prototype.

After each task was completed, the experts answered a number of formulated questions relevant to the evaluation heuristics. When all the tasks were completed by an expert, the expert would be asked to freely offer some comments based on the defined criteria. A detailed evaluation plan for the experts is presented in Appendix 5, and the document describing the theoretical framework is presented in Appendix 4.

The data was collected using an audio recorder, and analysed according to the heuristics. The findings were presented in Chapter Seven, and were expected to form a basis to improve the theoretical framework.

The next chapter describes the ways to analyse the interview data.

Chapter Four: Data Analysis for Establishing a Theoretical Framework

4.0 The findings into the investigation of various KOS

As emphasised in Section 3.2, it is important to investigate the main characteristics of different KOS. An investigation was conducted for this purpose, and presented in Appendix 1. Based on this investigation, seven points have been identified:

- Some databases are indexed by more than one controlled vocabularies. For instance, three controlled vocabularies are being used to index an online database called OMNI. These vocabularies include MeSH, UMLS, and National Library of Medicine. In this case, co-occurrence mappings could be established based on the subject-related metadata indexed by these three KOS. A metadata record could become an important intermediary for exchanging subject terminology between different knowledge organisation systems.
- Some common controlled vocabularies, such as UNESCO, LCSH, etc., are widely used by a number of online services, but other specific vocabularies are developed from these widely-used ones. For example, MeSH was developed based on the structure of LCSH, and similarly HASSET is based on the UNESCO Thesaurus. In this context, it is crucial to investigate the important features of these widely-used classification systems, such as DDC, UDC, LCC, LCSH, and UNESCO Thesaurus. In addition, it was found that a large amount of mapping work between a number of widely-used vocabularies had been established by different organisations. For instance, LCSH has been mapped to DDC. Likewise, Saeed (2002, p.575) proposed a framework in which "terms drawn from DDC indexes and IEEE Web Thesaurus were merged with DDC hierarchies to build a taxonomy in the domain of computer science".
- Some subject-specific thesauri are not only designed to provide tools for subject indexing and searching information, but serve as a guide to explore the knowledge in some specific fields. For example, the Gene Ontology thesaurus is aimed at describing different genes and gene product attributes in any organism,

and working with a specific database. In another case, the ASIS thesaurus can help information science students understand the landscape of information science terminology.

- It was found that two KOS from different domains differ greatly in respect of their syntaxes and semantics, but there are more similarities between KOS from the same domain. In other words, it is easier to merge a number of vocabularies within the same domain. For example, UMLS merges concepts from about fifty medical controlled vocabularies into a metathesaurus.
- There are a large number of in-house classification systems, such as subject-specific thesauri, and ontologies developed by different organisations. How to make these in-house knowledge organisation systems' interoperable with some generally-used KOS is a key issue for this research.
- It was found that a subject-specific thesaurus might be very context-based, and only designed and used for a specific database. However, a large number of common subject classification schemes and subject headings, such as DDC, UDC, LCC, LCSH, ACM, MeSH, etc., are widely-used by a number of information services. For this reason, it is a good starting point to create the mappings between these widely-used vocabularies to assess its usefulness before expanding the mapping work.
- Based on this investigation, only very few of these identified bibliographical databases adopt a fully-fledged ontology to describe the subject topics and their relationships. It was found that the use of ontologies for facilitating subject cross-browsing and searching is still in a very early stage. In this investigation, some ontologies, such as *CIDOC CRM*, *ABC Ontology*, etc., are used as metadata schemes to describe complex objects instead of subjects. Different objects could be categorised under predefined ontological classes, and a number of attributes could be used to specify a class. For this reason, in many projects, it was necessary to use the ontologies to describe complex information objects, such as museum objects, images, etc., but use various controlled vocabularies to describe the subject concepts and relationships.

4.1 Interview findings related to various library portals

As outlined in Section 3.3.2, nine expert interviews were conducted in this research to collect ideas for the development of a theoretical framework. The interview questions were outlined in Appendix 2. The interviews were recorded, and the recordings were transcribed into text. In order to analyse the text data, Altas Ti scientific software was used. A number of codes were identified in Table 3.2, and these codes formed the basis of the analysis of the interview transcriptions.

A range of steps were designated to analyse the transcription:

1. The important textual information in the transcriptions was highlighted as quotations;
2. The explanations and comments were given to each quotation to make the quotation clearer;
3. Relevant codes were assigned to the explanations and comments created;
4. The codes were then grouped by code family and sub categories;
5. The grouped explanations and comments were summarised.

The specific findings are presented as follows:

4.1.1 Metadata crosswalk (mapping)

Interviewees 1, 2, 3, 4, and 9 highlighted the heterogeneous nature of different information resources. It was generally felt that the content providers should offer their standard APIs to shared services such as SirSi Subject Rooms, Ex Libris MetaLib, etc. In the real world, however, the APIs that are offered varied from one content provider to another in terms of different access protocols and encoding formats used by different content providers (Interviewee 4). In this context, Interviewees 1, 2 and 4 pointed out that the different metadata standards, protocols, and formats used should co-exist for a long term, and that a library portal service should be able to cope with most of these standards, protocols, and formats.

Interviewees 1, 3 and 9 stated that there was a lack of mappings between different metadata standards used by different resources. For example, some databases offer

title/keyword, author/keyword or subject keyword search. Although there are rich sets of metadata elements within some databases, some rich metadata in these catalogues may not be exposed to the local search engines, and some may not be exposed to the Z39.50 servers. This depends on how the local catalogue staff map their item-level metadata to searchable indexes in the local search engine and also to the Z39.50 server. For this reason, Interviewees 1, 2, 3, and 4 highlighted the importance of mapping a central metadata scheme to different metadata schemes. In this context, one-to-many mappings could be conducted.

In the projects that Interviewees 1, 3, and 4 have been involved with, the typical approach was to create a common metadata scheme based on reviewing a variety of metadata schemes used by other information services, and the mapping work was done on the service provider side. In Interviewee 3's Project, an application profile based on the DC metadata scheme was developed, into which other metadata schemes were mapped (Interviewee 4). Based on the mappings between different metadata schemes through the application profile, a user can begin to interact with the Renardus federated search interface for cross-searching different information services, and jump to different user interfaces to view the details of the returned results. In the application profile, a specification of the syntax encoding scheme needed to be given to different content providers to follow when they create the mappings. Once the application profile was developed, the mappings from different metadata schemes to the application profile could be done by different local experts. By using the Z39.50 protocol, the service could cross-search different information resources. The project found that the DC metadata scheme was suitable as the basis for adaption into the application profiles for mapping different metadata schemes together. There were a lot of advantages to using DC, such as its multidisciplinary nature, the frequency of its update, etc.

Interviewees felt that an application profile can be made for local use to map heterogeneous metadata. Alternatively, a local library can develop an in-house metadata scheme, to which different metadata can be mapped. For example, it is possible to create a core ontology for mapping different metadata schemes in the same domain.

Interviewees 1 and 3 pointed out the importance of developing both a vendor-hosted and local-hosted service. In this approach, the vendor creates the mappings between different metadata schemes used by different data providers, and the local librarians could use the mapping created by vendor, and also create some further mappings based on their local information requirements.

4.1.2 Access protocols for metadata

Various protocols have been developed in different communities, such as Z39.50, SRW/U, OAI-PMH, etc. Interviewees 1, 2, 3, 4, and 9 outlined the difficulty in implementing Z39.50 protocols for the development of a meta-search service. Although a Z39.50-based search could offer a more accurate search, the number of information resources in each search query is very limited (Interviewees 2 and 9), the response time is relatively slow (Interviewees 1, 2, 3, 4 and 9), and therefore the use of Z39.50 is much more expensive in terms of the mapping effort, and technical cost (Interviewee 2). For example, the MetaLib federated search service, as a service using Z39.50 can use one query to cross-search at most eight information services simultaneously.

Apart from the Z39.50 access protocol, Interviewee 2 pointed out another protocol (OpenSearch) as having different searching features. One disadvantage of this protocol, as indicated by Interviewee 2, is that it cannot effectively conduct advanced searches on different metadata elements, such as subject metadata element, author metadata element, date metadata element, etc. In other words, OpenSearch protocol can only take a search into the general search boxes of different information services, and get results. However, using this protocol, there is no limitation on the number of information resources, which can be cross-searched, and a user can search all the required information resources in the same time.

Interviewee 1, 2 and 4 pointed out that by using the OAI-PMH to harvest different metadata into a centralised database, different metadata records harvested are more easily

controlled, because it is possible to use a centralised terminology to index these harvested metadata records. However, Interviewee 3 pointed out that in the real world, some content providers may not allow their metadata to be harvested. Thus, Interviewees 3 and 4 expected that in future the OAI approach and alternative distributed approaches should be developed together, and be made to complement each other.

By investigating a range of current protocols to access to KOS, Interviewee 5 felt that SRU/SRW was a very good protocol for exchanging information between different services in distributed information environments. In his project, by using an SRW server, to which different other services can set up relevant SRW clients, a centralised metadata database can respond to the subject query terms from different services, analyse the users' subject requirements, and finally let the users know what services are suitable for users' subject needs, what subject schemes are used by these services, and what terms in these subject schemes can be mapped to the users' query. This is explained in Figure 4.1. In addition, the service described by Interviewee 5 uses a centralised database to hold all the metadata used by the participating collections and all the mappings between them. By using a SOAP server, the terminological data in the database can be wrapped into the RDF format in the SRW server. The SOAP server, which is responsible for wrapping data in RDF, makes it easy to exchange the terminological information between the SRW server and different SRW clients.

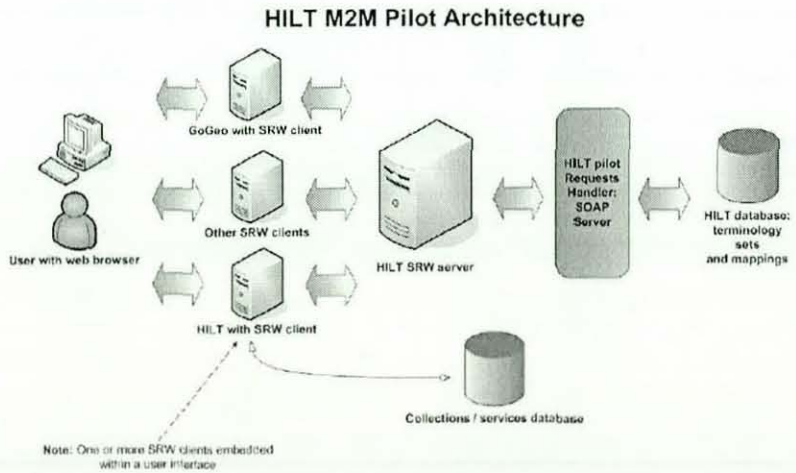


Figure 4.1: A SRW-based subject service

Based on using the SRW protocol, other APIs could be added into the SRW server side, which would provide more functions to the SRW clients. For example, a WordNet API could be added into the SRW server, and the related functions can be offered to expand a given term to return sets of grouped cognitive synonyms.

If all the metadata records could be converted into RDF format, Interviewee 7 emphasised the benefits of using the REST-HTTP Protocol and SPARQL to access RDF database. REST-HTTP and SPARQL is much more lightweight and less expensive than Z39.50 and SRW/U-based approaches. The endpoint of a service should be the web. For this reason, it is possible that if all the data were stored and accessible on the web, then it is easy for a meta-search service to build programmatic access interfaces to access the mapping databases across the web. However, Interviewees 6 and 7 pointed out the problems of using SPARQL. SPARQL is not good at indexing text-based information, as it is oriented towards searching the RDF graph. In other words, there is very limited support within the SPARQL query language for phrase queries, wildcard queries, proximity queries, range queries, and so on.

Interviewees 1 and 3 emphasised the importance of developing a multi-protocol environment for the development of a meta-searching service. Interviewees 1 and 3 pointed out that data conversion programmes are required to be developed to convert heterogeneous metadata records into a consistent library-based format, and present a combined view of the records from different resources to the end-users. This would mean that the retrieved records from different resources could be merged, de-duplicated, sorted, ranked, and filtered. More importantly, a data transmission programme should be developed as well. This programme can be used to translate a user's query into different forms of queries that different participant databases can recognise. Thus, Interviewee 3 concluded that a knowledge base acting as a mediator should be developed to store the connectivity details of different databases, outline how a federated search service can connect each database, and the metadata elements mapping instructions between different metadata schemes used by different information services. A knowledge base should set up the rules that transfer users' queries into the form that a database can understand.

Interviewees 1 and 3 continued to mention three important types of information that need to be developed in a mediator for exchanging information.

1. Descriptive information includes such general data as the full name of the specific resource, alternative names, the vendor's access policy (whether the resource is free or by subscription only), the type (such as a library catalogue, abstracting and indexing database, or Web search engine), the creator, and the publisher;
2. Rules for the transfer of a query. This information is provided on three levels: the method of transfer (such as server-to-server interaction or HTTP); the format in which the query is transferred (for example, it can be sent as a Z39.50 query or as an HTTP request, which is either formatted as a URL or sent in an XML document); and the construction of the query itself. The latter involves issues such as the adaptation of the query syntax, mapping of search indices, and character conversions.
3. Rules for the interpretation of the results obtained from the resource. This information resembles the query information and includes the rules that permit the interpretation of records that differ in their physical and logical structures, cataloguing formats (such as MARC and MAB), and character sets. Once the records are interpreted and converted into a unified internal format, machine-driven tools can display all the results to the user in the same manner, compare records that originate from disparate resources, provide services relevant to specific records, and more. With the interpretation of the results, the groundwork is set for the creation of an OpenURL for onward linking.

4.1.3 Subject-related use scenarios

In order to create a subject cross-browsing interface, Interviewees 1, 2, and 4 did not think DDC is a good tool for subject navigation. In this case, they also pointed out that developing in-house KOS is a more effective way to support subject cross-browsing. In this context, Interviewee 3 indicated that potentially, each organisation could develop its own vocabulary for subject browsing. Interviewees 1, 2, and 3 emphasised that the departmental structure and relevant modules could form a basis to create local KOS. In other words, based on the keywords incorporated within names of different departments

and subject majors, it is possible that university librarians can develop their own subject taxonomies.

Meanwhile, according to Interviewees 3, 4, and 5, most subject cross-browsing services just provide a starting point to help users begin their search, and may guide the users to the relevant collections. To some extent, these subject cross-browsing services cannot replace the role that each content provider's search interface plays. Therefore, it is extremely crucial to consider how to develop the interactive interface between the portal and different services that content providers provide. As Interviewee 1 stated, "The degree of the interaction between the users and some portal systems should be considered. In some cases, the users have no capability to interact with the portal systems. On the other hand, it seems to be impossible to create a totally interactive M2M interface. A balance needs to be made for this reason". Interviewee 6 had a strong feeling that in some projects, people were unrealistic and over ambitious in attempting to develop purely automatic cross-search and browsing methods, in which the end-users' help was not needed. It is therefore important to make a balance between the user engagement and the automatic process. In these cases, it is not realistic to develop something like a Google search engine, in which users can type search terms and then automatically conduct cross-searching without any user engagement. Thus, it is desirable to plan some interactive elements that users need to be involved in. One important element to facilitate user interaction is query expansion (Interviewee 6 and 7). Interviewee 6 developed a semantic expansion algorithm for spreading from a given concept over a thesaurus-based semantic network to yield a neighbourhood of concepts considered semantically close for retrieval purposes. According to Interviewee 6, this query expansion algorithm has been proven to support effective subject browsing.

In many cases, people still need to access different information resources separately, Therefore, Interviewees 1, 2, 3, 4, 5 and 9 highlighted that it is important to index different collections centrally. This can increase the response efficiency, as service registry systems can be developed to provide the collection-level metadata to the portal services. In this context, each of these digital collections may be classified according to

using one or more KOS. The selected hierarchy can be used as a starting point to help users find the relevant information. For this reason, through the structure of the used KOS, the end-users could be navigated to find relevant collections.

Interviewee 9 noted that different concepts in a subject browsing interface have different popularities. In social tagging research, "tag clouds" allow tags to be displayed in different formats for different purposes. In order to browse a KOS, it might be possible to use this type of display to present some concepts. For example, if there are a number of concepts related to 'history' within different geographic areas at the same level of granularity, the concept "the history of Europe" could be highlighted within the UK-based library portal services rather than the concept "the history of Asia".

Interviewees 4, 5, 8, and 9 noted that a term disambiguation function is very valuable. It is important to consider various ways to present a concept to the end-users, and realise that different concepts in different types of vocabularies should be presented in different ways. F or example, a term "weeding" could be classified under different categories within a subject vocabulary. In the field of computer science, the term "weeding" can be used to describe a type of collection development, but in the field of agriculture, the term "weeding" can be used to describe the activity of killing diseases and pests within the grass. Thus, it is important to present the relevant contextual information of a given term to the users, and then let the users understand the real meaning of a given term.

4.2 Interview findings related to various terminologies

4.2.1 Structural models

Because it was assumed in this research that a wide range of controlled vocabularies needed to interoperate with each other in order to provide the end-users with a subject cross-browsing service, Interviewees 4, 5, 6, 7, 8, and 9 highlighted that the use of a switch language, to which different other vocabularies can be mapped, is one solution. This is because:

1. It is very labour-intensive and time-consuming to create direct many-to-many mappings between different vocabularies (Interviewees 4, 5, 6, 8, and 9);
2. A number of vocabularies, such as LCSH, ERIC, HASSET, IPSV, etc., have been mapped to DDC as a switch language. It is easy to expand the mapping work based on the use of DDC (Interviewees 4, 5, 8, and 9);
3. With the continual development of the mapping data based on using the selected switch language, the mapping worker would become more and more familiar with the structure and principles of the switch language, and deal with more and more complex mapping situations. This would help improve the consistency and accuracy of the mapping work done by the workers.

Interviewees 3, 5, 6, and 8 emphasised the advantages of DDC as a spine to exchange terminological information between different KOSs. The advantages mainly include:

1. Using DDC web services as an entry vocabulary to index a number of collections is very useful (Interviewees 3 and 4). Based on this entry vocabulary, different relevant collections can be identified, and the users' subject needs can be clarified (Interviewees 4, 5 and 6);
2. DDC has been adapted in many digital systems on the internet (Interviewee 4). When developing a subject cross-browsing interface across different information services, it is not required to establish the mappings between these services using DDC ;
3. A large number of databases are not only indexed by DDC, but also indexed by other vocabularies, which offers a basis to create the co-occurrence mapping between DDC and other vocabularies (interviewees 3 and 8);
4. A large range of communities and organisations are using DDC (Interviewee 3). For example, tens of thousands of libraries are using DDC and so do 56 national bibliographies. For this reason, a large number of users might be familiar with the structure of DDC;
5. Because DDC is controlled by OCLC, DDC has an advantage in the speed and frequency of updating main subject classes (Interviewees 3 and 8);

6. Due to the efforts of OCLC, DDC has been represented in a range of widely-used interchange formats, which include MARC21, Zthes XML DTD, etc (Interviewees 4 and 6).
7. By mapping different KOSs to DDC, DDC can become a navigation tool to help users to find relevant parts of other hierarchies. Users can then be navigated by the identified parts of hierarchies mapped to DDC to find relevant information.
8. The use of DDC is controlled by OCLC. This means that different communities all use DDC consistently (Interviewees 3 and 6).
9. If a KOS has many facets, it will be difficult for it to become a mapping central spine. Because of the limited facets of DDC, DDC is a good candidate to become a mapping spine (Interviewees 3 and 5).
10. Because DDC is a decimal classification, Interviewees 5 and 9 developed a DDC-based truncation algorithm to find more general collections. The basic principle of this algorithm is to map a user's query term to DDC caption, get the relevant DDC notation, and truncate the specific DDC number to match the more general DDC concepts that are used to index collections. Using this algorithm, a number of relevant collections will be returned.

In general, DDC is a general subject classification scheme covering most subject areas. It can play important roles in cross-browsing by subject, information discovery by subject, and finding the relevant subject parts from other KOSs mapped to DDC (interviewees 3 and 6). In other words, DDC, as a spine, provides a starting point to help users begin their search, guide users to relevant collections (Interviewee 3, 4, and 5), and also direct users to local directories relevant to the users' topics. In fact, a lot of content providers may hope that the users do not solely stay in DDC spine to do all their information searching, rather, they may want users to use DDC to find their information resources and classification schemes, and search information in their own websites. In this context, DDC subject cross-browsing can play an important role in guiding users to find relevant subject-specific classification schemes from library portal systems.

However, there are some shortcomings to DDC as a central spine that were mentioned by Interviewees 4, 5, 6, and 8. Most importantly, the use of DDC as a central spine will cause one-step indirection problems (Interviewee 3 and 5), and the specificity may be lost in this process. One step indirection is where there are two thesauri which are A and B and where direct mappings between A and B is impractical. Therefore A is mapped to a switch language "S", and B is mapped to S. This enables the creation of indirect mappings between A and B through S. If required, users can be informed that there is a switch language in the middle.

Because DDC is a pre-coordinated classification scheme, it is very different to establish mappings between DDC and some other vocabularies (Interviewees 3, 5, 6, and 8). As interviewee 8 mentioned, using DDC to switch terminological information between different KOSs is not to switch individual concepts, but to switch the compound concepts. In this case, Boolean combinations should be used (Interviewees 4, 6, and 8). The BS8723 provides the clear guidelines on how to conduct the mappings from a pre-coordinated vocabulary to a post-coordinated vocabulary (e.g. UKAT), which include:

- Map the terms or notations representing the separate headings and subdivision of the pre-coodinate scheme (DDC) to the target vocabulary (ACM);
- Map any pre-coordinated strings that are enumerated in the source vocabulary. Some of these may map to pre-coordinated expressions in the target vocabulary, or they may be mapped to a combination of terms (ACM);
- In order to include additional pre-coordinated strings, mapping may be based directly on the entries in the existing catalogue rather than the nominal source vocabulary.

Other problems are that, in some cases, DDC might be too general to index some subject-specific collections, and it may be too general to be mapped to some subject-specific KOS, and it may not suit some user's subject needs. In other words, the mappings between DDC and other KOS may be not specific enough (Interviewees 4,5, 6, and 8).

Moreover, DDC cannot be immediately adapted to describe very new subject areas (interviewee 6), and cannot be changed for some particular applications, because DDC is a proprietary system. It is dangerous to add new numbers or captions to the published DDC, as this may cause incompatibility issues and copyright issues, although DDC has provided some instructions about its number building capability.

From the perspective of librarians, Interviewees 1, 2, and 4 also did not think DDC is good enough to be an appropriate option for subject browsing in a university environment. This is because the university students may not be familiar with the terminology provided by DDC. They recommend that in-house classification schemes need to be developed based on the departmental structure for an institute (Interviewees 1, 2, and 4). In this context, Interviewee 1 highlighted that it would be desirable to provide a subject-specific KOS as a browsing interface for some subject-specific portal.

With this consideration in mind, Interviewee 2 mentioned the possibility of letting local libraries develop or use their own terminology for their own users. In fact, a number of service providers, such as SirSi Rooms, give the local librarians the flexibility to develop their own portals (Interviewee 2). Similarly, it is expected that the service providers or the third party can provide a central terminology service so that different portal services (or even end-users) could develop their own terminology, which is based on the central terminology service. For this reason, it is reasonable to establish mappings between DDC and different KOS, and map all these controlled vocabularies into a super-vocabulary. Local subject librarians can adapt some parts of this super-vocabulary for their own particular requirements, or create the mappings between their own vocabulary and the super-vocabulary. Also, based on the super-vocabulary, the librarians could create their own social taggings, and map them into the super-vocabulary as well (Interviewees 3 and 6), see Figure 4.2.

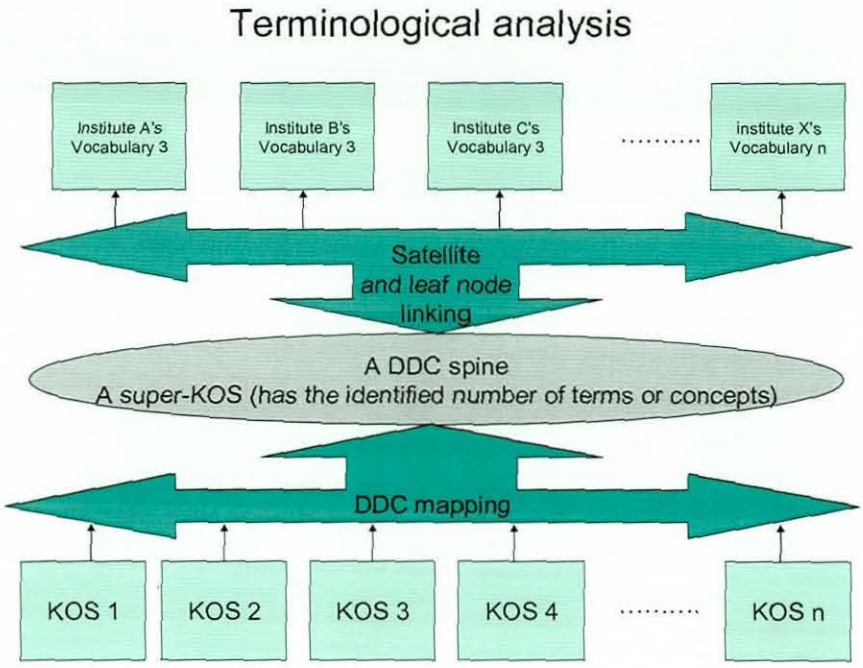


Figure 4.2: DDC-based mappings

Interviewees 3 and 8 mentioned that in a particular domain, it is possible to merge all the controlled vocabularies into a meta-thesaurus. For example, in the medical domain, information professionals have developed a conceptual spine called UMSL, into which about fifty medical controlled vocabularies have been merged. Likewise, in the pedagogical domain, the JISC has merged a range of pedagogical controlled vocabularies into a conceptual spine developed in the project. The concept spine can be developed based on an existing subject-specific thesaurus. For some subject-specific KOSs, a specific thesaurus might be able to be a conceptual spine. Different KOSs could follow the structure of this spine. In some cases, the KOS may need to be re-designed to adapt this spine. Furthermore, the selected thesaurus could also be migrated into a lightweight ontology that provides richer relationships between different concepts.

However, Interviewee 3 felt that it is impossible to merge all vocabularies across different disciplines into a meta-thesaurus. For example, it was stated that in the social science domain, it is impossible to merge all social science vocabularies together,

because these vocabularies greatly differ in subject areas, granularity, the use of language, pre-coordination/post-coordination, etc.

4.2.2 Ontological mapping

Ontologies can play different roles in improving the interoperability between different KOS, and facilitating a range of intelligent subject services (Interviewee 3, 4, 6, 7, and 8). In this context, Interviewee 6 mentioned that a decision on whether to create ontologies to facilitate a terminology mapping service depends on what purpose the service wants to achieve. It was also indicated that the use of ontologies can broaden the scientific modelling, support logic reasoning, and make the computer 'know' how a particular entity is related to the classification.

Most formal ontologies are not designed to facilitate subject cross-searching and browsing services, and they are not cost-effective and useful (Interviewee 6). Also, Interviewee 4 emphasised that mapping different KOSs to an ontology, to some extent, can improve the subject cross-searching service by adding reasoning functions, but that the cost is very expensive, and the retrieval speed may be very slow. Interviewee 4 also stated that the structure of an ontology is not good for subject browsing, because the relationships between different elements are too complicated.

It is obvious that one of the most important purposes of mapping KOSs to an ontology is to facilitate the provision of semantic reasoning. In this case, Interviewee 7 presented two basic methods to map KOS to the ontologies, which are described as follows:

1. Migrating: In order to provide thesauri with more description logic, it is necessary to add more semantic information on the structure of the thesaurus, and disambiguate the hierarchical relationships and related relationships in some meaningful ways.
2. Mapping KOS into an existing ontology: Another way to improve the description logic is by mapping a knowledge organisation to an ontology, and by using the ontology, the concepts in the thesaurus can be inferred.

In order to improve the semantic reasoning functions of a terminology service, Interviewees 4, 5, 6, and 8 emphasised which ontology could be the most appropriate one for mapping KOSs.

According to Interviewee 4, most upper-level ontologies are too generic for establishing mappings between different KOS, and are also not topic-based. However, mapping KOS to a core ontology might be an option, which could facilitate the provision of semantic reasoning, and terminology switching. The concepts in a core ontology are useful for providing a basis as a conceptual spine. However, most of these core ontologies, such as CIDOC CRM, are too generic and high-level. Other core ontologies, for example, the ABC ontology, cannot provide enough semantic capability to describe categorical information. In this context, it is necessary to develop some concept-based core ontologies to support mapping in future.

Interviewees 5 and 8 highlighted the importance of the WordNet ontology as a linguistic ontology in assisting terminology services recognise more terms. For example, if a user were to input an adjective format of the word "agriculture", and there was no term "agricultural" in the terminology service database, by mapping KOS into WordNet, the terminology service can recognise the term "agricultural", and finally provide the users with the term suggestions, e.g. "do you mean agriculture?".

Interviewee 8 did not think the CIDOC CRM could provide enough terms or concepts that different KOSs can map. However, Interviewee 6 thought it possible to create mappings between the CRM and a number of archaeological KOS. He felt that creating mappings between DDC and other controlled vocabularies was aimed at establishing detailed term-level connections between KOS, but creating the mappings between a core ontology and different knowledge organisation systems was aimed at making different controlled vocabularies independent and complementary, and able to deal with different aspects of information resources. The core ontology, in this sense, could play an important role in combining controlled vocabularies in sensible ways to be cross-searchable, and handling the facets between different controlled vocabularies.

In summary, it is important to bear in mind the purpose the search service wants to achieve—who is the service's end-user, and what service can be created. In other words, it is important to develop some viable application scenarios when developing a terminology service. However, in the current situation, it can be summarised that no appropriate ontologies could be applied to facilitate the subject cross-browsing function.

4.2.3 Mapping relationships and strategies

Interviewees 4 and 5 stressed the importance of clarifying different types of mapping relationships in a mapping project. By clarifying the mapping relationships, the users may be made to feel more comfortable to be navigated and jum to another concept scheme. This is because the system can tell users the mapping relationships between different terms from different KOS. On the other hand, Interviewees 4 and 6 emphasised that the types of mapping relationships vary from one context to another depending on the specific application. If mapping between vocabularies and an ontology, the relationships will become much richer. In this context, Interviewee 6 mentioned that there are two important points when identifying mapping relationships. The first is to define the degree of equivalence. For example, five types of mapping relationships have been identified in the Renardus Project, which include exact match, broad match, narrow match, related match, and partial match. The second is to use Boolean operators to combine a number of single terms to map to a compound concept. On the other hand, Interviewee 7 mentioned that the most important mapping relationships include broader, narrower, equivalent, and related, and the question of whether to use the Boolean combinations is still an open issue. In some cases, the Boolean combination may make the mapping work more complicated, and make it extremely difficult to evaluate the quality of the mapping (Interviewee 7).

Interviewees 5 and 9 indicated two different mapping strategies for developing a terminology service, which is based on the use of DDC as a central spine. In the first strategy, different KOS could be mapped from the top levels of DDC. This is known as high level mapping. Here, a user can begin to interact with the top three levels of DDC concepts to find mapped concepts from other KOS, and then the user could go into the

hierarchies of other vocabularies, and be navigated through those hierarchies to find relevant information. Although top level mapping has some limitations due to the lack of 'depth' in the mapping, it is relatively quick and cheap to perform. The second approach is to establish deep mappings. In this approach, mappings occur at deeper levels in the hierarchy of the classification schemes. This is a more complex approach and the mapping itself is a time consuming process. However, although this approach is more costly, it requires less user-interaction to be built into any resulting service as user navigation of hierarchies would be unnecessary.

In addition, when establishing the mappings, according to different types of KOS, Interviewee 8 emphasised the importance of distinguishing different ways to represent concepts. For example, the key way to represent a thesaurus concept is based on the use of preferred term, the way to represent a classification concept is based on the use of notation, and the method to represent a subject heading scheme is based on the use of pre-coordinated strings.

4.2.4 Mapping collaboration, mapping tools, and information architecture

The relationship between local libraries and Ex Libris is also helpful for the development of mappings between different KOS. It is desirable that different participants be involved in developing a mapping service through the intellectual mapping work. These might include content providers, portal service providers, libraries, and even end-users (Interviewees 1, 2, and 3).

In fact, some service providers, such as SirSi, and Ex Libris, have good relationships with different content providers, which potentially form a basis to fill the gap between the different controlled vocabularies used by library communities and commercial publishers (Interviewee 2). Both these organisations have contributed to the mapping work between metadata schemes used by these two communities (Interviewees 1, and 3).

On the other hand, the services that these service providers have established, such as meta-search engines, terminology server, service description registry, metadata scheme registry, OpenURL resolver knowledge base, citation analysis service, name authority service, etc, could potentially be reused by other services. It is therefore important to publish these services in the semantic web-enabled format (Interviewees 2 and 3).

In SirSi Rooms, librarians can create a subject-specific interface called SubjectRoom, where the local librarians could set up their local terminologies, and related functionality tailored to their local subject requirements (Interviewee 2). The experience of self-arranging the local terminologies is very valuable. Likewise, it is possible to allow the librarians to create the mappings from their own terminology to other vocabularies. In addition, it is necessary to provide some basic terminology resources and a range of guidelines to standardise the terminology generation for the Subject Room. The most important point is how to provide a basic and general terminology resource for the subject librarians.

From the content provider side, Interviewee 3 mentioned that it is very difficult to encourage content providers to get involved with these kinds of mapping projects, because they are protective of their schemes and interfaces. Thus, she suggested that different mapping methods and access protocols need to co-exist in a terminology service in order to access different content providers' terminology databases. Similarly, Interviewees 3, 5, 6, 7, and 8 emphasised the importance of distributed mapping work. In the project conducted by Interviewee 4, the mapping service just provided a central DDC spine and a range of guidelines for each of its participating providers, and the providers followed the guideline when creating mappings between their own KOSs and DDC provided. This project provides good experience on the viability of distributed mapping work. However, Interviewees 5 and 9 criticised this distributed model for terminology mapping, because it requires the efforts of different communities, which might cause inconsistency. Interviewee 9 pointed out that a centralised structure for terminology mapping is advantageous, and that local services are not required to create any mapping

work. What local services need to do is to combine API functions provided by a centralised terminology service in various ways to suit their local subject needs.

For this reason, in the project conducted by Interviewee 5, a centralised terminology database is used to store all the KOSs and the mappings between them. The terminology service provider has to be forced to create or collect all the mappings by themselves. In this context, they face a range of up-date problems. Interviewee 5 mentioned that the best method to solve an update issue should be based on a distributed solution and should depend on the individual suppliers providing and maintaining their mappings to the mapping service. Individual suppliers can provide their mappings to the central terminology service.

Interviewee 8 suggested that the terminology service should be created centrally, and hold different controlled vocabularies centrally. In this context, it is possible that in the future, a lot of content providers may not hold their controlled vocabularies locally, but submit their vocabularies to a central terminology service, through which they could manage their controlled vocabularies. Importantly, the central terminology service should support a range of M2M functions to be able to interact with different services and systems. When using a centralised database to hold all controlled vocabularies, local controlled vocabulary databases may still exist, and there might be some contradictions between different versions of a controlled vocabulary. Interviewee 6 emphasised the importance of creating a KOS registry to offer metadata and access to different KOS integrated in different terminology resources.

Interviewee 8 also presented a range of advantages of this kind of centralised terminology services. For example, if each content provider does not hold its own terminology locally, and is entirely dependent on the central terminology service, then different content providers could share a single controlled vocabulary, so there is no versioning problem. Content providers could also give the terminology service maintainers suggestions about how to update the terminology they use.

In addition, Interviewee 8 pointed out that a mapping service is another option, in which the mapping server does not hold the controlled vocabularies, but facilitates exchanging terminological data between different controlled vocabularies. However, when using a distributed mapping server to exchange the information between different KOSs, one important issue is the difficulty in maintaining and creating the mappings. It is necessary to create a mapping from a combination of terms in one KOS to a single term in another KOS. In this situation, the distributed mapping would be difficult. In addition, some content providers may not be willing to map their controlled vocabularies into the terminology service, and some of them may even think that DDC is not appropriate for their subject requirements. In this case, Interviewee 6 pointed out that it was necessary to develop or apply appropriate protocols and queries to directly access the remote KOS, which could automatically create the mappings on the fly. Because of the variety of different mapping work provided by different communities using different strategies, Interviewee 6 pointed out that it is important to be able to develop a metadata scheme to characterise mapping sets – provenance (source), method (intellectual, co-occurrence, other automatic, etc), and perhaps a quality indicator.

Interviewees 4, 5 and 9 mentioned that manually mapping different vocabularies from DDC is very labour-intensive and time-consuming and that it would be important to investigate the use of co-occurrence mappings. Some union cataologues, such as OCLC WorldCat, are good resources for co-occurrence mapping. It is important to encourage further collaboration between the terminology services and some union catalogue providers.

In summary, the most important point is to consider whether a mapping project is based on the distributed mapping work, whether the service can support versioning functions, and how to encourage different participants to get involved in distributed mapping work.

With the above in mind, it is important to develop or select some mapping tools to support distributed mapping work. Interviewees 4, 5, 6, and 7 highlighted the importance of the distributed nature of a mapping tool. This is a very complicated tool for

terminology mapping. Interviewees 4 and 5 highlighted that the mapping tools can be based on the use of some encoding formats to improve the reusability of the mapping work. In other words, when the mappings have been done in a database, it should be possible to use various relevant representation formats to wrap the mappings. Thus, it is important to consider whether the mapping tools can support the distributed mapping work, and whether the mapping tools can provide a flexible API to support more functionality in the future.

4.3 Interview findings related to semantic web-related issues

4.3.1 Protocols

The protocols to access a terminology service are similar to the protocols to access a metadata repository (Interviewees 4, 5, 6, 7, and 9), see Section 4.1.2. The main issue when accessing the terminological data focuses on whether the selected protocol is sufficient to support text-based retrieval or triple-based retrieval. The text-based query focuses on providing many powerful query types, such as phrase queries, wildcard queries, proximity queries, range queries and so on. The triple-based retrieval is based on facilitating logic reasoning among different vocabulary terms and their relationships.

As Interviewee 6 indicated, the SRW protocol, which is used to facilitate text-based retrieval, is not sufficient to support richer browsing interfaces, and the SKOS API (or some derivative) offers more capabilities since it is specialised for a variety of use cases, such as accessing vocabularies, and can therefore return composite elements corresponding to the use cases (Interviewees 6 and 7). For example, the SKOS API can provide functions to allow a user to get all data related to a concept, and allow a user to get a concept to relate to another concept and get the metadata about a specific concept scheme. Methods are available to get a concept (passing a URI and retrieve the concept data object), get a concept by preferred label, get the relatives of a given concept, pass a keyword to search a concept, etc. However, further development of the SKOS API is required. Care needs to be taken when using this API, because a lot of functions provided in the API are not mature (Interviewee 7). Also, Interviewee 7 added that the problem

with the SKOS APIs is that the APIs are very generic, and it is not targeted to any particular application. Different applications have different needs. In a particular case, this general APIs may be not appropriate for the particular needs. For example, in a project, SKOS Core Vocabulary may be adapted to have a label called "scientific label", but the relevant SKOS APIs may not be able to process this label. Interviewee 7 hoped to use a number of relevant RDF-based query languages, such as S PARQL, to further develop this APIs (Interviewee 7).

In addition, if the service uses some triple-based protocols to retrieve information, it is required to convert all the terminological information into RDF-based formats, which imposes high overhead (Interviewee 5 and 9).

4.3.2 Representation formats

Interviewees 3, 5, 7, 8, and 9 highlighted that it was impossible to use only one representation format to encode all the controlled vocabularies, because different controlled vocabularies have their own structures and syntax, and were developed in different organisations. More importantly, different representation formats can be converted into each other depending on specific requirements (Interviewee 6 and 7). In this context, Interviewees 6 and 7 point out that it is easy to use XSLT to covert the terminological data into the SKOS format. In the SWAD Project, Deliverable 8.8 called "migrating thesauri to semantic web" offers a range of stylesheets to convert different XML formats into SKOS. However, in fact, some content providers may not allow direct conversion of their thesauri into a consistent representation format, and there might be some copyright issues. In particular cases, it is possible to temporarily convert some restricted thesauri into an encoding format just for on-the-fly querying (Interviewee 5). The converted results would not be stored in the terminology service permanently.

According to Interviewees 6, 7, and 8, of utmost importance in developing an encoding format is whether the encoding format can properly represent concepts rather than terms. In real-world cases, there are some encoding formats that are designed to represent terms,

e.g. Zthes XML DTD. Zthes XML DTD cannot represent concepts and the relationships between them very easily. Compared with the Zthes DTD, SKOS is a concept-based representation language, and designed to encode the thesauri (Interviewees 3, 6, 7, 8, and 9). Interviewee 6 indicated the importance of developing a both concept-based and term-based encoding format to represent KOS data.

When applying SKOS standard in a project, it is worth noting that there are three important parts to SKOS, which are SKOS-Core, SKOS-mapping, and SKOS API (Interviewee 7). SKOS-Core has been developed and is relatively mature. Today, there are a number of tools compliant with SKOS-Core, such as THmanager, and Tema Tres. SKOS-mapping is still in the proposal stage (Interviewees 4 and 7). According to Interviewee 7, how SKOS-mapping can be derived from RDFS is still an issue that needs to be considered further. The SKOS-mapping vocabulary is not very mature, and a number of issues are being discussed and explored. When using SKOS-mapping to represent the mappings, it is important to avoid using Boolean combinations, because there is no clear suggestion on whether to use Boolean operators to combine the concepts within the instruction document of the SKOS Mapping Vocabulary (Interviewee 7).

Interviewees 4 and 7 pointed out the extensibility of SKOS. Because a lot of thesauri and subject classification schemes have their own particular structures, some representation formats may not be able to represent these KOSs properly. In this situation, SKOS has great extensibility, because it is derived from RDFS. It can be extended to describe particular subject structures, and even an ontology. In this context, combining SKOS with OWL is very promising for the further development of a terminology service mapping KOS to ontologies. In addition, Interviewees 7 and 9 emphasised that SKOS is a W3C standard that were designed to provide better support for a more distributed architecture.

On the other hand, Interviewee 7 pointed out a disadvantage of SKOS. He mentioned that because the SKOS is derived from RDF that is based on open data, and because RDF is not set up to express strict data constraints on the data model, there is no natural way to validate the values within SKOS data elements. However, XML (e.g. XML DTD, XML

Schema, etc) can validate the instances and document type. Therefore, using Zthes XML DTD and MARC is good for validating the data. RDF validation is not about validating RDF data, is about drawing inference from the RDF data, so when using SKOS, it is harder to constrain the data model, and validate the data using a standard tool. In most cases, someone may want to publish the data and do some quality control of the data they published.

Interviewee 6 pointed out that SKOS was not specifically designed to represent classification data. When using SKOS to encode classification data, it is important to adapt the SKOS to encode the classification work.

4.3.3 Query language

Because SKOS is based on RDF, Interviewee 7 strongly suggested the use of SPARQL with the SKOS data to provide subject cross-browsing. He mentioned that SKOS is a standard format, and SKOS is not only an exchange format, but also a descriptive language with great extensibility. When using SKOS, it is hoped to use the standard methods to interact with the SKOS data. SPARQL is a useful standard way to query the SKOS data (Interviewee 7). The limitation of SPARQL is that it is not good at indexing text-based information (Interviewee 6 and 7). In the real case, it would be preferable to develop direct programmatic access to different RDF data sources (Interviewee 7). It is important to create a SPARQL server between the library portal service and the SKOS data (Interviewee 7). See Figure 4.3.

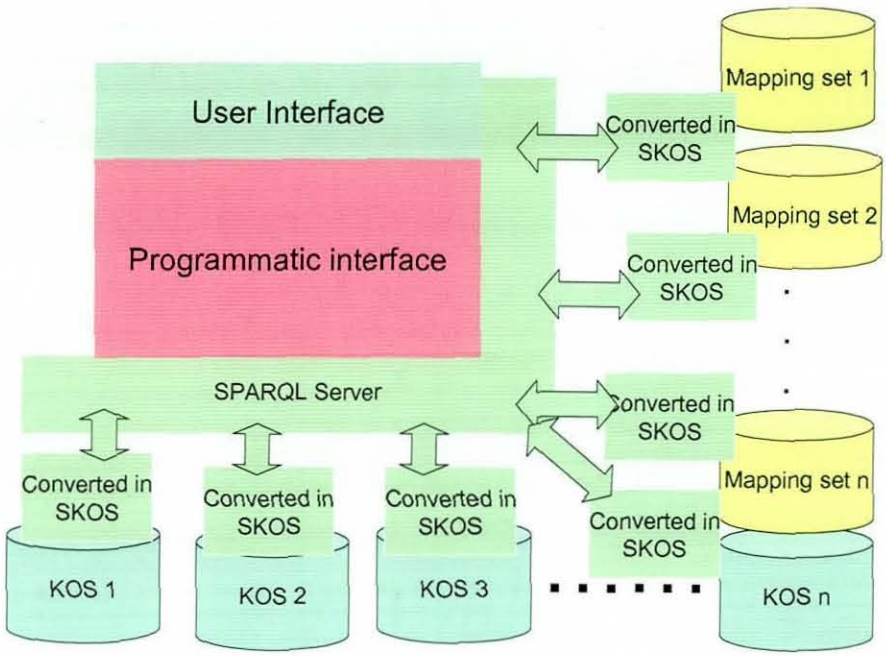


Figure 4.3: A SPARQL service

Apart from using SPARQL to access the RDF data, SKOS API (or some derivative) would offer more capabilities since it is specialised for a variety of use cases accessing vocabularies and can return composite elements corresponding to the use cases (Interviewee 6). According to Interviewee 6 and 7, in some cases, it is desirable to directly store the terminological data in a RDF triple store database, and use the SKOS API to query the database. The SKOS API makes the query expansion more flexible, and provides a richer browsing interface by using query expansion.

In addition, Interviewee 8 mentioned that SPARQL is designed to query the RDF language, and, for the purpose of retrieving the SKOS data, it could be used effectively. Compared with SPARQL, CQL is a query language associated with the SRW/SRU. A lot of SRW/SRU web applications are meant to be searched in a simple way. Therefore CQL is based on the field-based, Boolean retrieval system in some very defined ways, and does not have much capability to deal with knowledge-based semantic vocabularies.

It is true that SPARQL is not sufficient to support text-based retrieval (Interviewee 6). However, in this particular study, the main purpose is to provide a subject cross-browsing

interface for the end-users, which does not require the users to input a lot of text-based query. Rather, the users only need to select some pre-arranged subject term labels, and find relevant information. For this reason, the use of SPARQL is appropriate (interviewee 7).

4.4 Unanticipated issues

A number of research elements may be related to the development of a middleware system between different terminologies for providing subject cross-browsing services, some of which were not identified from the literature review. A number of unanticipated issues were found during the interviews, which are listed as follows:

1. It is very important for terminology services to develop a range of terminology resource selection criteria to decide which terminology resource should be added in their systems. Some terminology resources may not be worth being integrated into the portal system. It is important for a terminology service to collect a number of widely-used controlled vocabularies, and create mappings between them.
2. Although it is beneficial for some content providers to enable their terminology to be accessed, which can help the users find their information (Interviewee 5), in fact it is very difficult to map different KOS used by different content providers together, because resistance from them is a big issue (Interviewee 4). Thus, how a terminology service can get access to different vocabularies, by programmatic access to them, or by harvesting them, or by obtaining paper-based vocabularies, and then publishing them into digital formats needs to be considered;
3. On the end-users' side, to some extent, an institution can develop two portal systems based on different searching philosophies—a distributed cross-search service based on the Z39.50, SRU/SRW, etc, and an OAI data warehouse cross-search service based on the use of OAI-PMH. The OAI-based service can manage the information resources from institutional repositories, e-print archives, OAI-compliant databases, etc, and the distributed search service can deal with the non-OAI materials. Potentially, it is desirable that some terminological resources can

be harvested by the OAI-PMH so that mappings between the controlled vocabularies harvested can be established centrally. In summary, the OAI approach and the distributed approach should be developed together, and complement each other. (Interviewee 4)

4. Content providers are over confident about the KOSs they have, because they are continuously developing their KOS closely with the growth of their own metadata. Therefore, it is important to do some advocacy work to encourage them to share the controlled vocabularies with others (Interviewee 6).
5. Licenses outlining the use of controlled vocabularies vary from one organisation to another. Thus, when establishing a terminology service, it is necessary to create a terminology registry to record this license information. This will inform terminology service staff how to use and map these controlled vocabularies properly (Interviewee 7).

4.5 Summary of interview findings

4.5.1 Federated search

There are a number of theories and technologies for the development of a federated search service in a library portal system. In order to support a sophisticated and elaborate subject federated search, it is preferrable to map different metadata schemes used by different services into a central metadata scheme (Interviewees 1, 2, 3, 4, and 9). In particular, mappings between subject metadata elements need to be created to facilitate subject cross-searching. This helps expose different subject metadata elements to the cross-search engines. A data conversion and transmission programme should be developed to translate the users' query into different formats that specific databases can understand, convert heterogeneous metadata records into a consistent library-based format, and present a combined view of the records from different resources to the end-users (Interviewees 1 and 3). Because different online databases might accept different protocols to enable federated search engine query against themselves, it is necessary to develop a multi-protocol environment where a federated search engine could use different protocols to search different online databases (Interviewees 1 and 3).

4.5.2 Subject cross-browsing use case

It is inappropriate to directly use DDC as a subject cross-browsing structure (Interviewees 1, 2, and 3). It is important to give the local library enough flexibility to design their own subject portals (Interviewee 2). The local librarians should use their own terminologies to develop their own subject browsing interfaces (Interviewees 1, 2, and 3). In many cases, departmental structures and relevant modules could form a basis to create a local taxonomy (Interviewees 1, 2, and 3). By establishing the mappings or selecting some terminological information from these terminology services, the local subject librarians could also create mappings between their terminologies and several huge terminology resources, such as HILT, OCLC Terminology Service, etc (Interviewee 6).

The subject-browsing interface could just be a starting point to help users begin their subject navigation, and guide the users to jump to another concept scheme in another user interface (Interviewee 4). However, this would impose intensive user interaction efforts, and generate confusion over the different interfaces—this is not suitable for some inexperienced users (Interviewee 1). Thus, it is important to explore the possibility of making subject cross-browsing work with federated search engines that could offer a consistent interface for retrieving distributed information. It is expected that the users could retrieve relevant information from distributed information resources by just clicking concept terms in a local subject scheme.

Although the end-users in an institution prefer to interact with a subject cross-browsing interface using their own terminology, DDC provides an alternative browsing interface to support experienced users. These two subject cross-browsing structures (local terminology and DDC) could complement each other to suit different users' subject requirements.

4.5.3 Structural mode of terminology mapping

The central DDC spine approach used to create mappings between DDC and other vocabularies should reduce some of the mapping effort. Creating mappings between DDC and other vocabularies can result in the loss of precision and specificity, but in the current state, there was no other option that seemed more appropriate. In the current situation, there is no common ontology that could be used as a switch language (Interviewees 4, 6 and 8). Some ontologies, such as WordNet, etc, could be used to improve subject searching functionality, but not for subject browsing (Interviewee 8).

4.5.4 Mapping relationships

Broader, narrower, related, and equivalent relationships could be four basic mapping relationships (Interviewees 4, 5, 6, 7, and 8). When the users interact with the subject cross-browsing interface, it is preferable to let the users see the mapping relationships between terms from different vocabularies. In this case, the users could then make their own judgments as to whether the mapping provided could achieve their subject requirements. This would improve the interactivity between the users and the subject browsing interface.

Because of the pre-coordinated nature of DDC, it is important to consider ways to deal with compound concepts in DDC for the purpose of mapping. Boolean operators could be added to improve the accuracy of the mappings (Interviewees 6 and 8). The lack of real usage scenarios integrating the Boolean operators is the main issue. Thus, it is important to develop some use scenarios, in which Boolean operators are used to help create mappings.

4.5.5 Depth of mapping and elements to be mapped

Based on the use of shallow mappings, the users could interact with top levels of DDC as a starting point to find matched subject terms in other relevant subject schemes. The matched subject terms could help the users jump to other information resources using other schemes (Interviewee 4). However, when cross-browsing items in the collections, it

would not be useful for most users to start with a shallow mapping based on DDC, jump to another vocabulary for browsing, and re-conduct the search in another interface (Interviewee 6). This requires a lot of interaction efforts from users.

Despite this, shallow mappings might be useful at the level of discovering collections (Interviewees 4 and 6). Thus, it is argued that shallow mapping for subject cross-browsing is a good option for information collection discovery, but that deep mapping is appropriate to facilitate subject cross-browsing for retrieving item-level metadata records.

In different types of KOS, concepts are represented in various ways (Interviewee 8). When creating the mappings between one KOS and another, it is important to note these differences, and select appropriate elements to be mapped.

4.5.6 Mapping partnership

A variety of methods to create the mappings between different KOS have been conducted in different projects, and much mapping work has been done. This includes:

1. Co-occurrence mapping based on some union catalogues: OCLC uses its union catalogue WorldCat to generate mappings between LCSH and DDC, and LCC and LCSH (Interviewee 4);
2. Direct mapping: Mapping work has been done between GSAFD and LCSH (Interviewee 8);
3. Switch language: From DDC, mappings to the LCSH, LCC, MeSH, and NLMC have been done (Interviewees 4, 5, 6, 7, 8, and 9).
4. On-the-fly mapping: Different KOS could be put into a linked environment through relevant protocols, and different KOS could answer the queries for a given term. In this sense, mappings based on a given term could be created on-the-fly (Interviewees 4, 5 and 6).

Considering various terminology resources, three types could be summarised. First, a fair amount of terminological data has been stored in some terminology services, such as OCLC Terminology Service, HILT Terminology Service, STAR Terminology Service,

etc. It is important to develop programmatic interfaces to access these terminology services. Second, in any particular institution, a number of local KOS exist. It is also important to establish the mappings between these local KOS and the terminologies in different terminology services. Third, there are some KOS used by different information providers, which are not included in the terminology services, but these KOS need to be incorporated into the local information environment for subject cross-browsing.

A framework to create and collect mappings needed to be proposed. A variety of mapping data created or collected by different communities would form this structure. Figure 4.4 shows a basic structure of this framework.

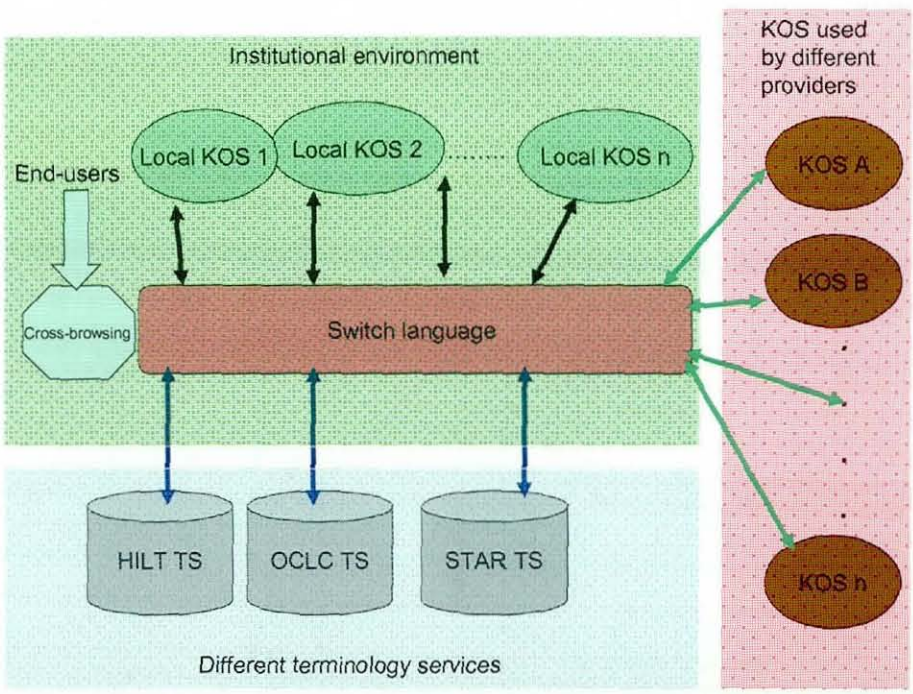


Figure 4.4: The integration of different terminology resources

In this mapping framework, obviously, a switch language is required to link different terminology resources, and exchange the terminological information between them. According to Section 4.5.3, DDC could be selected as the switch language to link different KOS and terminology services (Interviewees 4, 5, 6, 8, and 9). However, there might still be some KOS incompatible with DDC, and it may be difficult to establish the mappings between DDC and these KOS (Interviewees 4, 5, 6, and 8). In this case, it is

important to formulate appropriate queries to access these KOS and get the terminological data on the fly.

Because the local librarians from an institution or an information provider are familiar with their KOS, it is expected that the local librarians could establish the mappings between their local KOS and DDC (Derived from Interviewees 1 and 3), and the mappings constructed could be stored into a local database in an institution. In parallel, the terminology services could provide the mapping data to this institution through a middleware system. Thus, different types of mappings could complement each other to provide a fully-fledged subject cross-browsing service.

4.5.7 Protocols and encoding formats

Because different terminology services apply different access protocols, and different KOS used by different information providers are in different digital formats, it is impractical to force all KOS owners to use one access protocol and one exchange format for their KOS. In some cases, using XSLT could convert the KOS from one format to another (Interviewees 6 and 7). However, some KOS owners may not allow a third-party to hold their KOS in another database, or convert their KOS into another format. Also, the data conversion could cause data loss.

For this reason, it is necessary to develop a middleware system that could use different protocols to query all the terminological data from different KOS using different formats, and access protocols, and located in different servers (derived from Interviewees 1 and 3). In this middleware system, a data transmission programme could be developed to translate a user's query into different forms of queries that different KOS or terminology services could recognise, and gain the terminological information from different terminology resources (derived from Interviewees 1 and 3). When the terminological information is returned from different terminology resources, different exchange formats, such as MARC21 XML, Zthes XML DTD, SKOS, etc., might be candidates for representing the terminological information. Thus, a variety of data conversion

programmes could be deployed to convert different machine-readable terminological data into different human-readable presentation forms (derived from Interviewees 1 and 3).

4.5.8 Information architecture of the middleware system between terminologies

During the interviews, different experts had different opinions on whether a distributed model or a centralised model for this middleware system should be developed. However, it was widely-agreed that the centralised structure of a terminology service is advantageous as local services are not required to input a lot of intellectual effort, simply needing to combine functions in various ways to suit their local subject needs (Interviewees 5, 6, 8, and 9). This centralised structure could help achieve a long-term commitment for the development of a terminology service.

However, it is impractical for a terminology service to provide all mapping data. For this reason, a matrix that combines different mapping work from different communities was proposed in Section 4.5.6 to access different terminology resources. Section 4.5.7 indicated that the middleware system should be able to translate a user's query into different forms of queries that different KOS or terminology services could recognise, and gain the terminological information from distributed terminology resources. Thus, this middleware service should be based on a distributed model (Interviewees 1, 3 and 4).

These summarised points formed a basis for developing a theoretical framework. This framework will be discussed in the next chapter.

Chapter Five: Theoretical Framework Development

Based on the findings in Chapter 4, this chapter aims to develop a theoretical framework for a middleware system between terminology resources to facilitate subject cross-browsing functions in the context of library portal services, such as Ex Libris MetaLib, SirSi Rooms, etc. This chapter is organised as follows: Section 5.1 describes the strategy applied to establish the mappings between different KOS located in different terminology resources; Section 5.2 presents different technical components included in this framework system; Section 5.3 indicates the relationships between this framework system and other information services related to the development of library portal. These related services include meta-search engines, service registry, and distributed information resources; Section 5.4 discusses the issues related to the development of this framework.

5.1 Mapping strategy

5.1.1 Structural model for mapping

In this framework, it is important to exchange terminological information between different KOS. Although the method of one-to-one direct mappings between two KOS is very precise, it is difficult to be implemented in a large scale. Based on the findings in Section 3.2, a huge number of KOS need to be integrated into the middleware system in this research. A switch language approach is more appropriate. Based on the finding in Section 4.5.3, DDC as a switch language for creating the mappings could reduce intellectual efforts, and no other option has been identified, which seems better than DDC as a switch language. Thus, DDC is selected as a backbone structure, to which different KOS used by different information providers would be mapped.

Because a number of terminology services have been developed based on the use of DDC as a switch language, such as HILT TS, OCLC TS, etc., it is possible that this proposed middleware framework could be extended to permit the collaborative use of these existing terminology services for interoperability. Different terminology services could be accessed by this middleware system. It is therefore possible that each DDC concept

could be assigned an URI for identifying DDC concepts uniquely. Based on the use of same DDC concept URIs within different terminology services, different DDC-based terminology services could easily interact with each other for interoperability.

5.1.2 Elements to be mapped

According to Section 4.5.5, in different types of KOS, concepts are represented in various ways. In this research, the method to select the elements to be mapped is based on the BS8723-Part4. This is shown in Table 5.1.

Types of KOS	Conceptual elements to be mapped
Thesaurus	Preferred terms
Classification scheme	Notations
Taxonomy	Category labels, notations or identifier.
Subject heading scheme	Terms or pre-coordinated strings
Authority lists	Terms or identifiers
Ontology	Terms or identifiers

Table 5.1: Elements to be mapped

It is worth noting that different terminology resources, which can be integrated within this middleware system, may use different types of identification systems to identify their conceptual elements to be mapped. It is important to develop a unique identifier system for all the concepts beyond using the concept IDs from individual KOS. In semantic web environments, it is proposed that URIs should be assigned to most of these elements in different KOS. For this reason, the mappings between one particular KOS and DDC could be achieved through the use of concept URIs.

5.1.3 Treatment of compound concepts

Based on the findings in Section 4.5.4, DDC is a pre-coordinated controlled vocabulary that includes a number of compound concepts. When a post-coordinated vocabulary, in which most of its concepts are individual terms, is mapped with DDC, it is important to combine several relevant individual concepts in the post-coordinated vocabulary to map to one concept in DDC. For example, DDC concept "020.2854678 Internet—libraries" can be mapped to the combination of the UKAT concepts "libraries" and "internet". In theory, there are a variety of "combiners" that are able to combine different concepts from one vocabulary to another. These "combiners" mainly include Boolean Operators

(and, or, not), facets (time, place, people, event, etc), and ontological relations. Currently, how to use these "combiner" for mapping is still an open issue, and some combiners may make the mapping more complicated. With this in mind, these combiners are not used in this research. Instead, a number of relevant concepts are put into a "bag", and a bag of individual concepts can be mapped with an equivalent DDC concept. A bag of concepts could become a very abstract concept that may not have a clear meaning. In this case, it is not appropriate to use exact-match relationship to represent the mappings between a bag of concepts and a DDC compound concept. As a result, a major match relationship is employed in this research to represent a mapping between a compound DDC concept and a bag of several individual concepts. When the end-users get a list of mapped concepts through browsing, it is assumed that end-users can add appropriate Boolean operators to combine the terms in a comprehensive way to conduct their search further, or the users could delete some of terms in bag. See Figure 5.1.

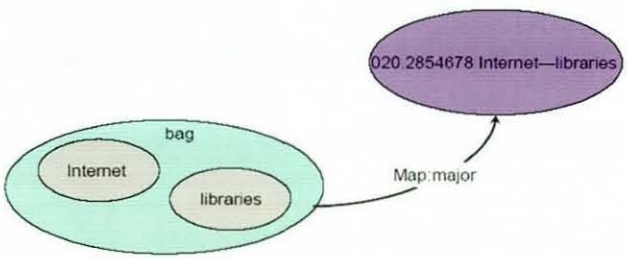


Figure 5.1: Treatment of a compound concept

5.1.4 Mapping relationships and logics

The findings in Section 4.5.4 suggested that broader, narrower, related, and equivalent relationships could be four basic mapping relationships. However, in many cases, these four relationships are not focused on dealing with the compound concepts. For this reason, a new mapping relationship called "major match" is introduced. The specific description of these relationships is listed as follows:

1. Exact match: DDC contains a concept identical in scope to the concept in a specific KOS used by other resources (BS8723-part4);

2. Broad match: a concept in the target vocabulary is a superset of a concept in the source vocabulary (Renardus mapping report);
3. Narrow match: a concept in the target vocabulary is a subset of a concept in the source vocabulary (Renardus mapping resport);
4. Major match: This is used to map one compound concept to more than one post-coordinated concepts in another vocabulary, because it is not suitable to use exact match to establish the mappings between a bag of post-coordinated concepts and one compound concept in different KOS. In this situation, the compound concept should linguistically equal to the combination of post-coordinated concepts;
5. Minor match: This relationship is used to state an associative mapping link between two conceptual resources in different concept schemes. (Alistair and Matthew 2004).

When mapping data is based on the co-occurrence mapping work, there is no specific mapping relationship identified between concepts in different vocabularies. In this situation, it is possible to put some additional intellectual mapping effort on selecting appropriate mapping relationships. This makes the co-occurrence mapping data more explicit and usable.

However, because the actual mapping work is more complicated, and a wide variety of situations may emerge, it is necessary to establish some logic to deal with different situations, and ensure the mapping is consistent. Derived from Doerr's (2001) theory, the mapping logic can be described as follows:

1. If a concept (C1) in the source vocabulary has an exact equivalence to one concept in the target vocabulary, the mapping between these two concepts could be created directly, and no other mapping need to be created further;
2. If a concept (C1) in the source vocabulary has no exact equivalence to any concept in the target vocabulary, then it is expected to find a number of concepts (Ca, Cb, Cc) in the target vocabulary that can be combined together. The combination of these concepts (Ca, Cb, Cc) in the target vocabulary can be linguistically equal to the concept (C1). In this case, it is desirable to put these

relevant combined concepts of the target vocabulary into a "bag", and make this bag major-map to the concept C1;

3. If a concept (C1) in the source vocabulary has no exact equivalence to any concept in the target vocabulary, and it is difficult to find a number of concepts in the target vocabulary that can be combined together to major map to C1, then for this concept C1, it is desirable to find at least one broader equivalence or at least one narrower equivalence in the target vocabulary. The broader equivalence should be minimal, and the narrower equivalence should be maximal;
4. If a concept (C1) in the source vocabulary has no exact equivalence to any concept in the target vocabulary, we cannot find a number of concepts in the target vocabulary that can be combined together to major map to C1, and we cannot find at least one broader equivalence or at least one narrower equivalence in the target vocabulary, then it is expected to find some related concept(s) in the target vocabulary, and minor-map to the concept C1. The number of related concepts in target KOS can be one or more. If the number is more than one, we need to combine the concepts into a bag, and minor-map this bag to the concept C1;

Based on the finding in Section 4.5.4, it is desirable to allow the users to see the mapping relationships between terms from different vocabularies, and make a *judgement on if the mappings provided are useful for them. In this framework, the* mapping relationships will be displayed with the mapped concepts.

5.1.5 Mapping between DDC and local taxonomies

Based on Section 4.5.2, it was found that many users find it difficult to follow DDC structure to find relevant information, and institutions should use or develop their local taxonomies for subject browsing. A subject cross-browsing service should be an application that seamlessly combines different terminological resources with a locally-tailored browsing structure. Thus, this framework plans to establish the mappings between different vocabularies and a local directory through DDC spine, and present a local directory to end-users for subject cross-browsing structure. However, mapping local

taxonomies to different KOS via a DDC spine may cause one-step indirection problems and the precision may be lost in this process. In order to minimise the loss of prescision, only "exact-match" is allowed to be used when mapping local taxonomies to DDC. No other mapping relationship is used in this stage. Appendix 4.1 is an example, in which a number of mappings between an information taxonomy and DDC were established.

Derived from the findings in Section 4.1.1, this framework was developed as a vendor-hosted and a localhosted terminology mapping service. In other words, the vendor will provide DDC-based cross-browsing APIs and detailed information of DDC concepts to each library. Each library has its own ability to map its own taxonomy to DDC. As shown in Figure 5.2, the vendor will host a terminology mapping database, in which mappings established from different KOS to DDC are stored, and each organisation can also host the same DDC-based terminology mapping database as the vendor's. In addition, the mappings established from DDC to its local terminology should be stored in the local database. In this way, DDC will become a mediator between the local terminology and different KOS sources.

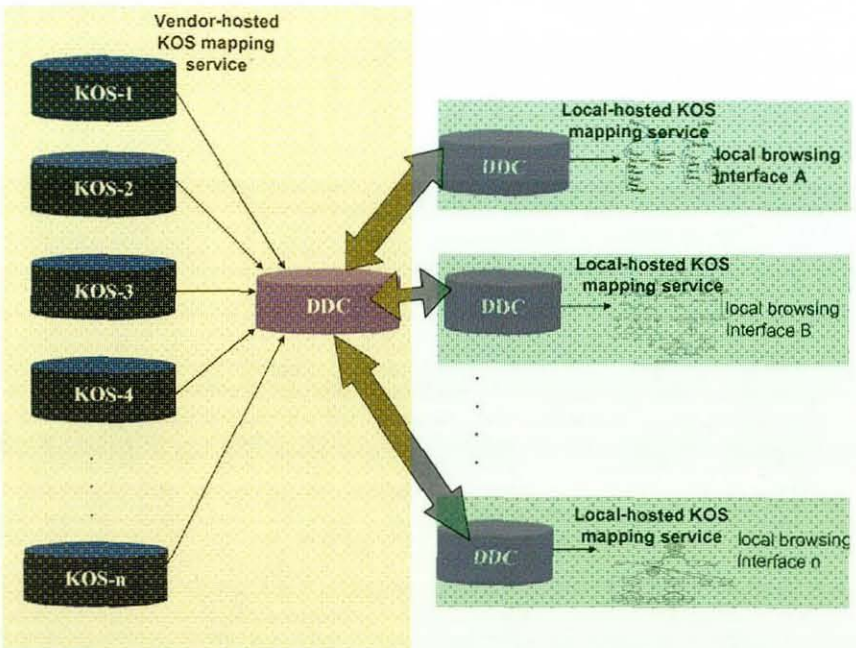


Figure 5.2: A locally-hosted and vender-hosted architecture

Local users, therefore, can not only use the local taxonomy for their subject navigation, but also browse DDC to find the relevant information. This would be extremely useful for users who are familiar with DDC structure.

5.1.6 Mapping work collaboration

As argued in Section 4.5.6, establishing the mappings is a tedious process, and it is helpful to distribute the work into different groups. All the groups can work together based on a range of consistent mapping guidelines. In this framework, it is proposed that different participants could be involved in developing a mapping service and establishing intellectual mapping work. These participants might include KOS owners, terminology service providers, middleware service providers, and libraries. Thus, in this framework, it is proposed that

1. A number of terminology services could create and provide mapping data to this middleware. However, different terminology services may use different mapping strategies to create their mappings. It is important for some terminology experts employed by this middleware system provider to validate the mappings provided by different terminology service providers.
2. Some KOS owners can contribute to generating the mappings between DDC and their own KOS for this middleware system. However, in fact, it was found in Section 4.4 that most KOS owners feel over confident on their own KOS. It is difficult to encourage them to create mappings between their own KOS and other KOS.
3. According to all the mappings, a local librarian can establish his/her own mapping between his/her own local taxonomy and DDC with validation by terminology experts.

5.2 Framework architecture

Based on the findings in Section 4.5.7 and Section 4.5.8, as well as the semantic heterogeneity between different KOS, there are also a number of other factors that challenge interoperability between different KOS. These factors mainly include:

1. Multiple formats;

2. Multiple access protocols;
3. Multiple metadata schemes to describe different KOS;
4. Multiple terminology service systems where different KOS are located.

According to Section 4.4, in many cases, terminology resources providers may not allow a third-party to hold their KOS and mappings, or convert it into a unified format. For this reason, the proposed middleware framework is developed to cross-access terminology data from terminology resources that use different formats, access protocols, schemes and that are located in different terminology services.

With this in mind, the basic principle of this framework is to employ a variety of agents to query various terminology resources. As shown in Figure 5.3, there are several steps that are described as follows:

1. Different KOS from different terminology resources will intellectually be mapped into DDC based on the use of designed concept URIs;
2. The mapping data is put into SKOS-Mapping format. This is because SKOS is not only an exchange format, but also a descriptive language with great extensibility. It can represent various mappings more precisely. This is derived from the findings in Section 4.3.2;
3. An RDF API called JENA RDF API is adopted to establish a DDC-based cross-browsing API above various SKOS-mapping datasets between DDC and different vocabularies. In JENA, the SPARQL query could be constructed to query against the SKOS mapping datasets. This is derived from the findings in Section 4.3.3. When a mapping between a DDC concept and another non-DDC concept is found through using DDC-based cross-browsing API, a URI of the mapped non-DDC concept should be retrieved, the URI could be used as a query to search against relevant terminology resources, and then detailed terminological information about this concept and its semantically closed concepts could be returned from the terminology resources, and be presented to the users;
4. Below the different SKOS-mapping datasets, a knowledge base is developed to store the interaction information about the type of protocols that distributed

terminology resources support, the formats that the terminology resources use, the form of the queries to access different KOS, and the formats of results that are retrieved based on the relevant protocols. The knowledge base will inform the system how to use different agents to access different terminology resources. In other words, this knowledge base is responsible for translating a user's query into different forms of queries that different KOS or terminology services could recognise, gain the terminological information from different terminology resources. Also, this knowledge base can be used to convert the returned results into a consistent format, and display the converted results to the users. Based on the knowledge base, a number of appropriate agents are employed to translate a user's query, and a number of XSLT files should be developed for data conversion. This is derived from the findings in Section 4.5.1;

5. Because different KOS owners have different licenses regarding the use of their KOS, it is possible that some KOS can be stored and set up in the local environment of this middleware system, but some must be accessed remotely.

The framework is shown in Figure 5.3.

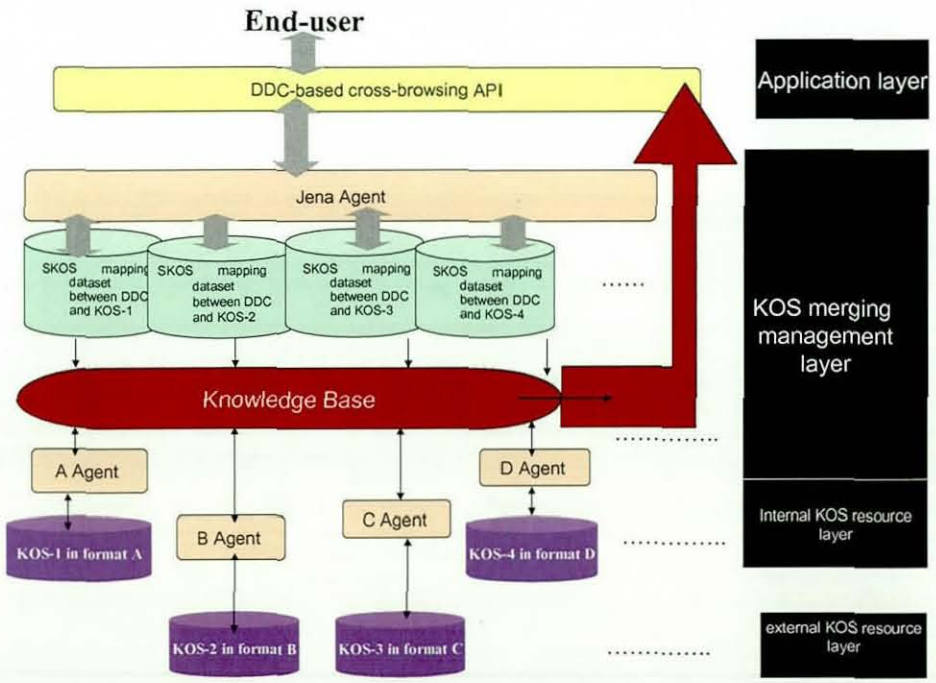


Figure 5.3: The basic architecture of the research framework

In Figure 5.3, it is assumed that the owner of KOS-2 does not allow a third-party to hold its KOS, so its KOS has to be accessed remotely. Three terminology services need to be accessed remotely through their access protocols. But the owners of KOS-1 and KOS-4 allow other systems to hold their data in another place. In the situation, KOS-1 and KOS-4 can be set up in the local system. This framework is able to access a number of terminology servers, such as HILT, OCLC Terminology Service, STAR Terminology Service, etc. For example, a SRW client could be set up in this middleware framework to access the HILT SRW server, and get the wrapped SKOS data. When the SKOS data is gained, a SKOS API-based agent would be employed to process the returned SKOS data, and convert the data into human-readable format.

5.3 Framework and meta-search engines

As highlighted in Section 4.5.2, it is desirable to make the subject cross-browsing service interact with federated search engines. By now, this proposed subject cross-browsing service can only return the relevant conceptual terms, but end-users may be more concerned with gaining the relevant metadata records through subject cross-browsing services. It is important to make the subject cross-browsing framework interact with meta-search services provided by library service vendors.

In a meta-search service, a user can send a query to numerous information resources. The query is broadcast to each resource, and results are returned to the user. Thus, it is useful that through interacting with the subject cross-browsing interface, the mapped conceptual terms from a particular KOS can become queries against the specific databases that are indexed by this KOS. In this context, a database registry that records the usage of KOS in different databases should be developed.

The purpose of this registry is to make the mapped conceptual terms from one particular KOS become meta-search queries against the specific databases that are indexed according to this KOS. See Figure 5.4 as an example. The specific steps are described as follows:

1. Users interact with a subject cross-browsing interface, and get a number of mapped conceptual terms from various KOS;

2. The database registry is used to make the federated search service use mapped terms to search against the relevant databases using these mapped terms. Thus, a number of item-level metadata results could be returned from different databases;

3. The item-level metadata results returned will be converted into a consistent format, and presented to the end-users;

4. A ranking algorithm should be developed based on the five different types of mapping relationships.

In this manner, an end-user can get the item-level metadata results by using the subject cross-browsing service and meta-search engines.

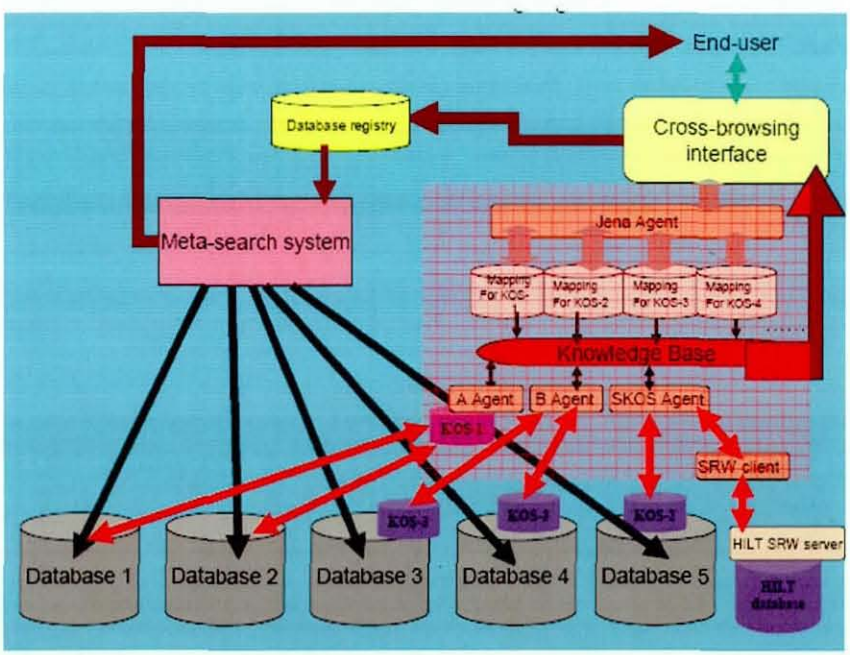


Figure 5.4: M2M interaction between the framework and a metasearch engine

In the database registry, a simplified RDF schema has been developed to describe the use of various KOS in different databases that can be accessed by a specific meta-search engine. The specification of this schema is listed in Table 5.2:

Element	Explanation
lsls2:database	Is a rdfs:class
lsls2:isindexedby	Is a rdfs:property lsls2:isindexedby rdfs:domain lsls2:database lsls2:isindexedby rdfs:range skos:conceptScheme this element is aimed to record the usage of various KOS in different databases that can be accessed by a specific meta-search engine.
dc:title	Database name
dc:subject	Database subject coverage, and recommend to use DDC number
dc:date	Database starting time
dc:format	Document format in the database
dc:language	Database language
dc:copyright	Copyright
dc:publisher	Database publisher
dc:identifier	Database URL
dc:type	The type of database

Table 5.2: The RDF schema used to describe the usage of various KOS in different databases that can be accessed by a specific meta-search engine.

A simple metadata record using this schema is listed as follows:

```
<?xml version="1.0" encoding="UTF-8" ?>
    <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-
syntax-ns#" xmlns:rdfs="http://www.w3.org/2000/01/rdf-
schema#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:lsls2="http://www-staff.lboro.ac.uk/~lsls2/scheme#">
  <lsls2:database rdf:about="http://pubs.cs.uct.ac.za">
  <dc:title>UCT CS Research Document Archive</dc:title>
  <dc:description>UCT CS Research Document Archive is an
institutional repository in the department of computer science,
at University of Cape Town</dc:description>
  <dc:language>en</dc:language>
  <lsls2:indexedby
rdf:resource="http://www.acm.org/class/1998/acmccs98-
1.2.3.xml" />
  </lsls2:database>
</rdf:RDF>
```

In this manner, by using the subject cross-browsing service and the database registry, a metadata search engine can recognise how to broadcast mapped concepts returned from the subject cross-browsing service to query different databases indexed by using these mapped concepts. It is a simple process that requires a fairly low degree of interaction

from users to progress from the stage where they browse the hierarchy to getting the relevant information items. In addition, the databases can be also indexed according to the local taxonomy, which forms a basis to provide a subject browsing interface to find relevant databases.

5.4 Further consideration

In this framework, it is proposed to provide the users with a subject cross-browsing interface. Different from other subject cross-browsing services that guide the users to identify relevant collections, users could be navigated by this particular service to find item-level metadata records from distributed information resources. Based on the description in Section 5.1.5, it is easy for end-users to browse their own terminology, but get the item-level results from distributed information resources. However, there are a number of issues that need to be considered.

Firstly, because it is necessary to exchange terminological information between a local taxonomy and different KOS through a DDC spine, it is difficult to solve the indirection problem. It is possible to lead to the loss of precision during this stage. For example, a term "house cat" in a local KOS can only be mapped to DDC concept "cat", but in another vocabulary, there is an exact term called "house cat". In this situation, precision is lost. The relevant solution will be discussed in Section 8.2.

Secondly, establishing and maintaining the mappings is possibly the most important barrier to the development of the middleware system. Using existing mappings from other terminology services seems acceptable. Thus, it is important for this middleware system to create a programmatic interface to access different terminology services. However, mappings from different terminology services may be inconsistent, because different terminology services use different methods to create or collect their mapping data. It is important to explore the solution to this inconsistency problem.

Thirdly, in this research, a collection registry and meta-search engine are proposed to interact with the framework. It is possible that other shared services could interact this

framework in Machine-to-Machine ways to improve the functionality of a library portal service. Thus, it is important to explore if other shared services could work with this framework to enhance the functionality of this framework.

Fourthly, in this middleware, the returned terminological information comes from different types of KOS, such as classification schemes, subject headings, thesauri, etc. It is important to consider various ways to present a concept to the end-users, and realise that different concepts in different types of vocabularies should be presented in different ways.

Fifthly, in some cases, a number of KOS in a terminology service might not be able to be mapped from DDC. In this situation, it is possible to use the caption term in a subject cross-browsing as a query against the terminology service, and get the results on the fly (see Section 4.5.6).

With these issues in mind, it is necessary to develop a software prototype system based on this theoretical framework. This prototype system is used to assess if the proposed theoretical framework could meet certain desired goals. Also, it is important to use this prototype system as a testbed to address the issues listed above and find out the potential problems with the relevant solutions. Hence, the next chapter will introduce the main principles for the development of a software prototype system based on this theoretical framework.

Chapter Six: Software-based Prototype System

This chapter aims to describe a software-based prototype application that seamlessly combines the terminological resources from more than one KOS into a switch scheme for subject cross-browsing. This prototype system is based on the developed theoretical framework in Chapter 5. However, this prototype system is a simplified version, because it is impractical to incorporate all related KOS into a system in any study. A long-term commitment is very important for the development of this kind of middleware service between different terminologies. Ideally, there might be a number of teams with sufficient expertise that have the responsibility for maintaining and creating the mapping, providing different terminology mapping datasets for the future.

6.1 Selection of vocabularies

Based on the description in Section 5.1.1, DDC is applied as a backbone structure to exchange the terminological information between different KOS. Also, by browsing DDC structure, users could be navigated to find relevant concept information across different KOS. This prototype has been developed in the field of library science and computing science (DDC 000-099). The selection of different vocabularies for mapping was purposive. DDC is a pre-coordinated classification. It is important to explore methods to create the mappings between a pre-coordinated KOS and a post-coordinated KOS, and investigate the ways to combine several individual concepts into a bag, and map a compound concept into this bag. Thus, a post-coordinated thesaurus "UKAT" was selected as a KOS to map from DDC, and a classification "ACM Computing Classification" is also mapped from DDC. These two KOS greatly differ in their degrees of coordination, and semantic structures. This is aimed to test if the identified mapping strategy in Section 5.1 is appropriate for establishing mappings between a variety of KOS.

The proposed theoretical framework planed to establish the mappings between DDC and the local directories, and present a local directory to end-users. Thus, it is assumed that an information science taxonomy developed by Hawkins (2003) is effectively used in a particular organisation as a subject browsing structure. It is required to map this

taxonomy to DDC for subject browsing. The mappings have been established as shown in Figure 6.1.

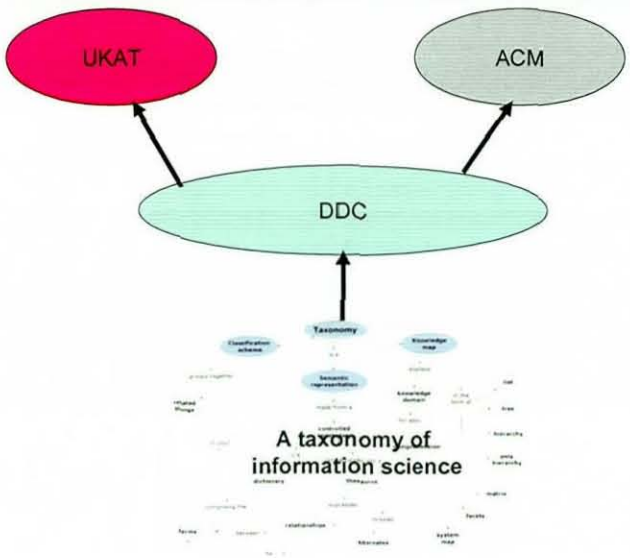


Figure 6.1: The mappings between the vocabularies used in this prototype

As mentioned in Section 2.4.1.3, it is important to identify mapping directions before conducting mappings. Clear mapping directions can form a basis to create clear mapping guidelines for different mapping participants. As shown in Figure 6.1, in this research, when establishing the mappings between UKAT, ACM, and DDC, DDC was selected as a source vocabulary, and UKAT and ACM are target vocabularies. However, when the local information science taxonomy became a structure for subject browsing, DDC became the target vocabulary, to which the information taxonomy is mapped.

6.2 Mapping strategy

Based on the description in Section 5.1, a number of rules are used to create the mappings between DDC and other selected vocabularies. These rules are listed as follows:

1. Based on the description in Section 5.1.2, the preferred terms in UKAT are identified as the basic elements for mapping, and the notations in DDC, ACM,

and the information science taxonomy, are identified as the basic mapping elements;

2. A bag is used to combine a number of related UKAT concepts, and it can be mapped from a compound concept;
3. The five mapping relationships in Section 5.1.4 are used to create the mappings;
4. The mapping logic described in Section 5.1.4 is used during the construction of mappings;

However, this research is an individual work, and it is impractical to invite a number of participants to collaborate for the construction of various mappings. Thus, the feasibility of mapping collaborative workflow developed in Section 5.1.6 would be verified by asking a number of relevant questions to the expert evaluators. These questions were incorporated in Appendix 5.5 "final general interview questions".

6.2.1 Mappings between DDC and UKAT

Because DDC is a pre-coordinated classification scheme, and UKAT is a post-coordinated thesaurus, the basic mapping elements in DDC would be DDC notations, and the mapping elements in UKAT would be the preferred terms. It is important to develop methods to handle the mappings between a compound DDC concept and several UKAT terms. In this situation, based on the description in Section 5.1.3, a bag can be used to combine the UKAT concepts. The mapping between a DDC concept and the bag could be established, as shown in Figure 6.2. It is assumed that when a number of terms in a bag are retrieved and presented to users, users would be able to further adjust their query by deleting some of irrelevant terms in the bag or by adding the Boolean operators to further combine the terms.

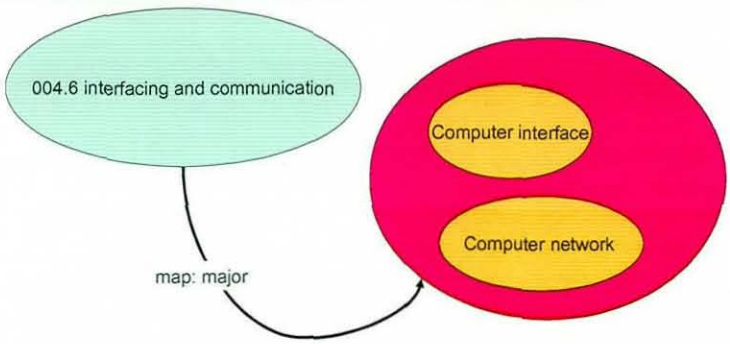
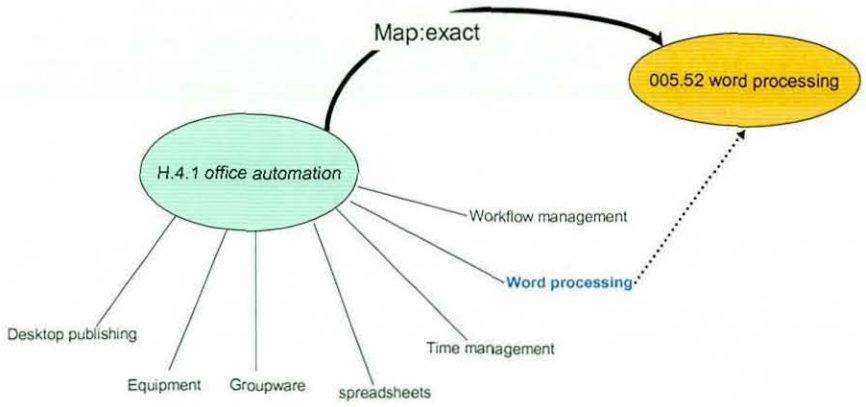


Figure 6.2: The way to handle a compound concept

The mappings between UKAT and DDC in the field of information and computer science were established and listed in Appendix 3.1.

6.2.2 Mappings between DDC and ACM

In ACM, there are some concepts, each of which includes a number of sub-terms without notations to explain the concepts. When establishing the mappings between DDC and ACM, it was found that some of these sub-terms could exactly map to the equivalent concepts. In this case, it was decided that if a concept (C1) in ACM includes a number of sub-terms, and one of these sub-terms is equivalent to a DDC concept (Ca), the exact mapping relationship can be used to map DDC concept (Ca) to the ACM concept (C1). Based on this principle, Figure 6.3 shows an example of a mapping established between DDC and ACM.



The terms under the concept "H.4.1 office automation" are not concepts in ACM
These terms have no notations.
Logic: If a concept (C-1) in ACM includes a number of terms, and one of these terms is equivalent to a DDC concept (C-2), then we exactly map C-1 to C-2.

Figure 6.3: An example of a mapping established between DDC and ACM.

Based on this principle, it was found that most ACM concepts could exactly match to the equivalent DDC concepts. The mappings between ACM and DDC can be found in Appendix 3.2.

6.3 Technical architecture of the prototype

One purpose of prototyping is to test if the selected technologies could be feasibly used in the middleware system between terminologies. Hence, a computer-based prototype system, which applies most of the technologies identified in the theoretical framework, was developed in this research.

6.3.1 Data formats

In this research, DDC data was made available in SKOS-Core format. However, it was found that SKOS-Core format is not specifically designed to represent classification data. For this reason, the SKOS-Core vocabulary needed to be adapted to suit different types of controlled vocabularies. A notation label was derived from <skos:preLabel> to represent

the classification concept. In this way, a classification concept could be presented as follows:

```
<skos:Concept rdf:about="http://www-
staff.lboro.ac.uk/~lsls2/ddc.rdf/069.51">
<skos:notation>069.51</skos:notation>
<skos:inScheme rdf:resource="http://www-staff.lboro.ac.uk/~lsls2/ddc.rdf" />
<skos:altLabel xml:lang="en">Selection, acquisition, disposal</skos:altLabel>
<skos:broader rdf:resource="http://www-
staff.lboro.ac.uk/~lsls2/ddc.rdf/069.5" />
</skos:Concept>
```

The UKAT thesaurus is originally encoded in SKOS-Core format, and each concept has a URI. The ACM Classification Scheme is in a simple XML format. In the ACM dataset, there is no URI to identify each concept. Based on the description in Section 5.1.2, when mapping ACM to DDC, a URI was designed and assigned to each concept of ACM in the mapping dataset, see Table 6.1.

DDC and ACM mapping data in SKOS-Mapping	ACM XML data
<skos:Concept rdf:about="http://www-staff.lboro.ac.uk/~lsls2/ddc.rdf/006.35"> <skos:notation rdf:datatype="http://iaaa.cps.unizar.es#notation">006.35</skos:notation> <skos:inScheme rdf:resource="http://www-staff.lboro.ac.uk/~lsls2/ddc.rdf"/> <skos:prefLabel xml:lang="en">**Natural language processing**</skos:prefLabel> <skos:broader rdf:resource="http://www-staff.lboro.ac.uk/~lsls2/ddc.rdf/006.3"/> <map:exactMatch **rdf:resource="http://www.acm.org/class/1998/acmccs98-1.2.3.xml/I.2.7"** /> //assigned URI for concept "natural language processing"// </skos:Concept>	<node id="I.2.7" label="**Natural Language Processing**"> <isComposedBy> <node label="**Discourse**" /> <node label="**Language generation**" /> <node label="**Language models**" /> <node label="**Language parsing and understanding**" /> <node label="**Machine translation**" /> <node label="**Speech recognition and synthesis**" /> <node label="**Text analysis**" /> </isComposedBy> </node>

Table 6.1: The mappings between a DDC concept and an ACM concept

In this context, three sets of mappings between the taxonomy of information science, UKAT, and ACM through DDC have been established in three SKOS-Mapping data sets.

A RDF Class called "conceptbag" was defined in this research, and used to combine a number of individual concepts together. Figure 6.4 show the way to use a RDF bag, and establish the mapping between the RDF bag and a compound concept.

```
<rdf:RDF xmlns:rdf="&rdf;" xmlns:rdfs="&rdfs;" xmlns:dc="&dc;" xmlns:dct="&dct;"
xml:base="http://www-staff.lboro.ac.uk/~lsls2/map1">
<rdfs:Class rdf:ID="conceptbag">
            <rdfs:label>conceptbag</rdfs:label>
            <rdfs:subClassOf rdf:resource="&rdf;Bag"/>
            <rdfs:comment>This class is used to combine a number of individual
concepts together</rdfs:comment>
      </rdfs:Class>
<skos:Concept rdf:about="http://www-staff.lboro.ac.uk/~lsls2/ddc.rdf/004.03">
   <skos:notation
rdf:datatype="http://iaaa.cps.unizar.es#notation">004.03</skos:notation>
   <skos:inScheme rdf:resource="http://www-staff.lboro.ac.uk/~lsls2/ddc.rdf"/>
   <skos:prefLabel xml:lang="en">Computer science—dictionaries</skos:prefLabel>
   <skos:broader rdf:resource="http://www-staff.lboro.ac.uk/~lsls2/ddc.rdf/004"/>
<map:majorMatch>
     <map1:conceptbag>
       <rdf:li rdf:resource="http://www.ukat.org.uk/thesaurus/concept/1582"/>
       <rdf:li rdf:resource="http://www.ukat.org.uk/thesaurus/concept/6118" />
     </map1:conceptbag>
   </map:majorMatch>
</skos:Concept>
```

Figure 6.4: Using RDF bag to combine individual concepts together

A Jena RDF API was employed to process the three SKOS-Mapping data files at the same time. Jena is a Java API for semantic web applications, in which the SPARQL queries could be built to query against the RDF data, and some related functions in Jena could also be used to query the RDF data. The Jena API was applied with the Java applets in the environment of NetBean 5.5.1. A Java applet is a program written in the Java programming language that can be included in an HTML page.

In the theoretical framework, it was assumed that a middleware system should be able to cross-access a variety of terminology services using different access protocols and query languages. For example, many current terminology services are using the SRW/U model with CQL as the query language. It is possible for this prototype to adapt the SRW/U model for accessing the relevant terminology services, such as HILT, OCLC TS. In

addition, with the development of semantic web technologies, a number of new protocols and formats would be developed. A knowledge base was developed to store connectivity details of different terminology services. This knowledge base could transfer the users' queries into various structures of queries that different terminology services could understand, and convert the returned terminological records into a consistent format. In this prototype system, the developed knowledge base could transfer a user's query into a SPARQL query to search the UKAT data in SKOS-Core format, and also transfer a user's query into a XML-based query to search the ACM data in XML format.

A DDC-based cross-browsing interface has been developed, and was presented to the end-users. The end-users could select the concepts from DDC and get the mapped concepts from either ACM or UKAT. The specific steps are listed as follows:

1. When users select Concept A from browsing DDC, the Jena agent processed the two mapping datasets, and return the relevant URIs of the concepts mapped to Concept A of DDC from UKAT and ACM.
2. Based on the URIs returned from the mapping datasets, two other agents were employed to deal with the UKAT data and ACM, and get more detailed information (e.g. preferred term, non-preferred terms, broader terms, related terms, etc) about the mapped concepts;
3. The returned terminological information will be converted into a consistent format, and presented to the end-users.

Two APIs that are employed to process the ACM and UKAT are: a W3C DOM API that is employed to develop relevant functions for querying the ACM XML data, and an ARQ API that is employed to relevant functions for querying the UKAT SKOS data. All these agents were implemented in Java environment, and embedded into Java applets. The basic diagram of the prototype is shown in Figure 6.5.

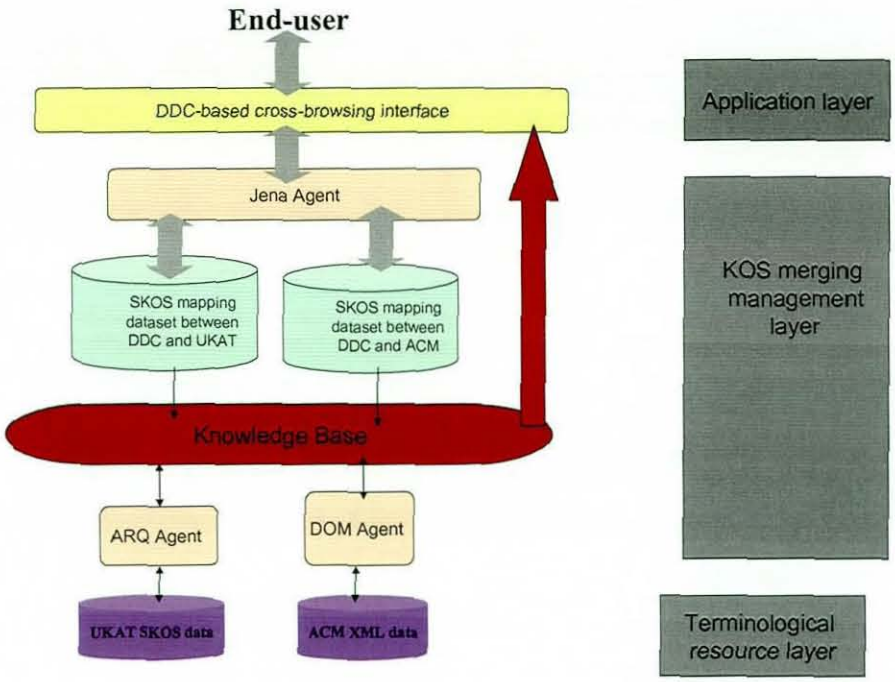


Figure 6.5: The basic architecture of the simplified prototype system

From the perspective of librarians, it could be decided that DDC is not good enough to be an appropriate scheme for subject cross-browsing in a library portal system. This may be because the inexperienced user may not be familiar with the terminology provided by DDC. In this case, the information science taxonomy was mapped to DDC. The mapping data was made available in SKOS-mapping format. Above the mapping data, a programmatic interface was developed to provide subject browsing based on the use of local vocabulary. The users began by interacting with the taxonomy, gain the mapped DDC concepts as mediators, and finally accessed the mapped concepts from other vocabularies through DDC. However, it was found that some specific concepts in this information science taxonomy cannot map to the relevant DDC. There are three main reasons for this. First, DDC used in this research was not complete. Second, concepts in the information science taxonomy are newer than DDC concepts. Third, in some parts of this taxonomy, the granularity is more detailed than DDC's relevant part. Fourth, because of the pre-coordinated nature of DDC, it is difficult to map a KOS to another pre-coordinated vocabulary. Thus, it was decided to use two levels of this taxonomy to map to DDC. It was found that about 90% of concepts in two levels of this taxonomy can map to the relevant DDC concepts.

6.3.2 Queries

In most subject browsing services, the users are not required to input their textual information to conduct their search. Instead, the users just need to identify some pre-arranged subject term labels, and click some of the labels to find relevant information. Thus, it is suitable for this prototype to use SPARQL query language to manipulate the mapping data in SKOS-Mapping format. A URI was given to each of DDC concepts. When a user clicks a DDC concept label, the SPARQL query would be generated to search the mapping dataset. Figure 6.6 shows a SPARQL query to search a DDC concept "004.678" with exactly mapped concepts from other KOS.

```
"PREFIX skos: <http://www.w3.org/2004/02/skos/core#>"+
    "SELECT  ?notation ?caption ?mappedconcept
 WHERE{{ "http://www-staff.lboro.ac.uk/~lsls2/ddc.rdf/004.678"
skos:notation ?notation}
{ "http://www-staff.lboro.ac.uk/~lsls2/ddc.rdf/004.678" skos:altLabel ?caption}
{"http://www-staff.lboro.ac.uk/~lsls2/ddc.rdf/004.678"
map:exactMatch ?mappedconcept}}"
```

Figure 6.6: A SPARQL query of a DDC concept

Based on using SPARQL to query DDC-mapping data, a number of mapped concepts from other vocabularies (e.g. UKAT, ACM) could be returned. These mapped concepts might be located in other terminology services in different formats. Thus, it is important to use other query languages to query against the dataset of other KOS.

In this prototype, because the UKAT data was in SKOS format, the SPARQL query language could still be used to further query the UKAT data for more terminological information. However, the ACM data is in a XML format. It is important to employ other types of queries to search the ACM dataset. In this case, W3C DOM API was employed to query the ACM dataset, and get the results from the ACM dataset.

6.4 Prototype demonstration

In the prototype, two user scenarios with relevant interfaces were designed. In the first user scenario, a list of DDC concepts was presented to the users. Each of these concepts

includes a DDC notation and a relevant caption. The users could browse DDC concepts through DDC hierarchy, identify the relevant concept, and click a "search" button to gain the mapped concepts from other vocabularies. Two checkboxes were also presented, which were used to let the users identify which mapping dataset is required. In this particular prototype, "mapping data between DDC and the UKAT" and "mapping data between DDC and ACM" are provided. Figure 6.7 is presented to show the basic elements in this interface.

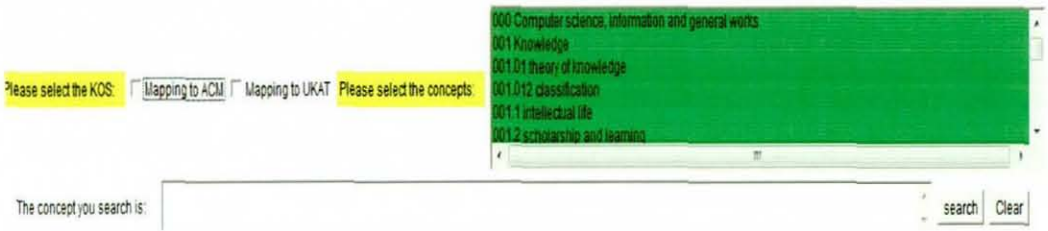


Figure 6.7: User-based subject cross-browsing interface within the prototype system

When the users selected an appropriate DDC concept to search relevant mapped concepts in another vocabulary, the mapped concept would be returned to the users. Then, the users could identify if the returned concept were suitable for their further search. Because concepts from different types of vocabularies should be presented in different ways, it was important to consider using various methods to represent the concepts from different vocabularies. In this case, the preferred term and non preferred terms were used to present a UKAT concept, and a notation with the relevant caption was used for an ACM concept. In addition, in ACM, there are some concepts including a number of sub-terms to explain the concepts, but these sub-terms have no notation, and are not concepts in ACM. In this case, these sub-terms with the concept notation and caption were presented to the user so that the users could get a better understanding of what the returned concept means. Figure 6.8 shows the ways to present the mapped concepts from both the ACM and the UKAT to the end-users, and the selected DDC concept is "004.6 interfacing and communication".

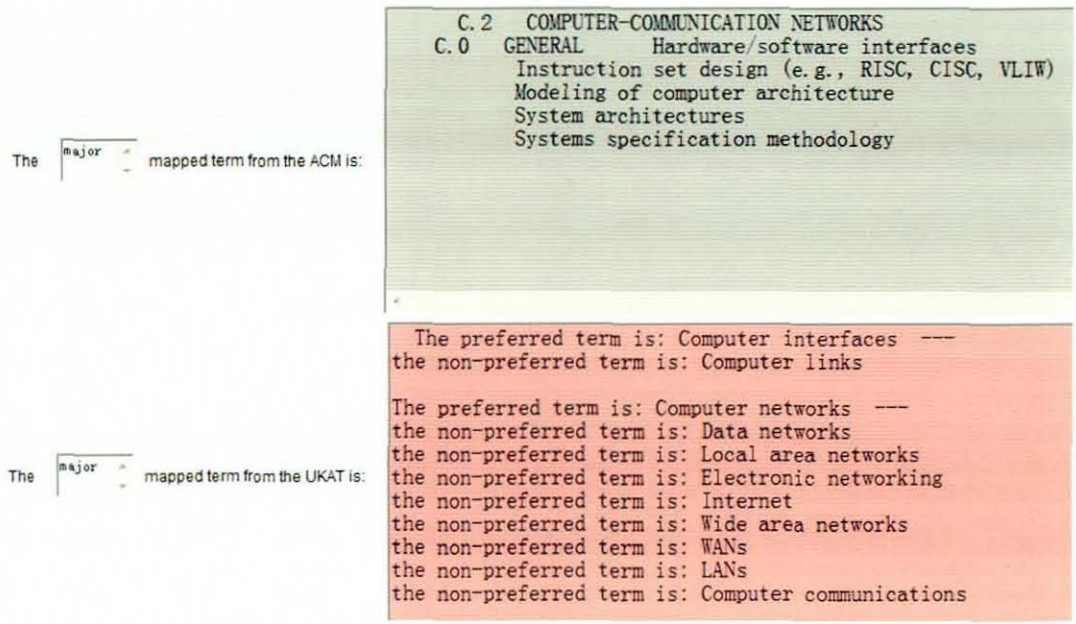


Figure 6.8: The ways to present the returned mapped concepts from this subject cross-browsing interface

In another interface, an information taxonomy was presented to the users for subject cross-browsing. When the users clicked the relevant terms in this taxonomy, the mapped DDC concepts would be returned, and the mapped DDC concept would become the query to search the mapped concepts from the ACM and UKAT. Figure 6.9 is an example of the search. In this example, the users would begin by clicking the term "2.1 thesauri, authority list" in the local taxonomy, and get the mapped DDC concept "025.49 controlled vocabulary". Subsequently, the mapped DDC concept would become the query to search against the mapping dataset between DDC and UKAT, and mapping dataset between DDC and ACM. The relevant UKAT concept "controlled languages", which has been mapped from DDC, would be returned. Finally, the UKAT concept "controlled languages" would be used to query against the databases that have been indexed according to the UKAT for getting the item-level metadata records.

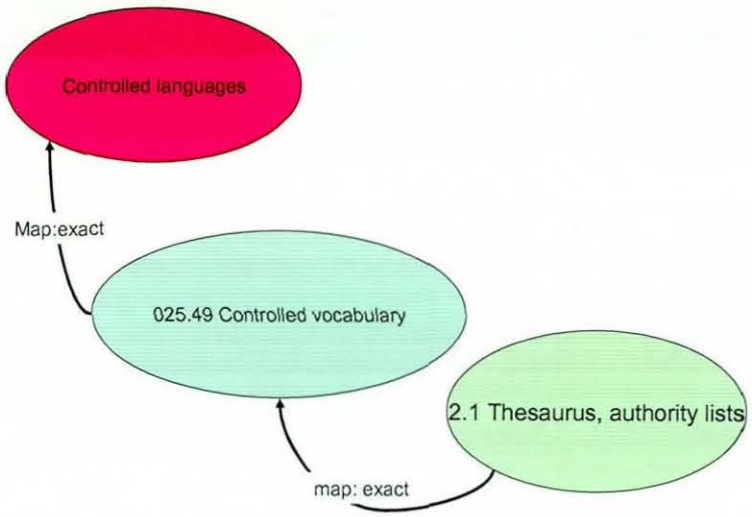


Figure 6.9: An example of a mapping between a local taxonomy and UKAT through DDC

6.5 Conclusion

Based on the development of this prototype, it was found that most of selected techniques discussed in Chapter 5 have been effective to retrieve the terminological information from different terminology resources. A number of points are concluded as follows:

1. It is feasible to establish mappings between UKAT and DDC, and between the ACM and DDC, based on the mapping strategy defined in Section 5.1, although the accuracy of the mappings still needs to be evaluated further. However, the indirection problem caused by the use of DDC-based switch language is not solved;
2. It is true that SPARQL is not focused on searching text-based data. However, it is not necessary for users to input the text-based query to search the information in a subject cross-browsing service, and the users only need to click a number of pre-arranged term labels, and then find the relevant information. Thus, the use of SPARQL is appropriate to facilitate the subject cross-browsing services;

3. This prototype developed a number of data transmission and conversion programmes. It is technically reasonable to translate a user's query into different forms of queries that different KOS or terminology services could recognise, gain the terminological information from different terminology resources, and finally convert the returned results into a consistent format. It would be useful to extend this ideas to incorporate more terminology services and other terminology resources into this prototype system;
4. In this prototype, when users interact with the local taxonomy, there is a long process that requires a fairly high degree of interaction from users to progress from the stage where they browse the hierarchy to getting the relevant mapped terms from various KOS through DDC spine. It is desirable to develop some approaches to improve the interactivity between the users and the systems.
5. A concept in any type of KOS could be presented in a variety of ways. It is important to further explore the methods to friendly present a concept to the users, and make the users fully disambiguate the returned conceptual term.

Chapter Seven: Evaluation Data Analysis

7.1 Evaluation method

As discussed in Section 3.4.4, an expert-based evaluation method was selected to assess the framework, and six experts were asked to walk through the software-based prototype system, and provide feedback to improve the system. The experts in the evaluation are not those who were previously interviewed. Before assessing a terminology mapping-based framework, it is important to identify a reasonable analytical starting point. As discussed in Section 3.4.4, six heuristics were highlighted, which were:

1. Semantic extensibility;
2. Technical adaptability;
3. User interactivity;
4. Cultural feasibility;
5. Technical viability;
6. Quality of established mappings.

Through analysing these heuristics, it is obvious that the evaluation needed to consider accurately both the technical and semantic approaches used for the development of the middleware system and whether these approaches are applicable in the long term. The evaluation was impartial and was conducted in a manner encouraging the expert evaluators to be open and frank.

Based on these heuristics, a range of specific objectives of this evaluation can be summarised, and can be linked to relevant heuristics. These objectives include:

1. **Semantic extensibility and quality of established mapping**: To examine the effectiveness of the methods to establish the mappings between DDC and two selected vocabularies;

2. **User Interactivity**: To determine whether the selected methods to present the ACM concepts and UKAT concepts are suitable for helping users disambiguate the mapped concepts;
3. **User Interactivity**: To determine whether the use of a local taxonomy as a subject cross-browsing interface is more suitable than the use of DDC as a browsing interface;
4. ***Semantic extensibility and quality of established mapping***: To explore whether there are appropriate methods to solve the indirection problems arising from the use of the DDC spine;
5. **Cultural feasibility**: To test the viability of the cooperative work between this mapping middleware and other services;
6. **Technical viability and technical adaptability**: To test the technical feasibility of the information architecture that is used for the development of this middleware system;
7. **User interactivity**: To examine user interactivity with the subject cross-browsing interface, and test if users could use the subject cross-browsing interface to find metadata records.

Based on these objectives, a number of codes were developed and employed to analyse the data collected through the evaluation. These codes are listed hierarchically in Table 7.1.

Table 7.1: Code used in the analysis of the evaluation data		
Codes	**Sub-codes**	**Explanation**
Mapping strategy	Structural model of mapping	To test if DDC is an appropriate spine to exchange terminological information between KOS
	Mapping relationships and logics	To examine if the 5 identified mapping relationships are appropriate in various cases
	Mapping collaboration	This is related to the question of who should create the mappings for this middleware.
	Depth of mappings	Different mappings may facilitate different subject services in different perspectives.
	Correction of the mappings: • Mapping between DDC and UKAT • Mapping between DDC and ACM • Mapping between DDC and local taxonomy	This is related to the question whether the established mappings are correct.
Methods to present different concepts	Ways to present classification concepts	This refers to what terminological information for various concepts should be presented to the users so that the users could understand the context of a given concept.
	Ways to present thesaurus concepts	
	Ways to present concepts from other kinds of KOS	
The use of local taxonomy for subject cross-browsing		To test if the local taxonomy is more friendly than DDC. To explore some potential methods to solve the indirection problems driven by the use of DDC spine.
Technical architecture	Standards • SKOS; • SPARQL; • SKOS API; • Etc	To test if the used standards is appropriate
	Knowledge base for accessing various terminology services	To test if the knowledge base is extensible enough to access most of different terminology services
Automated approach	Automatic mapping	To test whether some automated approaches could further improve the framework
	Query expansion	
Interaction design issues		To test whether the users could successfully interact with the prototype system
Other issues		

In order to help the evaluators become familiar with the browsing functions of the prototype, a number of use demonstrations and an explanatory document of the theoretical and technical approach/interface were produced (Appendix 4 and 5). The use demonstrations reflected the current capabilities of this middleware system for providing mapping data between different terminologies. After each demonstration, the evaluators were asked to identify relevant problems, and express their comments.

After all the demonstrations, a number of interview questions were devised to collect qualitative data from experts (See Appendix 5). The interviews were recorded, and transcribed to text. All the interview questions were mapped to the heuristics described in Section 3.4.4. In parallel, a document called "Introduction document" was designed to explain the basic design principles of the theoretical framework and relevant prototype system, describe the defined demonstrations, and present the interview questions. This document was sent to each of evaluators before they interacted with the prototype, which helped the evaluators to understand the theories used to develop the framework. This document is given in Appendix 4. The findings are reported as follows:

7.2 Findings of the evaluation

7.2.1 DDC as a switch language

Because DDC is a pre-coordinated classification scheme, and does not have sufficient granularity in some subject areas (e.g., information science, medical science, etc), all evaluators agreed with the expert interviewees that generating mappings between DDC and other vocabularies can result in the loss of precision and specificity. In some mapping projects, the mapping staff may lack knowledge of DDC, and they may misunderstand the meanings of some DDC captions (Evaluators 1, 4, and 5). For example, DDC concept "005.10288 maintenance and repair" is in the field of computer science, and some mapping staff may wrongly map this DDC concept to the UKAT term "maintenance" in the field of construction engineering. It is therefore essential to provide an appropriate DDC training programme to the mapping staff.

The copyright of DDC was emphasised by Evaluator 1. If an organisation maps their KOS to DDC, OCLC, as DDC owner, would allow this organisation to do that, but they would not allow them to use the mapping in their service. They need to submit the established mappings to OCLC, and OCLC will sell the mappings back to them.

On the other hand, all evaluators agreed with the expert interviewees that using DDC still makes sense in the medium term. DDC is a well-known classification, and is widely used by a huge number of organisations. DDC covers most subject areas, and is encoded in the MARC21 XML format for the purpose of exchange.

All the evaluators predicted that well-developed post-coordinated KOS would be more suitable as switch languages than DDC. UDC could be an alternative to DDC as a switch language because UDC has greater notational synthesis capability than DDC (All evaluators). UDC is more detailed, and could be adapted to incorporate more subject concepts by freely synthesising concepts. Two important issues with UDC were pointed out by Evaluator 5. The first one is that UDC is less well-known than DDC, because UDC is mainly used in European libraries. The Second issue is that the financial support for UDC is less than for DDC, It could be said that UDC, to some extent, is suffering from poor management problems. Some other faceted vocabularies could potentially be adopted as a switch language. One of these vocabularies is BLISS, which has great notational synthesis ability and great granularity (Evaluators 2 and 5). However, this vocabulary is still in the developmental stage. In some subject-specific domains, a number of subject-specific schemes could be employed as switch languages for exchange the terminological information between different KOS in the same domain (Evaluator 4). For example, in most UK government-based projects, IPSV (Integrated Public Sector Vocabulary) was applied as a switch language, to which different other government-related vocabularies could be mapped.

Besides using a switch language to improve the interoperability between different KOS, the construction of a meta-thesaurus to standardise the existing concepts from different terminologies was emphasised by Evaluator 2.

7.2.2 Mapping relationships and logics

In the five selected mapping relationships, the differences between major match and minor match may not be clear for most users (Evaluators 1 and 5). In this study, major match is used to represent the mapping between a compound concept and a combination of several concepts, and minor match is used to represent the mapping between two related terms. However, in the SKOS-Mapping vocabulary, major and minor matches are used to represent co-occurrence mappings. The difference between these two relationships may cause users and mapping staff to be confused when using these two relationships.

To solve this problem, several suggestions were provided by evaluators. Firstly, it was suggested to present information about the definitions of these mapping relationships to users, before they interact with the browsing service (Evaluators1, 3, and 5). In this context, the users could understand the real meanings of the mapping relationships. Also, the term "related match" should replace the term "minor match" in this framework to avoid confusion (Evaluators 1, and 5). In addition, because different mapping projects may be based on different mapping methods, such as intellectual mapping, automatic mapping, and co-occurrence mapping, as mentioned in Section 4.2.4, it is important to introduce new metadata elements to characterise the established mappings through these methods (Evaluators 1 and 5). For example, when handling some automatic mapping data, it is possible that each mapping relation can be tagged with a relevance rating (high, medium, and low) to make the quality of the mapping relations more explicit (Evaluators 1, 2, 3, and 6).

Secondly, it is desirable to use different mapping logic to establish different mapping sets for the same concepts in different KOS. For example, when mapping "025.0654 Chemistry—information system" to relevant concepts in UKAT, in parallel to using a bag to make a combination of relevant concepts, it is also possible to use the term "chemistry" in UKAT as a broad term of "025.0654 Chemistry—information system" in DDC, and use the term "information system" as a broad term of "025.0654 Chemistry—information system". All these mapping sets could be presented to the users so that the

users could select the most appropriate mapping set for their subject requirements (Evaluators 1, 5, and 6). Figure 7.1 shows a variety of mapping sets between the concept "025.0654 Chemistry—information system" in DDC and relevant concepts in UKAT. In addition, it is important to know if each local search engine used by different online databases has the algorithms to deal with Boolean operators (Evaluator 5). If a search engine does not support Boolean search, it is meaningless to let users add Boolean operators to further conduct their searches. Today, a number of search engines used by different collections, such as, Google, Ask, etc., do not have the ability to handle the Boolean operators.

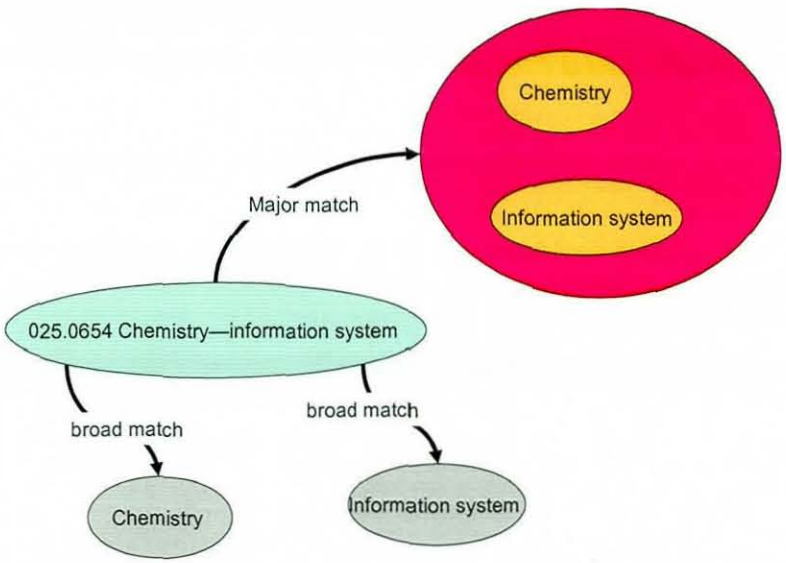


Figure 7.1: A variety of mappings established between a DDC concept and two UKAT concepts

In order to present the mapping data in more user-friendly ways, it was suggested that relevant visualisation technologies could be helpful to show the explicit mapping relationships and the mapped terms to the users, and improve the interactivity between the users and machine (Evaluator 2).

Thirdly, because a number of users may not be interested in the specific mapping relationships used to establish the mappings, and they may not be concerned with the

vocabulary that the collections were using, it was suggested that it would be useful to allow users to click the mapped terms without knowing much about specific mapping relationships and then use guidance to present all possible combinations of the mapped terms to the users (Evaluators 1, 2, 4, and 6). Google-style "Do you mean..." plus all possible Boolean combinations of mapped terms is very useful for this purpose. For example, it is possible to use "Do you mean Chemistry, information system, Chemistry and information system, Chemistry or information system, Chemistry not information system, information system not chemistry?" to present all possible combinations of mapped concepts in UKAT to the user.

7.2.3 Mapping collaboration

Creating and maintaining the mapping is possibly the most important barrier to the development of this framework (Evaluators 1, 3 and 5). The importance of re-using existing mappings from other terminology services has been highlighted (Evaluators 1, 3, and 5). Because this framework is proposed to encourage different participants to contribute the mapping data, it is helpful to develop some API functions within the middleware system to allow other parties to submit their mapping data (Evaluators 1 and 3). As the mappings might be from different terminology sources, it is important to be able to develop a metadata application profile to characterise various mapping sets – provenance (source), method (intellectual, co-occurrence, other automatic, etc), perhaps with a quality indicator (Evaluator 3).

Also, before establishing the mappings between two KOS, it is important for the mapping staff to identify the mapping direction. In some cases, the mappings established from Vocabulary A to Vocabulary B might be greatly different from the mappings created from Vocabulary B to Vocabulary A. For example, the concept 'Computer' in Vocabulary A might be matched to the concept 'Information System' in Vocabulary B, but the same concept 'Information System' in Vocabulary B might be matched to another

concept 'Data base' in Vocabulary A. In many mapping projects, such as, KoMoHe Project[1], etc., two vocabularies were mapped bilaterally.

In order to further develop the mappings in different terminology services, a suggestion was made by Evaluator 5 that one central team with sufficient expertise should be formed and should have the responsibility for maintaining the mappings from different terminology, and improving the consistency of the mappings from different terminology resources.

Evaluators 4 and 6 emphasised the importance of developing programmatic interfaces to cross-search different terminology resources. These terminology resources might include terminology services, the KOS in various digital formats located on the Web, the local KOS in local databases, etc. In this context, a query input by a user could cross-search a number of terminology resources, return a number of relevant concepts, and then present them to the user. It is important to offer functions to limit users to search particular vocabularies across these terminology resources, and develop query expansion algorithms to expand the returned concepts (Evaluator 4). In this context, the terms from different KOS could be mapped together on the fly through the query.

A number of automatic mapping techniques were introduced by Evaluator 6. Automatic mapping solutions can be derived from various techniques, such as lexical techniques, ontology mapping techniques (that is based on reasoning over different ontological instances), statistical techniques, etc. On the other hand, Evaluator 6 emphasised that most established automatic mapping tools performed insufficiently in real cases, and most terminology mapping projects still depends on intellectual mapping techniques to create their mapping data.

[1] A German project supervised a terminology mapping effort, in which 'cross-concordances' between major controlled vocabularies were organized, created and managed.

7.2.4 Mapping depth for subject cross-browsing

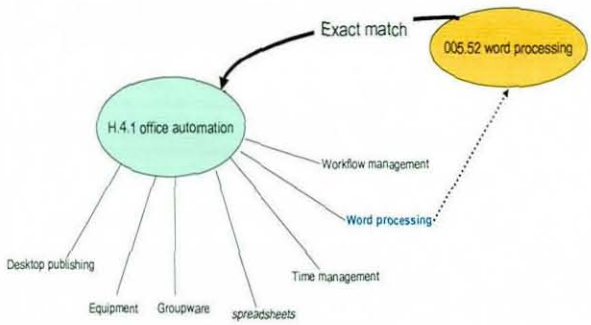
There was widespread support for creating deep mapping (all evaluators). It was difficult to see any benefit, beyond the superficial, in limiting activity to shallow mapping. The risk is that users may be frustrated by access to shallow mapping and make no further use of the service (all evaluators). For example, if a user uses the term "weeding" as a search term, and gets a very general mapped concept, such as, "computer science", he/she would have no confidence in the service. Some query expansion algorithms, however, could be used to expand the mapped concepts based on the established shallow mappings, and improve recall, but even by using expansion it is still difficult to improve precision. Also, in this framework, the returned terms from subject cross-browsing services could be used as a query through the meta-search engine to cross-search different databases. In this context, it is not very useful to automate the shallow mappings to make the shallow mapped terms cross-search different databases (Evaluator 4).

Evaluators 1 and 2 pointed out some use scenarios of using shallow mappings. In these scenarios, it is possible for users to separately browse the structures of different controlled vocabularies. DDC might be a good starting point to combine various KOS together at the top levels of DDC. In this manner, several vocabularies could be mapped to the top levels of DDC. A user could then begin to interact with the top levels of DDC concepts to find the other relevant vocabularies, and then the user could jump into a specific structure of another vocabulary, and be navigated through this structure to find relevant information.

In addition, because mapping is not a simple job, and there are a number of uncertain issues about how to create the mappings, it is important to take further initiatives in mapping practices, evaluate the established mappings and incorporate the lessons into the future mapping projects (Evaluators 5 and 6). For this reason, the established mappings could form an evaluation basis to measure the costs and benefits before expanding the mapping work.

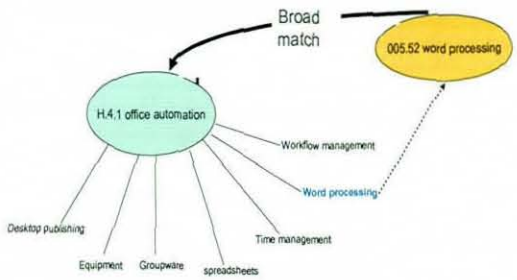
7.2.5 Correction to the established mappings

One main mapping error was identified. It was noted that in the ACM computing classification, there are a number of terms without notations, and these terms cannot be treated as elements for mapping (Evaluators 1, 2, 3, 5, and 6). For example, in ACM, the term "word processing" was used to explain the concept "H.4.1 office automation", but it cannot be treated as a mapping element. For this reason, it is important to use the "broad match" relationship instead of the "exact match" relationship. See Figure 7.2 and Figure 7.3.



The terms under the concept "H.4.1 office automation" are not concept in ACM These terms have no notations.
Logics: If a concept (C-1) in ACM include a number of terms, and one of these terms is equivalent to a DDC concept (C-2), then we exactly map C-1 to the C-2.

Figure 7.2: Old mapping



The term "word processing" under the concept "H.4.1 office automation" in ACM is not a concept. This term has no notation.

Logics: If a concept (C-1) in ACM includes a number of terms, these terms have no notation, and one of them is equivalent to a DDC concept (C-a), then the broad match relationship is used to map C-a to C-1.

Figure 7.3: Improved mapping

7.2.6 The use of bag to combine concepts

It was widely-agreed that using the bag to combine a number of concepts together is a good idea (all evaluators). Based on their subject needs, users can select some terms from a bag, or use appropriate Boolean operators to combine some of selected terms in the bag to further conduct their searches (Evaluator 2). However, when a number of mapped terms are returned and listed, many users might not realise that they could use Boolean operators to further combine these mapped terms. It is important to integrate some

checkboxes into the subject cross-browsing user interface. The checkboxes could let users select appropriate Boolean operators to combine the mapped terms and further conduct their subject meta-search (Evaluator 3).

7.2.7 Methods to present various concepts

In the developed prototype system, the preferred term was used to present a UKAT concept, and the notation with the relevant caption was used for an ACM concept. In ACM, there are some concepts including a number of sub-terms to explain the concepts, but these sub-terms have no notation, and are not concepts in ACM. These sub-terms with the concept notation and caption are presented to the user so that users can get a better understanding of what the returned concept means.

However, it was found that the terminological information used to present the mapped concepts in this prototype system was not sufficient for the purpose of term disambiguation (Evaluators 1, 2, 3, and 4). When users get the mapped subject terms through subject cross-browsing, they may not be sure if the mapped terms can represent their subject needs. In this case, it would be useful to show users a neighbourhood of the mapped concepts considered semantically close (Evaluators 3, and 4). This could help users get a better understanding of mapped terms, and give users more options to expand their search. With this consideration in mind, Evaluators 1 and 3 suggested showing a thesaurus preferred term with its non-preferred terms, broader term, narrow terms, scope note, and related terms. Also, in some cases, the sibling terms of a preferred term might be useful for expanding the user's query (Evaluator 3).

In classification schemes, some captions are not the most explicit terms to represent the concepts, and it is not reasonable to use captions as a search term to conduct understandable searches (Evaluator 6). Thus, it is important to present the whole hierarchy where a given concept is located to the users (Evaluators 1 and 6). For example, when presenting the concept "B.1.5 Microcode applications", it is desirable to show the hierarchy "B.hardware/B.1 controlled structures and microprogramming/B.1.5 Microcode

applications" with the concept "B.1.5 Microcode applications" to the end-users. In other words, all the semantically close concepts can be presented to provide the detailed information. This would help explain a mapped concept so that the users can understand the mappings better. However, if there are too many terms returned, the users may feel confused. In this context, it is important to use some technologies to highlight the main mapped term, and give the users the ability to distinguish the difference between the mapped term returned and the semantically close concepts. Figure 7.4 shows the method used to highlight the mapped term, and make a distinction. In this case, the mapped concept "'B.1.5 Microcode application" from the ACM classification was highlighted.

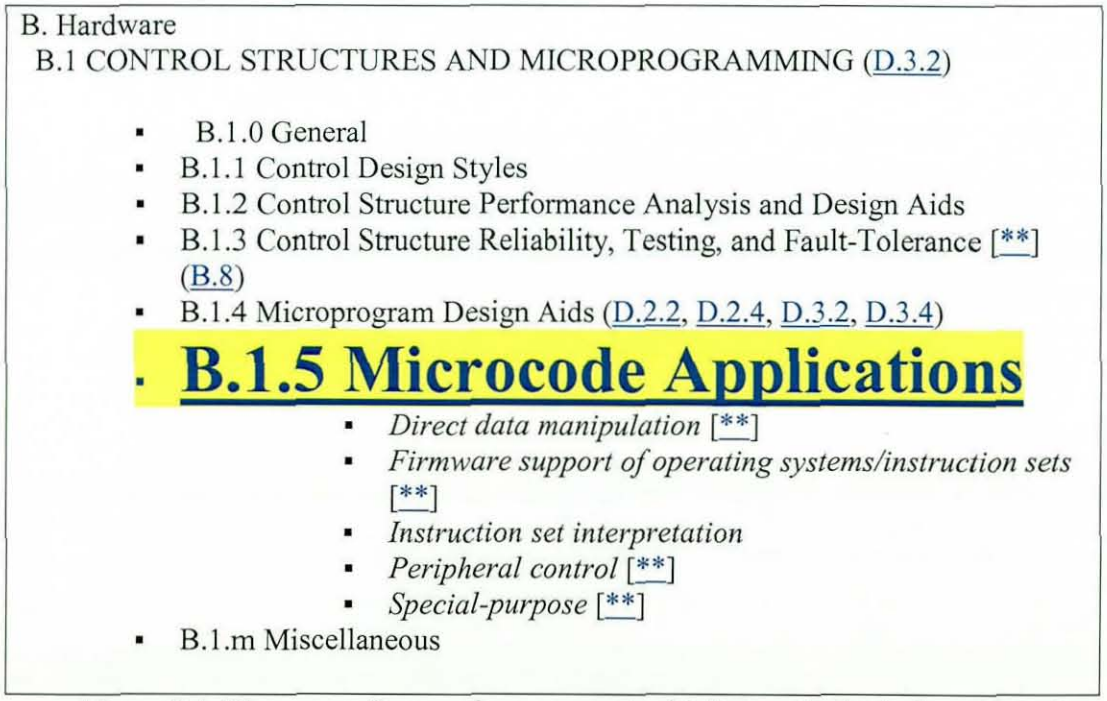


Figure 7.4: The ways of presenting a concept with its terminological contextual information

It was noted that most end-users are just concerned with the metadata records finally returned rather than the real mapped terms used to cross-search different databases (Evaluators 2 and 5). When the returned metadata records are not relevant, users might want some more subject terms to further refine their search. In this case, it would be helpful that this middleware system could simply present a number of mapped terms

from various KOS and terms considered semantically close to these returned mapped terms, and let the users select some of them without knowing where these terms come from. Using a "Do you mean...?" sentence plus all mapped terms from different vocabularies might be an appropriate way to present these terms to the users.

7.2.8 The use of local taxonomies for subject cross-browsing

In this framework, it was agreed by all evaluators that a local taxonomy is more focused and specific to reflect subject areas that the end-users are interested in. Most users like to start their subject browsing through their local taxonomies. However, using a local taxonomy through DDC would lead to the loss of precision, and make the search more general or narrower than what the users initially require (all evaluators).

In order to solve the indirection problem, a number of suggested solutions have been given, and listed as follows:

1. It is possible to use the UDC or other faceted general vocabulary as a switch language to replace DDC (Evaluator 3). These faceted vocabularies should have more flexible notation synthesis capability and more detailed granularity. Thus, it would be easy to combine different concepts into a new compound concept to support complex mapping work.
2. Instead of using the switch language, it is possible to establish many-to-many mappings between different KOS (Evaluator 5). However, this approach would be very expensive;
3. Because different mapping relationships were used to establish the mapping, it is important to make all these relationships with the mapped terms transparent to users (Evaluators 3 and 4). Based on this approach, the users can know the mapping relation from the local taxonomy to DDC and from DDC to other vocabularies, and make a judgment on whether the mapped terms are appropriate to further conduct their search.
4. It is possible that the mapped terms returned are not always totally appropriate. In some cases, though the mapped terms may be relevant, it might be important to give users some other (optional) chances to further refine the search term by local

browsing or by expanding the mapped concepts within the vocabulary, and then reformulating the subject search before searching within collections. For this reason, it would be helpful to develop some query expansion algorithms to return more terms semantically close to the mapped concepts, and show these expanded terms to the users. (Evaluators 1 and 6) The users might then be able to find some more appropriate terms to further refine their search. Evaluator 6 proposed a specific algorithm, which makes the system automatically decide the direction of expanding the mapped concepts in a KOS depending on analysing different mapping relationships used to establish the mappings between the local taxonomy and DDC, and mappings between DDC and other vocabularies. This algorithm is presented in Table 7.2.

Table 7.2: An algorithm to make the system automatically decide the direction of expanding the mapped concepts in a KOS

Concept in the local vocabulary	**Mapping relationship to the switch language**	**Concept in the switch language**	**Mapping relationship to the mapped term**	**The direction of expanding the mapped concept**
Cat	Broader	Pet	Exact: pet	To be expanded to present the narrow concepts of the mapped concept (pet).
Pet	Narrower	Dog Cat Pig	Exact:dog Exact: cat Exact: pig	To be expanded to present the broader concepts of the three mapped terms (dog, cat, pig).
Catalogue	Related	Cataloguer	Exact: cataloguer	To be expanded to present the related term of the mapped term (cataloguer).
Catalogue	related	Cataloguer	related: cataloguing	To be expanded to present the related term of the mapped term (cataloguing)

5. When the mappings between the local taxonomy and DDC and mappings between DDC and other vocabularies are established, all the mapping sets would form a basis to further develop the direct mappings between the local taxonomy and

other vocabularies without using DDC as a spine (Evaluator 5). In this context, the query expansion would be used to expand more semantically close terms, and find more appropriate terms to facilitate the mapping work.

7.2.9 Technical architecture

Knowledge base

It is widely-accepted that developing different applications to query different KOS in different terminology services is a scalable approach (Evaluators 2, 3, and 4). The knowledge base, which is used to translate a user's query into different forms of queries that different KOS or terminology services could recognise, is a very powerful way to technically incorporate a lot of terminology services (Evaluator 4).

However, with the increase of the number of different terminology resources, the information in the knowledge base would become larger and larger, and the knowledge base would become difficult to maintain (Evaluator 4). Because both the middleware system and the meta-search engine use different protocols and APIs to process the data from distributed resources, the response time might become slow (Evaluator 3). When new terminology services using new protocols and formats are added to this framework, it would take a lot of effort to update the knowledge base. With this in mind, Evaluators 5 and 6 implied that a number of terminology services should be developed based on the centralised model, which could consolidate a long term commitment to create, collect and maintain the mapping, and guarantee the consistency and quality of mapping work, and this distributed middleware system should be devised to comprehensively access a small number of important centralised terminology services, such as, HILT, OCLC TS, STAR, etc. When only a few terminology services are proposed to be integrated within this middleware, it would be easier to develop and maintain the knowledge base of this middleware.

Encoding formats

Because a variety of KOS, such as classifications, thesauri, taxonomies, subject headings, authority lists, ontologies, etc., differ greatly in their structures, different encoding formats should be applied to represent different KOS (Evaluators 2 and 4). These formats

might include Zthes XML Schema, SKOS, MARC21, DD8723, etc. Each format was designed to represent one particular type of vocabularies. In one case where there is a need to encode a particular vocabulary, it is desirable to use the most appropriate encoding format to encode this vocabulary.

Because different terminology services, which are proposed to be integrated within this middleware, use different encoding formats to wrap their terminology information, Evaluator 2 pointed out that this middleware system should be designed to be able to process most of these data formats, and convert terminology data from different terminology services into different other formats for interoperability. The converted data would be returned to the client side of this middleware, and the client would process the data, and present the data to the users.

7.2.10 Automated approach

It is important to consider the methods to develop additional automated approaches to reducing the current human-intensive effort. A number of automated tools have been developed to assist creating the mappings (Evaluators 1, 2, 4, and 6). However, some of these tools are very simple, which might lead to bad mappings. Thus, it is important to achieve a good balance between using human intelligence and machine intelligence (Evaluator 2). In practice, automated mapping tools are usually used to help find the important mappings that are not discovered by the human mapping workers. The human mapping workers should also check and correct the bad mappings created by the machine.

7.2.11 User interaction issues

A number of suggestions to improve the interactivity between the users and the middleware system were collected, and are presented as follows:

1. Most users are not concerned with which vocabulary a mapped concept comes from, and it would be preferable to let the users select mapped terms without knowing information about the vocabulary (Evaluators 1 and 2). Using "Do you mean..." plus the mapped concepts from different vocabularies would be useful. This would make this middleware system more similar to the typical google-like web services;

2. Because it is a "long journey" from the user's browsing of the local taxonomy to get final metadata records from distributed information resources, a user needs to get involved in a number of detailed interactions across different stages (Evaluators 3 and 4). It was suggested to use visualisation technologies to improve the interactivity between the users and machine (Evaluators 2 and 3). Through the visualised data, users could more fully understand the relations between different terms from different KOS.

3. When cross-searching a number of databases in a library portal, it is important to allow users to remove irrelevant databases before they conduct the search. In this case, it is hoped that all databases could be indexed according to the local taxonomy, and all KOS could be indexed based on the local taxonomy as well. Through the local taxonomy, the users could not only find the relevant item-level metadata records, but also find the relevant KOS and databases to further refine their searching. See Figure 7.5.

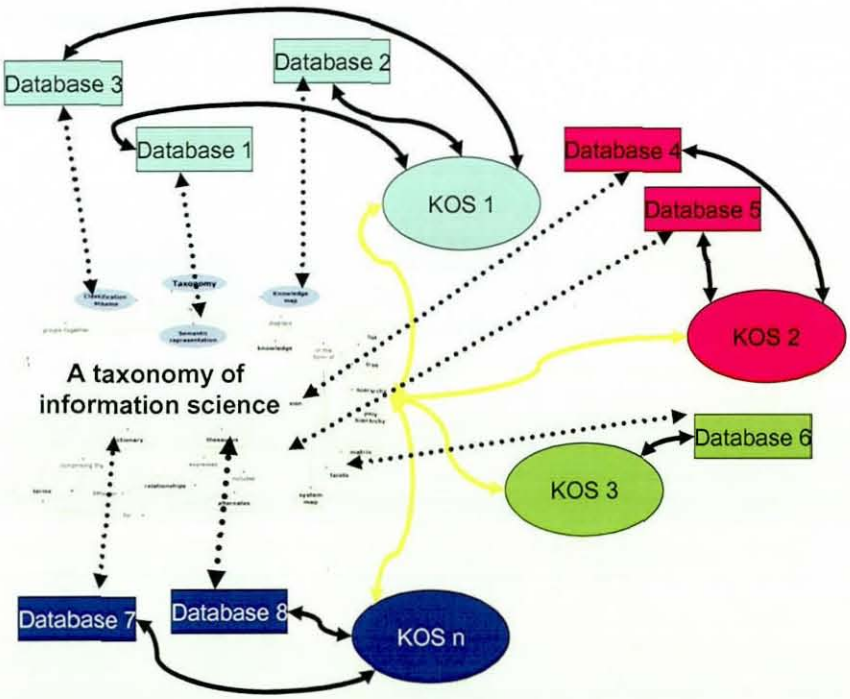


Figure 7.5: Various resources that are indexed by a local taxonomy

In this figure, the broken line means that the databases are indexed by the local taxonomy in the collection level. The curved lines are used to show that the metadata items in each of databases are indexed by a specific KOS. The yellow curved lines are used to demonstrate that all the KOS in this framework are indexed by the local taxonomy. Through the taxonomy, the users could be navigated to find relevant databases indexed according to this taxonomy, and relevant KOS for further subject browsing various databases. Because all the KOS are indexed according to this local taxonomy, it would also be useful to develop a KOS registry with a number of sophisticated metadata elements to describe the main characteristics of various KOS.

7.3 Conclusions

As described in Section 7.1, seven evaluation objectives were set up. This section will discuss how the evaluation findings match to the achievement of these objectives.

7.3.1 To examine the effectiveness of the methods to establish the mappings between DDC and two selected vocabularies

In the evaluation, the quality of the established mappings was reviewed by the experts. As mentioned in Section 7.2.5, one main mapping error for the mappings between ACM and DDC was identified, and the relevant solution to this problem was indicated. Because of the confusions in the minor mapping relationship, it was suggested by the evaluators to use related match relationship to replace minor match. In addition, it was suggested that some other mapping relationships need to be further explored to represent the automatic mappings, and co-occurrence mappings.

Using the bag to combine a number of concepts was accepted by all the evaluators as a good approach to establishing the mapping between a number of post-coordinated concepts and a compound concept. In parallel to using the bag to combine the single conceptual terms, a number of other methods to handle the mappings between a number of single conceptual terms and a compound concept were suggested in Section 7.2.2. It is agreed by a number of evaluators that within a mapping project, different mapping strategies/logic could be used together to present the mapped terms in different ways, and

let different users select the mappings tailored to their special subject needs. This could help the users understand the established mapping data in more sensible ways.

7.3.2 To determine whether the selected ways to present the ACM concepts and UKAT concepts are most suitable for helping users disambiguate the mapped concepts

Based on the evaluation findings, the methods of presenting various concepts in this framework need to be improved, and it was suggested that more contextual terminological information should be added to the mapped concepts returned from this middleware system. For example, the broader term, narrower terms, related terms, non-preferred terms, scope note, and even sibling terms of a mapped concept should be retrieved from this middleware, and presented to the end-users. Based on this kind of thesaurus-like display of concepts, the end-users could easily disambiguate the returned concept, and understand real meaning of this concept.

In many cases, users may be very familiar with a Google-style search. It is possible to present the mapped terms to the users without using any semantic relationships, and use "do you mean" plus all possible mapped terms to let the users select the most relevant subject terms.

7.3.3 To determine whether the use of a local taxonomy as a subject cross-browsing interface is more suitable than the use of DDC as a browsing interface

All evaluators agreed that a local taxonomy is a reflection of how end-users within a department consider their subject areas, and is more focused and specific. Another advantage of using a local taxonomy is that it is easy to integrate different information resources and the local collections into a single platform to improve the interoperability. On the other hand, some evaluators had a strong impression that using standard classification schemes for subject browsing might be suitable for some other users, such

as librarians, subject experts, etc. Thus, it is useful to give a variety of subject browsing structures to different users, and let them select the most relevant browsing structure for subject browsing.

7.3.4 To explore whether there are appropriate methods to solve the indirection problems arising from the use of the DDC spine

In this evaluation, all evaluators highlighted the indirection problems arising from the use of the DDC switch language. Based on the evaluation findings, the solutions to this problem focus on the use of query expansion algorithms to expand the mapped terms to return a number of terms considered semantically close. In this sense, the end-users could look into all these returned terms and find more relevant subject terms to further conduct their subject search. In addition, based on the established mappings between a local vocabulary and DDC, and the established mappings between DDC and other vocabularies, mapping staff could find new direct mappings between the local vocabulary and the vocabularies used by other information services.

7.3.5 To test the viability of the cooperative work between this mapping middleware and other services

This middleware framework proposed to:

- re-use the mappings from different terminology services,
- establish mappings between different vocabularies from different terminology services and a DDC spine,
- and enable the local subject librarians to create the mappings between their local taxonomy and DDC.

However, it was found by all the evaluators that different organisations may create the mappings in different ways, and it is difficult to ensure the consistency and quality of the mappings. For this reason, two methods were suggested in the evaluation and these two methods should be implemented together. In the first method, one central team with

sufficient expertise should be formed and should have the responsibility for maintaining the mappings from different terminologies, and improving the consistency of the mappings from different terminology resources. In the second method, it is necessary to create a metadata scheme to characterise various mapping sets from different terminology resources. In this manner, the central team described in the first method could look into the metadata records describing the mappings set, and then decide appropriate methods to improve the consistency of the mappings from different provenances.

7.3.6 To test the technical feasibility of the information architecture that is used for the development of this middleware system

Because different terminology services, such as OCLC TS, HILT, STAR, etc., all use *different access protocols, query languages and encoding formats, it was accepted by all* the evaluators that using the knowledge base is a powerful way to translate a user's query into different forms of queries that different KOS or terminology services could recognise, gain the terminological results from different terminology resources, and convert the returned results into a consistent format for the client of this middleware system. However, with the increase of the number of integrated terminology resources, how to extend the knowledge base to incorporate more terminology resources is still an important issue.

7.3.7 To examine user interactivity with the subject cross-browsing interface, and test if the users could use the subject cross-browsing interface to find metadata records

The use scenarios tested in the evaluation were accepted by all the evaluators, and a number of new features were suggested to add into the scenarios for improving the interactivity between the system and users. These new features focus on adding more contextual information to the returned mapped concept for improving the term disambiguation, and using Google-style "do you mean" plus the mapped terms to simplify the user browsing interface.

In addition, more new use scenarios were suggested in the evaluation. One of the suggested scenarios was based on the use of a local taxonomy to index the collections integrated within the library portal. In this context, this local taxonomy could behave as a subject collection finder to help users identify relevant collections. Another scenario focused on using the local taxonomy to index various KOS used by different collections, or creating the shallow mappings between the local taxonomy and the KOS used by different collections. In this case, the local taxonomy could be a starting point to guide the users to jump to other vocabularies.

7.4 Final conclusion

As described in Section 7.3, most of the evaluation objectives were achieved, and a number of solutions to the relevant problems were given. For this reason, next chapter will focus on discussing how the new findings and technologies could improve the framework.

Chapter Eight: Discussion

As discussed in Chapters 4 and 7, nine expert interviewees and six evaluators were in agreement with different stages for the development of the middleware frameworks. This chapter discusses the implications in the findings of the different stages of this research. The discussion will be related to the original research questions formulated in Section 2.7. These research questions are given again here:

1. What is the most appropriate approach to improving semantic interoperability between different KOS used by different collections? Is the identified approach appropriate and sufficient to offer a subject cross-browsing service?
2. Who should create the mappings for the service? How should the mappings be created?
3. Which KOS structure is the most appropriate to facilitate the subject cross-browsing services?
4. What use scenarios could this subject cross-browsing service offer?
5. What technologies are suitable for the development of this service?
6. What other services could be integrated with this subject cross-browsing service?
7. What functionality could this subject cross-browsing service offer?

This chapter will be based on a combination of the literature review, findings from the research, and self-reflection.

8.1 Approaches to improving interoperability between different KOS

In order to facilitate subject cross-browsing between different KOS from various terminology resources, two types of interoperability barriers need to be removed. The first type of interoperability barrier refers to the technical heterogeneity between different terminologies. As implied in Section 2.4.3 and Section 7.3.6, in a digitalised environment, different terminology resource providers may use different formats and access protocols to publish their terminologies on the web. It is important to technically integrate these terminology resources. As mentioned in Section 4.5.6, the basic method is to establish a

technical architecture, in which different terminology resources can be linked to each other, and be accessed using relevant protocols and query languages. A query input by a user could cross-search a number of selected terminology resources, and various relevant concepts from different KOS could be returned, converted, and presented to the users. A detailed technical architecture will be discussed in Section 8.5.

The second type of interoperability barrier is related to semantic interoperability between different KOS from different terminology resources. As mentioned in Section 2.4.1 and Section 4.0, the construction of terminology mappings is the most direct method. In this approach, generating the mappings is the most problematic issue. Mapping activity is a long way from reaching a critical mass and will continue to be a burden on those concerned with developing cross searching and browsing services. Ensuring the quality and consistency of the mappings established is the most important part of the development of subject cross-search and browsing services. A number of factors challenging mapping work need to be considered before developing subject cross-browsing services. These factors include the structural models used for mapping, the types of mapping relationships, the mapping logics, collaboration in mapping work between different participants, the ways in which compound concepts are handled, the top-level metadata schemes to describe different KOS, etc. For this reason, the following sections will discuss these different factors related to establishing the semantic mappings between different KOS.

8.1.1 Structural models for establishing the mappings

It is true that establishing many-to-many direct mappings between different KOS is a very precise method to facilitate subject cross-browsing and cross-searching (Section 2.4.1.1 and Section 7.2.8). Based on this approach, a user can be navigated by any vocabulary to get directly mapped terms from all other KOS without interacting with a mediator. However, this approach is very labour-intensive and time-consuming. In a very large information environment where there are a huge number of KOS, this approach is not applicable, because establishing direct mappings is very costly.

One solution is to select or designate a particular vocabulary as the switch language, to which different subsidiary vocabularies can be mapped (Section 2.4.1.1, Section 4.2.1 and Section 7.2.1). In this approach, the applied switch language behaves like a mediator to exchange the terminological information between different KOS. In order to select or designate an appropriate switch language, a number of requirements can be concluded:

1. As noted in 4.2.1, it is important to use a switch language having great granularity and covering most subject areas. Because different KOS may differ in their subject areas and granularity, a switch language has to incorporate the subject fields that these KOS cover. In this context, a number of general classification schemes, such as DDC, UDC, LCC, etc., might be appropriate. These general classification schemes have been developed for many years, and cover most subject areas. It is difficult to find a general thesaurus to cover most subject areas.
2. As indicated in Section 4.2.1, the switch language chosen should be well-known across different communities. Because in this research, it is necessary to map a variety of KOS to the switch language, and the mapping work may be distributed to mapping workers in different communities, it is necessary that different mapping workers should have an in-depth knowledge of the switch language. In addition, when users from different communities interact with the subject cross-browsing interface, the users may need to interact with the switch language. It would be necessary for users to be familiar with the switch language. In this context, a well-known KOS, such as DDC, LCSH, etc., would be helpful to facilitate mapping generation and user interaction. For example, tens of thousands of libraries are using DDC as do 56 national bibliographies.
3. Based on the findings in Section 4.2.1, it would be preferable that a switch language should be encoded in a well-defined interchange format. When using a switch language to exchange terminological information between different KOS, a wide range of information systems may need to use the data from the switch language, and present the data to their users. It is important to employ well-defined formats for representing the switch language, and then different library portal systems can process the encoded data for their local subject needs. For example, DDC has been encoded in the MARC21 format, and is easy for other

library portal systems to process the encoded DDC data and present the DDC data to their users.

4. As found in the evaluation, a switch language should have excellent concept synthesis capability. Different KOS that need to be mapped through the switch language may differ in their degrees of coordination. In these KOS, some pre-coordinated vocabularies might consist of some complex compound concepts, each of which combines a number of concepts. In this case, a switch language should have the ability to synthesise its own concepts into a compound concept matching against those complex compound concepts from other KOS. In some post-coordinated vocabularies, there are some simple concepts, and a switch language should have the ability to split its compound concepts into several simple concepts that can be matched against the simple concepts in the post-coordinated vocabularies. UDC is an appropriate option. In addition, some other faceted vocabularies, such as BLISS, BSO, etc., might be good options. Compared with these faceted vocabularies, DDC only has limited notational synthesis capability. Within DDC, only very broad concepts can be combined into a compound concept (Section 2.2.1.4).

Based on these requirements, Figure 8.1 represents the four categories used to stand for the four requirements described above, and how different vocabularies are mapped into each category.

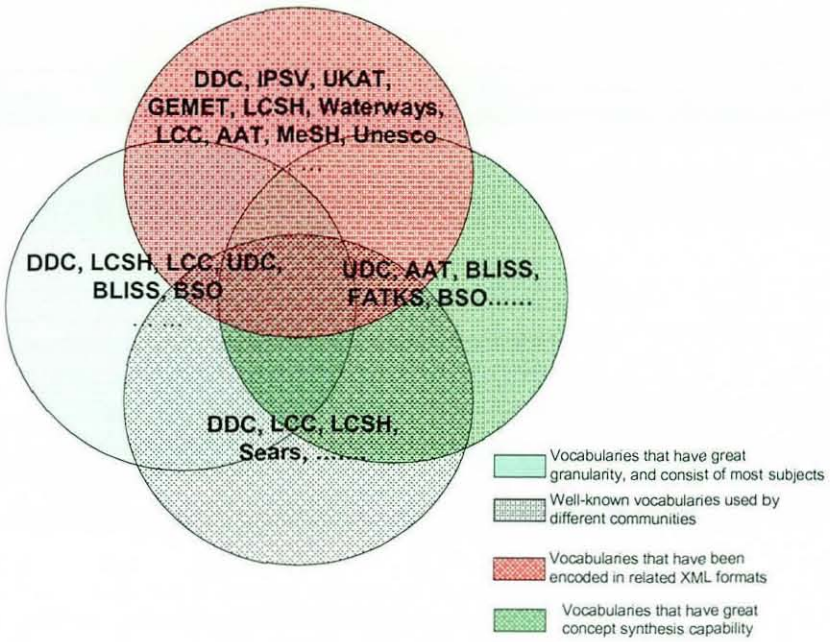


Figure 8.1: Vocabularies are mapped to the four types of requirements

Based on this figure, it was found that it was difficult to find a vocabulary as a switch language to achieve all the four requirements. A number of faceted classification schemes, such as UDC, BC2, etc., with their great notational capability and detailed subject coverage may have great potential to be powerful switch languages in future. However, the disadvantages of these classification schemes are that they currently lack any encoding work that represents complex faceted classification data (e.g. UDC). The basic RDF classes and properties in SKOS format are not sufficient to incorporate complex structures of a given subject matter within a faceted classification. For this reason, it is helpful to extend SKOS format to include more properties and classes to represent facets and particular relationships of a given faceted classification.

In addition, many faceted classifications (e.g. UDC) are not widely-applied by many US communities. Thus, some advocacy work would need to be conducted to encourage the use of UDC world-wide, and further financial support should be provided to allow UDC to be encoded in semantic web-enabled formats, such as SKOS, etc.

In conclusion, using DDC as a switch language still makes sense in the medium to long term. This is because:

1. As mentioned in Section 2.2.1.4, although DDC is not a faceted classification scheme, DDC can also offer a limited capability of notational synthesis. In DDC, it is possible to handle the complex multi-dimensional concepts found in works by combining notations into a concept. For example, it is possible to combine the 026 (library) and 780 (music) into a compound concept 026.78 (music library).
2. DDC is not only a widely-used classification scheme by many academic libraries throughout the world, but also has been applied as a switch language by a number of terminology services such as the HILT terminology service, OCLC terminology service, Renardus, etc. If a project uses DDC as its switch language, it is easier for this project to exchange its terminological information with different terminology services.
3. DDC has been encoded in MARC21 XML data format, and is well-maintained by OCLC. It is easier for a third party to buy the DDC, and convert DDC into its own local system.
4. Many metadata records have been indexed not only by DDC, but also by other vocabularies. In this context, the co-occurrence mappings have been constructed.

8.1.2 Mapping relationships

Considering different mapping relationships, in most cases, it was widely-agreed to use four types of mapping relationships to create the mapping. As stated Section 7.2.2, these relationships include broader, narrower, related, and exact relationships. In these relationships, the exact match relationship should be used to state that two concepts have a similar meaning. However, it is inappropriate to use this exact match mapping relationship to represent the mapping between two synonyms. For this reason, in the latest development of SKOS Project[2], a new mapping relationship called close match was derived from the exact match. The close match relationship is used to map two concepts that have sufficient similarity that "they can be used interchangeably in some information retrieval applications", and the exact match relationship is used to map two concepts

[2] http://www.w3.org/TR/2008/WD-skos-reference-20080125/

considered equivalent. Thus, when establishing mappings between synonyms from different KOS, it would be sensible to consider using the close match relationship in future.

In addition, there have been some co-occurrence and automatic mappings established, but it is difficult to use the mapping relationships described above to represent these mappings. For example, there might be lots of metadata records in a metadata repository that may be indexed using both the term "philosophy" from KOS A and the term "religion literature" from KOS B, but it is difficult to use the defined relationships (broad, narrow, related and exact match) in this research to represent the mapping between these two terms. For this reason, it is necessary to define new mapping relationships to particularly represent co-occurrence mapping. As mentioned in Section 7.2.2, it was very useful to assign the relevance ratings to this kind of co-occurrence and automatic mapping.

8.1.3 Terminology mapping registry

As described in 8.1.2, there are a variety of methods to create mappings between KOS. Therefore, as indicated in Section 4.2.4, it is possible to develop a terminology mapping registry service using a range of metadata elements to record specific characteristics of different terminology mapping data. Based on accessing a terminology mapping registry, the developers of the subject cross-browsing services could discover, investigate, and evaluate different terminology mapping resources, and then comprehensively select some investigated terminology mapping resources for the development of their own subject cross-browsing services. In other words, this terminology mapping registry can provide best practice guidance about how to reuse various established mapping sets for the development of different subject services.

8.1.4 Mapping depth

According to Section 4.5.5 and Section 7.2.4, deep mappings and shallow mappings can provide different use scenarios for subject cross-browsing. The scenarios using the deep mapping approach may often require machine automation when cross-searching different

information resources. When a user selects an appropriate term from its local vocabulary, the machine needs to take the responsibility of processing the established mappings, and employing the mapped terms for retrieving item-level records.

In the shallow mapping approach, the relevant scenarios focus on guiding the end-users to put their own effort and intelligence into selecting appropriate subject hierarchies and separately browsing each controlled vocabulary used by different collections. In this approach, users are asked to select appropriate vocabularies to browse, and consider a number of subject terms within a variety of vocabularies for subject cross-browsing. This approach is intended to help users discover relevant collections and KOS used by these discovered collections. Users could then further interact with the collections and KOS.

Generating mappings is very labour-intensive and time-consuming, and it is important to encourage relevant communities to conduct a variety of mapping exercises, which include deep mapping and shallow mapping. This will enable evaluation of the value of each mapping method in terms of cost and benefit to be conducted before further developing the real mapping service. Because it may be relatively easier and cheaper to create shallow mappings, and shallow mappings, in most cases, could form a basis to further create the deep mappings, establishing the shallow mappings between different widely-used vocabularies could be a good starting point for the mapping effort to measure the costs and benefits before expanding the mapping work. Figure 8.2 shows that the mapping work expansion should begin with establishing shallow mappings.

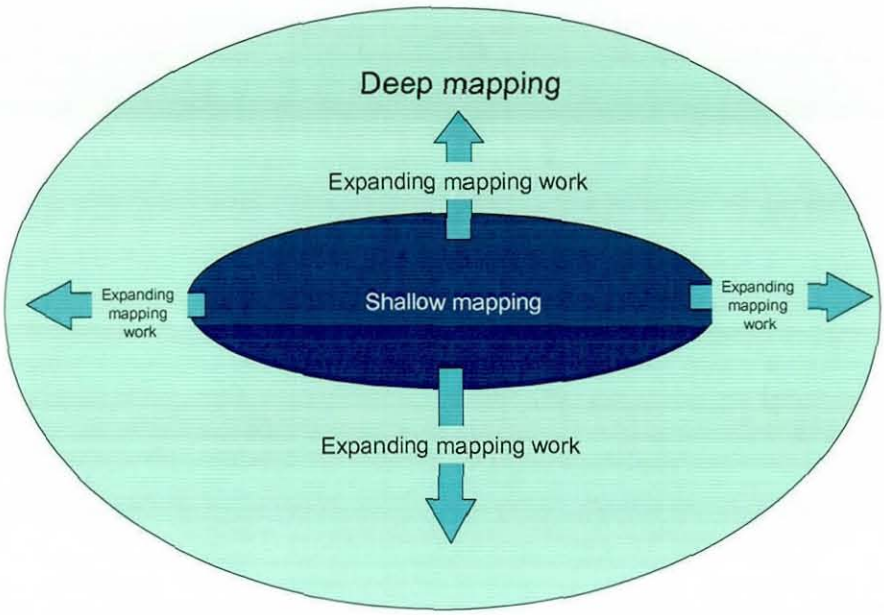


Figure 8.2: The expansion of mapping work based on established shallow mappings between widely-used KOS

8.2 Who should create the mapping?

8.2.1 Mapping creators

A variety of participants could potentially create the mappings for this middleware framework.

1. Collection providers: Most collection providers have their own controlled vocabularies. The cataloguing and indexing staff from these collection providers might be able to create some mappings between their vocabularies and DDC. Most of the collection providers are over confident with their own vocabularies, and they are not really aware of the real benefit of the use of mappings to develop subject cross-searching and browsing services. Therefore, it would be difficult to encourage them to contribute to the mapping work of the middleware framework. Furthermore, because most individual collection providers may lack both the expertise and the will in the area of creating the mappings, which might lead to

inconsistency between the mappings created by these different collection providers, it is not strongly suggested to encourage a lot of collection providers to establish the mappings between DDC and their own vocabularies.

2. Terminology services: A large amount of mapping data is being created by a number of terminology services. A terminology service, as one central team with sufficient expertise, might be able to provide consistent mapping data between different terminologies for this middleware framework. Eventually, there could be a number of terminology services on the web storing different sets of mapping data. In this context, the key role played by any middleware framework is to build a programmatic access interface to access a number of terminology services across the web. For example, it is possible that two terminology services, either of which could use the standard DDC as a switch language, could use the DDC concept URI as the basis for exchanging the terminological information between the two terminology services through the middleware. However, because different terminology services may use different methods to create the mappings, such as automated approach, intellectual mapping, statistical mapping, etc., the mapping data created may be very inconsistent. Therefore, when developing a middleware service to cross-access different terminology services, it is important to conduct some value-added work to improve the consistency between the mappings created by different terminology services.
3. Local institutions using the library portal software: Local librarians, who are responsible for managing the local taxonomies in library portal software, could establish the mapping between their local taxonomy and DDC. This is because: 1). They are familiar with their local vocabularies, and most librarians are familiar with DDC. In theory, they should have enough knowledge to create the mapping between their own KOS and DDC, although in practice, some local subject librarians may face cultural or financial barriers to create this kind of mappings; 2). They wish to offer subject cross-browsing and searching services to their users, so they have their own motivation to create mappings for their subject services.

4. Middleware service itself: As a third party to provide subject cross-browsing service to different library portal systems, the middleware framework provider itself could create the mappings between the DDC and other vocabularies, because in most cases, the services wishing to offer enhanced subject searching would like to contribute the mappings. In addition, this middleware could also be responsible for improving the consistency between the mappings created by different terminology services and local librarians. For this reason, an investigation into the mapping methods used by different terminology services should be conducted. In addition, when the mappings between the DDC and other KOS, and the mappings between the DDC and the local taxonomy are established, the middleware providers could use all the established mappings as a basis to create direct mappings between the local taxonomy and different KOS.

Thus, terminology services, local librarians, and middleware service providers could collaborate to create and improve the mappings from different perspectives. Figure 8.3 shows different roles that could be played by these different parties for providing the mapping data to the users.

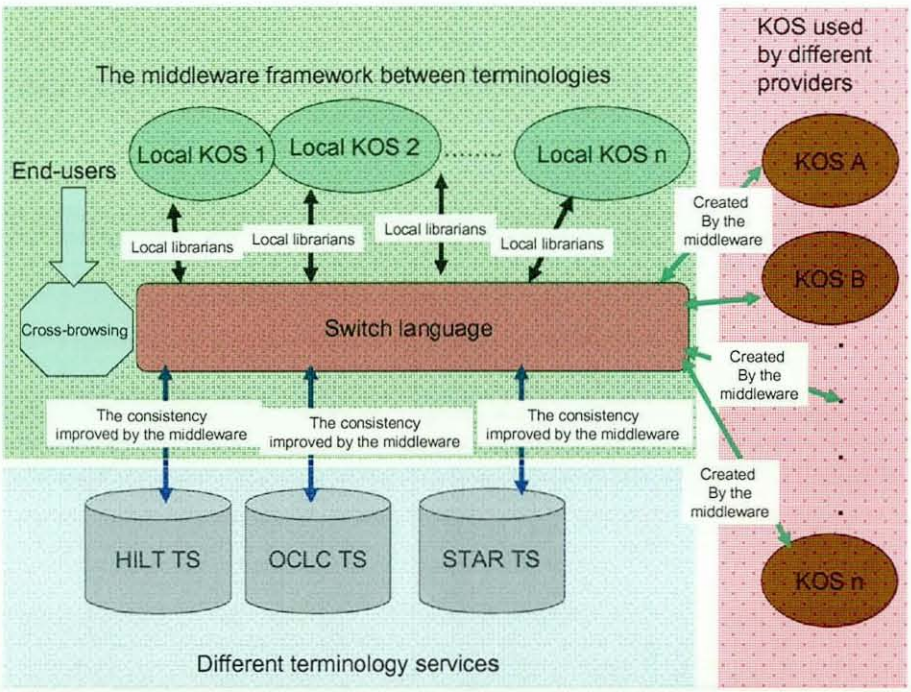


Figure 8.3: Mapping collaboration

8.2.2 Mapping tools

As discussed in Section 4.2.4, it is important to establish a distributed mapping tool to help mapping workers create their mapping data. A mapping tool should be developed based on a well-defined encoding format for shareability, such as SKOS. In practice, SKOS-Mapping Vocabulary was widely-used in different mapping projects for the development of mapping tools, such as Explicator Project, STITCH Project, HILT Project, etc. However, in these projects, people used different mapping relationships and logics. For example, in the STITCH Project, no relevant RDF class, such as RDF bag, or RDF collection, was used to combine individual concepts, but in other projects using SKOS-Mapping Vocabulary, it was possible to use a RDF bag or collection to combine individual concepts, and establish mappings between a RDF bag and a relevant compound concept.

With this in mind, before using a SKOS-based mapping tool to create mappings, it is important to provide relevant functionalities in the tool to enable developers to adapt and further define the SKOS-Mapping Vocabulary to satisfy their local mapping requirements.

8.3 KOS for subject cross-browsing

As discussed in Section 4.1.3, different institutions may like to use different subject structures to facilitate their subject browsing, and even within a single department, different users may like to use different KOS structures for subject browsing. A number of controlled vocabularies can provide the basic subject structures for browsing. These controlled vocabularies include thesauri, classification schemes, subject lists, local/in-house controlled vocabularies, etc. The classification schemes are more useful for providing a browsing function through a hierarchical structure, but other controlled vocabularies are more appropriate for indexing and searching. In this context, it is helpful that a single department could develop or use different controlled vocabularies for different purposes, and have different browsing structures for different kinds of users. The subject browsing structures might include:

1. Local taxonomy based on the departmental structure within an institution: This may be useful for some inexperienced users who are only familiar with their own local taxonomy based on their departmental structure;
2. Subject-specific classification scheme: This would be helpful for some expert users who need more specific concepts to conduct their search. For example, in the medical area, the NLM has been used as a browsing interface by a number of services;
3. General classification scheme: Most users with a library background might like to use a general library classification as a browsing interface to conduct their search.
4. Other KOS: some thesauri or subject headings could also be used as a browsing structure to help users identify the most relevant subject concepts to their subject needs.

Based on using these various KOS, when users interact with their own library portal system, the system can offer several options to allow users to select the most suitable subject structure from a number of KOS based on their subject requirement.

As described in Section 4.2.1, using a switch language to exchange terminological information between different KOS can lead to loss of precision. Section 7.2.8 listed five basic methods to address this problem. This section will discuss the main issues within these identified methods, and indicate the trends for the further development of subject cross-browsing services based on the use of switch languages. It is possible in future that some faceted vocabularies with good concept synthesis capability would be candidates to become a common switch language instead of using DDC (discussed in Section 8.1.2), but using DDC as a switch language still makes sense in the medium term. Thus, this section will discuss the important issues, which are related to the use of DDC as a switch language.

Based on reviewing the methods to solve the indirection problem outlined in Section 7.2.8, it can be summarised that three main sources of intelligence could be added into the middleware system to reduce the loss of precision derived from the use of DDC as a switch language. These intelligence sources consist of:

1. **Intelligence from the mapping staff**: It is very valuable to establish direct mappings between the local taxonomy and different KOS. In this context, the users can get more accurate mapping data from the middleware without interacting with the DDC spine. When highly in-depth mappings are established between the local taxonomy and other KOS through DDC, and accurate mapping relationships are assigned to the mappings, it is easy to use the established mappings to further create more direct mappings between the local taxonomy and different KOS. In this case, a great amount of human intelligence and effort from mapping workers is required.
2. **Machine intelligence**: A number of machine-based automated algorithms could be developed to improve the functionality of the middleware system in different ways. Firstly, it would be helpful to use an automated string matching algorithm to match the terms from different KOS, and produce the automatic mapping. Although the automated mapping data might not be accurate enough, the mapping data could provide different perspectives for retrieval purpose. For example, the automatic mapping tools could find some mapped terms that are not easily discovered by human mapping staff. Secondly, when the mapped terms are returned, a query expansion algorithm could be used to expand from the returned terms over the thesaurus-based network to produce a neighbourhood of subject terms semantically close for retrieval purpose. In this situation, the end-users may be able to select more appropriate terms from these expanded terms to further conduct their search. Thirdly, as described in Section 7.2.8, it is possible to develop some automatic algorithms to enable the machine to decide the direction of expanding the mapped concepts in a KOS. This would depend on analysing the different mapping relationships used to establish the mappings between the local taxonomy and DDC, and mappings between DDC and other vocabularies. For example, when a term "white-colour cats" in the taxonomy is mapped to the term "cat" in DDC by using broad-match relationship, and DDC concept "cat" is mapped to the term "cat" in UKAT by using the exact-match relationship, it is desirable to use the appropriate algorithm to decide to expand the term "cat" in UKAT to incorporate all its narrower terms.

3. **Intelligence from the end-users**: It is important to develop a user-friendly interface encouraging users to use their own intelligence to refine their search. In most cases, end-users might only be concerned with the metadata results finally returned rather than the mapped terms used for searching. However, if they find the returned items irrelevant to their subject requirements, it would be very helpful for the users to get some suggested terms from the middleware system. In this situation, it is possible that the users could use their intelligence to consider and compare these suggested terms, and make judgement on selecting the most appropriate subject terms to refine their searches. Potentially, the terms selected by the users might be more accurate than the mapped terms returned. In addition, a number of visualisation tools should be developed to improve the interactivity between the KOS-based system and end-users, in which the end-users might be encouraged to add their human intelligence into the system to improve the human-computer interaction.

Therefore, a comprehensive combination of these three types of intelligences is needed to further solve the problems driven by the use of switch language. A number of use scenarios could be developed using combinations of various intelligences. These scenarios will be listed and discussed in the next section.

8.4 Use scenarios facilitating subject cross-browsing

8.4.1 Scenarios to create direct mappings between the local taxonomy and different KOS

When the mappings between the local taxonomy and different KOS are established through DDC as a switch language, it would be helpful to use the existing mappings to further develop the direct mappings between the local taxonomy and different KOS. During this process, it would be helpful to apply an appropriate query expansion algorithm to expand the mapped terms from the KOS used by different databases to return a number of concepts considered semantically close. This may enable the mapping workers to find more appropriate terms from the expanded concepts, and select some of the expanded concepts to establish more accurate direct mappings between the local taxonomy and different KOS. However, most query expansion algorithms are used to

expand a subject term to incorporate all the concepts considered semantically close to it, some of which might be too general or too specific to the mapping work. Thus, it is important to develop an algorithm to make the machine analyse the mapping relationships used to establish mappings between the local taxonomy and DDC, and mappings between DDC and other KOS, and decide the direction of expanding the mapped concepts. As described in Section 7.2.8, a specific semantic expansion algorithm could be developed.

With this in mind, Figure 8.4 shows an example of a query expansion algorithm semantically expanding a mapped KOS concept to return a number of its narrower concepts, and the mapping workers selecting one of the narrower concepts to establish the direct mapping with the concept in the local taxonomy.

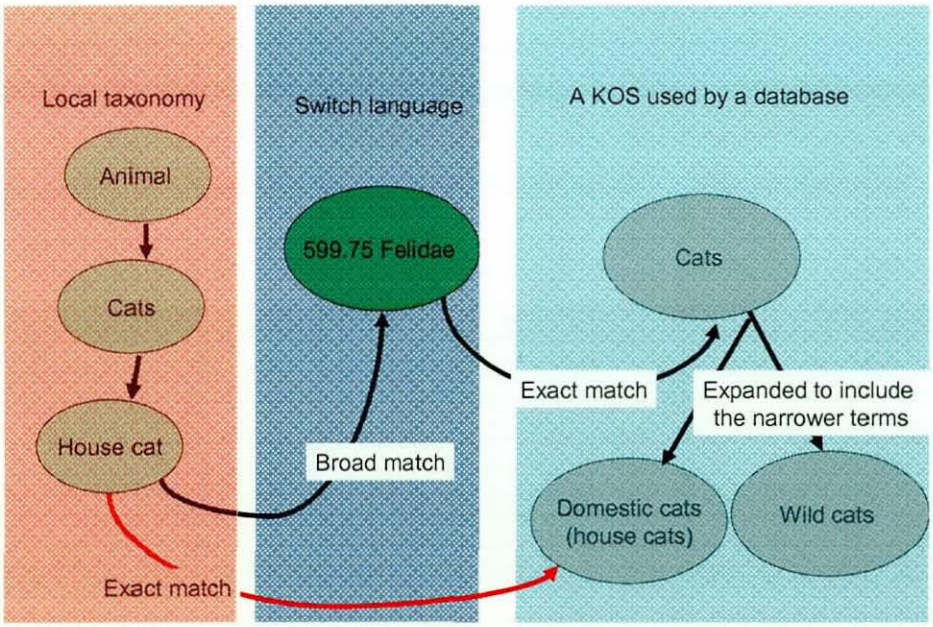


Figure 8.4: Discovering a direct mapping based on established mappings

In this figure, the black curved lines represent the mappings already established between the local taxonomy and switch language, and between the switch language and an external KOS. The red curved line represents the mapping discovered by the mapping

worker, which is based on analysing the established mappings between the local taxonomy and the switch language, and between the switch language and the external KOS.

Based on using this method to establish direct mappings without using DDC, through their local taxonomy, users can get more accurate mapped terms for conducting their subject searches. However, before a lot of mappings between DDC and a lot of KOS are finished, it is difficult to use this method to further develop the direct mappings. Thus, it is important to use some automated tools to assist the end-users, and encourage them to add their own intelligence into the system. This will be discussed in the next section.

8.4.2 The proposed functions of the middleware framework

This section will discuss the basic functions that the middleware can provide to facilitate subject cross-browsing. Different functions could be combined in different ways to offer different subject cross-browsing use scenarios. In addition, it is worth noting that the functions are very similar to the relevant functions provided by other terminology services. It is desirable that different terminology services can harmonise these functions in more standard ways. This will save a lot of effort in re-implementing the relevant functions separately for each terminology service. These harmonised functions are listed and discussed below:

Search_collection function: When a user selects a concept from the cross-browsing structure, it is expected that the user can access the relevant collection to this concept. This function would be used to help the users find relevant collections based on the subject term selected from the local taxonomy. However, in most cases, the users' requirement will be to map a very specific subject search at a very deep level of granularity up to a collection classified at a shallow level and then down again, within the local scheme used, to a level of granularity appropriate to the original query. In other words, most collections are indexed using very general concepts, but users like to select more specific terms to find relevant collections. For this reason, it is desirable to develop a truncation algorithm to successively truncate the specific concept notation that the users

select in the local taxonomy to include a number of more general terms. These general terms can be used to find relevant collections indexed by using the general concepts.

Get_mapped_term function: When a user selects a concept from a KOS as a search term, this function would be applied to get the mapped concept of this selected concept from another given vocabulary. For example, when users click a concept in their local taxonomy, using this function can help users find the relevant mapped concept from DDC, and when the mapped DDC concept is gained, using this function can help users find the mapped concepts from other vocabularies, which have already been mapped with DDC.

Get_expanded_term functions: It is important to use appropriate query expansion algorithms to expand a given concept to include a number of other concepts considered semantically close. However, different concepts in different terminology resources may differ in their data formats, access protocols, and the database systems where they are located. Thus, in this framework, it is important to develop or apply a variety of queries to search and expand different terminology resources. For example, it is appropriate to use CQL-based queries to formulate this query expansion function to search against the terminology resources using the SRW/U protocol, and it is suitable to apply SPARQL queries to formulate this function to search against the terminology resources using the SPARQL protocol. It is worth noting that in this framework, the query expansion functions are mainly used to expand the mapped terms within their own vocabulary.

In addition, when a number of semantically expanded terms of a given mapped concept, such as the narrower terms, broader terms, related terms, sibling terms, etc., are returned, it is crucial to develop ranking algorithms to rank these expanded terms. The ranking algorithm can be based on the measures of semantic closeness between terms. For example, in most cases, it is possible to assign a higher mark to the broader concept of a given concept than its narrower concepts. More importantly, because there are some proposed algorithms that could be designated to analyse the mapping relationships used to establish mappings between the local taxonomy and DDC, and mappings between DDC and other KOS, and decide the direction of expanding the mapped concepts (See

Table 7.2), the decided direction of the semantic expansion can also help the ranking algorithms give some relevant subject terms higher marks. In addition, in this function, it would be hoped that users could define the direction of query expansion. For example, when a mapped term is returned, the users can let the machine expand this mapped term to include all its related terms, or broader term.

Based on using the "get-expanded-terms" function, it is possible to offer users a number of terms considered semantically close to the returned mapped terms. As discussed in Section 8.3, these expanded terms can encourage users to use their intelligence to re-formulate the search terms, and improve the precision of information retrieval.

Get_hierarchy function: When users want to browse the structure of a specific KOS that is recorded in the middleware, this function would be used to query the terminology resource where this specific KOS is located, retrieve the terminological data about this KOS from the terminology resource, use the appropriate algorithm to display the retrieved data in a hierarchical tree, and present the hierarchical KOS information to the users.

In some cases where only shallow mappings are established between the vocabularies, it is important to separately display each of the KOS structures, and let users be able to browse different vocabularies. In this case, users can start with the local taxonomy, use the "get_mapped_term" function to get the shallow mappings, and use the highly-mapped terms to retrieve and display the KOS structures from different terminology resources. Thus, this function would be helpful for dealing with shallow mapping data between different vocabularies, and finding the most relevant subject terms to search collections.

Search_vocabulary function: This function would be used to limit users to search several specific KOS integrated within the middleware. In this context, a user could input some text into a search box, select a suitable KOS to search, and use this function to get the relevant subject concepts from the selected KOS. Furthermore, when the relevant concept is returned from a selected KOS by using this function, the "get_expanded_term"

function can be used to semantically expand the returned concepts, and let the users select a number of terms considered semantically close to the returned concepts.

Get _collections function: When users identify the most appropriate mapped terms from a specific vocabulary to cross-search a number of relevant collections, this function would be used to return the information about all collections that use this vocabulary, and are cross-searchable by the applied meta-search engine. Thus, the users can select some of these collections, and use the identified mapped terms to cross-search the selected collections.

It is worth noting that the names of the functions described above are just used to simply represent the meanings of the functions, and are not related to any technical standards. All these functions are supposed to be provided by this middleware framework. Different library portal systems could set up the clients to use these defined functions to provide the subject cross-browsing. Because different terminology resources, such as some terminology services or different KOS in different formats, might use different protocols, functions, formats, and query languages to set up their services, it is important to map the defined functions in this middleware to different functions provided by different terminology resources.

In this context, the knowledge base of this middleware was developed for this purpose. Within this knowledge base, different clients could be set up to manipulate the data from different terminology services, and different queries could be established to query against different KOS in different formats. Figure 8.5 presents the basic principle of this knowledge base working with other terminology resources.

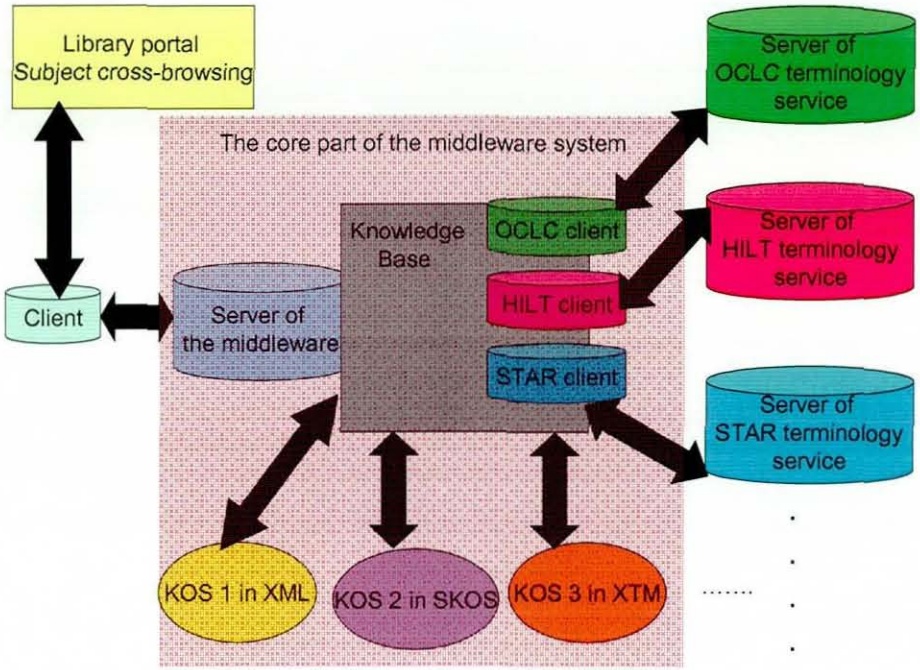


Figure 8.5: The basic architecture of the knowledge base

Taking the HILT Project as an example, the functions defined in this middleware could be linked to the HILT's functions. The HILT functions with use cases are:

- "HILT's get_ddc_records" function: When a user inputs a query term and use the "HILT's get_ddc_records" function, the HILT terminology server will return all DDC notations and captions that match the users' query term. Neither broad concept nor narrow concept of DDC concept matched is returned. The user can select appropriate DDC concepts from a number of relevant DDC concepts to conduct a further search (disambiguation). In this case, based on this HILT function, it is possible to use DDC concept notation as the basis for exchanging the terminological information between the HILT terminology service and this middleware framework.
- "HILT's get_all_records" functions: When a user inputs a term and uses the "HILT's get_all_records" function, the HILT terminology server will return all relevant DDC concepts that matched the user's term and the mapped concepts in other KOS. Thus, the user can select either appropriate DDC concepts or the mapped concepts in other KOS to conduct a further search (disambiguation, query expansion by mapping). In this context, the middleware could harvest these

mapped terms, and use the "get_mapped_term" functions to help the users find the mapped concepts from other vocabularies.

- The HILT function to present a KOS in the hierarchical way: A user can select to browse a number of KOS on the fly in the HILT. This function could be mapped to the "get_hierarchy" function to collect the terminological data from the HILT database, and use this HILT function to present the subject browsing structure of a selected KOS to the users.
- "HILT's get_filtered_set" function: When a user types a query term and select a vocabulary (e.g. LCSH) to query, the HILT terminology server will return all the relevant concepts from the selected vocabulary that match the query term. This function could be linked to a "search_vocabulary" function to limit to search a specific KOS within the HILT database.
- Two more functions were designed within HILT to support query expansion. "HILT's get_parents", a function to retrieve all the broad terms of a given concept, and "HILT's get_children", a function which retrieves all the narrower terms of a given concept. These two functions can be linked to the "get_expanded_term" function in this middleware to help the end-users to expand the mapped terms. However, because HILT has no function to expand the related terms of a given concept, in the current state, it is impossible to retrieve the related terms of a given concept from the HILT database.
- HILT collection finder function: When a DDC concept is provided, this function could guide the user to find the relevant collections to this given DDC concept. Within this function, a DDC truncation algorithm would be applied to find more general collections. This HILT function could be linked to the "search_collection" function to find the relevant databases indexed in the JISC Collection Registry.

Based on reviewing the different functions provided by the HILT terminology service, it can be found that most of functions developed in this middleware can successfully link to the functions provided by HILT.

8.4.3 The proposed scenarios from the users' perspectives

This section will discuss how an end-user can actually interact with this subject cross-browsing service to find relevant information. Four specific user scenarios are suggested:

1. When users begin to use the subject cross-browsing service in a library portal system, they are asked to select a KOS browsing structure from a number of vocabularies. These vocabularies include DDC as a switch language, the local taxonomies, and various KOS used by different collections. When the users select a vocabulary structure, the "get_hierarchy" function will be used to extract the appropriate data from the terminology resource storing this selected vocabulary, and present the hierarchical information about this vocabulary to the users. If users select a particular KOS structure (rather than DDC and local taxonomy), which belongs to the vocabulary used by some other collections, the "get_collections" function will be used to get al.l the collections using this particular vocabulary.
2. If the users choose DDC as the browsing structure, and click the relevant DDC concept, the "get_mapped_term" function will be used to get al.l the mapped terms from different vocabularies, and the ""get_expanded_term" function will be used to expand all the mapped terms from the different vocabularies. It is expected that the users select some of these mapped or expanded terms to further conduct their search. After the users choose the terms from different mapped or expanded terms, the "get_collections" function will be used to get al.l the collections using these selected mapped terms, and let the user know which collections they can use the selected mapped terms to cross-search. It is assumed that the users themselves could filter some collections that use these terms but are irrelevant to the users' subject need. In parallel, when the users select DDC concept, the "search_collection" function, which is assisted by a notation truncation algorithm, will be used to get a number of collections that are relevant to the selected DDC concept. In this case, users could go into these collections' own interfaces to locally search the collections.
3. If users choose the local taxonomy for the subject cross-browsing interface, and they select relevant subject terms, the "search_collection" function will be used to

get a number of collections that are relevant to the selected subject terms, and the "get_mapped_term" function is used to get the mapped DDC concept. The mapped DDC concept can be semantically expanded by the function "get_expanded_term", and a number of terms considered semantically close to the mapped DDC concept could be returned to the users. Users could select some of these returned DDC concepts. When the users select appropriate DDC concepts, the "get_mapped_term" function will be used to get al.l the mapped terms from different vocabularies, and the "get_expanded_term" will be applied to expand these mapped terms. The users could re-choose the terms from these mapped terms, and then cross-search the collections using these terms.

4. In most cases, users input a term to cross-search different collections, and they might not want to use the subject cross-browsing service. For this reason, users could get more terminological support from this middleware system without using the subject cross-browsing interface. In this situation, when the users input a term to cross-search, this term will become the query to cross-search different KOS integrated within this middleware. In other words, the "search_vocabulary" function will be used to cross-search different vocabularies, and return the detailed terminological information about the terms that are matched to the users' query. The returned terminological information might consist of the scope note, narrower terms, broader terms, preferred terms, non-preferred terms, notations, and related terms of the matched terms. The "get_expanded_terms" could also be used to expand the terms to include more relevant concepts. During this process, "get_collections" could be used to get the collections using the terms that the user finally select. Thus, the users could select some of the returned terms to re-formulate their query to cross-search the collections using these terms.

In order to achieve all the functions and scenarios discussed above, it is important to identify appropriate technologies to develop this framework. The next section will discuss the technological issues for the development of this framework.

8.4.4 A KOS browser

Based on the functions proposed in Section 8.4.2 to access different KOS data, a KOS browser service should be developed and integrated within a library portal system. Different KOS could be shared within this KOS browser service. Through the KOS browser, users could access, browse, look up, and exploit different KOS and their mappings. This KOS browser service should not only include a browser for human subject cross-browsing, but also a shared server that other services can access for getting terminological data. For example, when users identify appropriate subject terms from particular KOS within the KOS browser service, the KOS browser should automatically use the selected term to interact with the cross-search engines provided by a library portal system, and provide integrated access to different online databases.

Nowadays, a number of applications have been developed to facilitate KOS browsing within library portal systems, such as /facet[3] browser, the KOS browser developed in the STAR Project[4], the facet KOS browser developed in the STITCH Project[5], etc. In some cases, visualisation techniques could be integrated within these browsers to enable efficient retrieval.

Through the KOS browsing service, mapping staff can browse, select, find and semantically expand subject terms from different KOS. Therefore, it would be helpful to integrate mapping functions/tools within this KOS browsing service. In this context, mapping staff can easily create terminology mappings within the KOS browsing service, and submit the established mappings to the middleware system.

[3] http://slashfacet.semanticweb.org/
[4] http://reswin1.isd.glam.ac.uk/STAR/SKOS_WS_EH/SKOS_WSClient.htm
[5] http://www.cs.vu.nl/STITCH/KB_Rijks_demo.html

8.5 Technical architecture

8.5.1 Identification mechanism

A concept is defined as "an abstract idea or notion used to represent a subject topic" (Miles and Brickley 2005). In most cases, a number of terminological information elements can be used to represent a concept. These terminological elements may vary in different terminology resources. When establishing the mappings between concepts from different terminology resources, it is helpful to use a unified mechanism to give these heterogeneous concepts clear identifiers.

In the context of the semantic web, URIs are used as an identification mechanism to provide persistent concepts from different vocabularies. For this reason, it would be desirable in this framework to give URIs to all concepts from different terminologies so that the mappings could be technically established based on the use of URIs rather than other terminological elements. However, many KOS providers do not have persistent identifiers in place. An identification resolving mechanism needs to be developed to translate different concept identifiers into consistent URIs. When users select a mapped concept from the subject cross-browsing interface as a search term against other terminology resources, the URI of the mapped concept could be translated to a different identifier for the relevant terminology resources.

8.5.2 Access protocols

Section 2.5.3 suggested that different terminology resources may use different access protocols and query languages to let other services query their terminological information. A number of main access protocols were discussed in Section 4.3.1, such as SRW, SRU, SKOS API, SPARQL, etc. Based on reviewing the characteristics of these protocols, it was summarised that there are three basic requirements related to selecting access protocols for the development of terminology services:

1. An access protocol should incorporate rich query languages to facilitate reasoning over the semantics within the vocabulary. Because each vocabulary is a very rich

semantic network, it is important that a query language within a protocol should recognise and process the semantics, and provide a number of semantic-based functions, such as query expansion, semantic reasoning, etc.

2. It is desirable that access protocols should be lightweight. In SOAP-related protocols, because each SOAP operation needs to be predefined, and be given an URI, using a SOAP server is quite 'heavy' due to the overheads imposed by SOAP. For example, when using SRW and SOAP within the HILT Project as protocols for the development of a terminology service, a user's query would be translated into CQL requests that comply with the SRW protocol to query against the SRW server, then the CQL request would be taken by the SOAP server and mapped into appropriate SOAP functions. Based on these SOAP functions, the SQL request is sent to the relevant terminology database and the results returned are wrapped into relevant encoding formats using the SOAP protocol. In this example, using SOAP and SRW is quite demanding as the requests have to access two M2M layers. It would be preferable to develop direct programmatic access to the terminology database. Compared with the SOAP-related protocols, RESTful protocols, such as SRU, SPARQL, etc., make it much less expensive to develop the basic functionality of a terminology service.
3. As mentioned in Section 4.3.1, an access protocol should support powerful text-based query. This is because terminology services and controlled vocabularies are text-based systems. It is necessary to use a text-based query language within an access protocol to facilitate many text-based query types, such as phrase queries, wildcard queries, proximity queries, range queries, string truncation queries, and so on. As discussed in the early part of this section, some semantic-web-based protocols, such as SPARQL, are not powerful in facilitating text-based queries.

It was found that there is no protocol that could fully satisfy all the requirements listed above. Thus, it is important to combine these RDF-based protocols, query languages or APIs with text-based query languages within a terminology service. For example, in the STAR Terminology Service, the SKOS API is used to process the SKOS data stored in the database to support the semantic query expansion, and the

SQL query language is used to query the full-index added for partial matching on literal strings. See Figure 8.6.

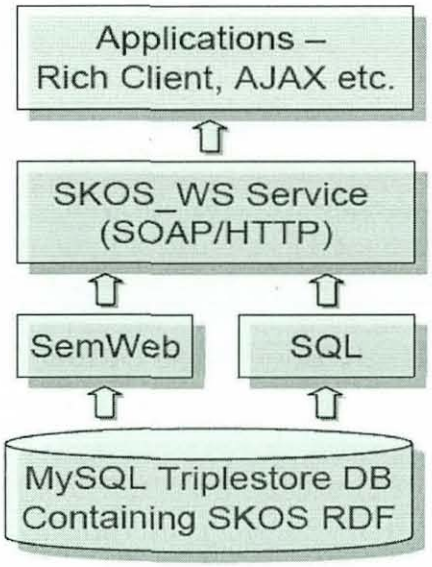


Figure 8.6: The technical architecture of STAR Project (source: Tudhope *et al.* 2008)

As stated in Section 7.2.9, this framework was intended to integrate a variety of protocols to access different terminology resources. With the increase in the number of different terminology resources, the information in the knowledge base would increase, and the knowledge base would become difficult to maintain. Thus, it would be helpful if most terminology services could reach an agreement to use some standard protocols and data formats so that a middleware framework could easily formulate their client requests to cross-access these terminology services.

8.5.3 Representation formats

Different terminology resources may use different interchange formats to exchange their terminological information between the servers and clients. Different encoding formats were discussed in Section 2.5.2. It is preferable that a middleware system, as a terminology server, could use the most accurate format to encode appropriate terminology data for different clients. For example, it would be suitable to use the DD-8723-5 to encode thesauri data, use MACR21 XML for classification to encode

classification data, and use SKOS to represent the concepts and relationships that have been treated as web resources in the semantic web.

As discussed in Section 4.5.7, a middleware framework between terminologies should have the ability to process these different formats, and, through the framework, different representation formats should be converted into each other depending on the specific client requirements. Because a client may be able to process a few encoding formats, and different clients may have different preferences for processing different encoding formats, it is important for this framework to add more encoding formats to represent different vocabularies for different clients, such as MARC21, Zthes XML Schema, etc.

8.6 Other services

In this section, three services were discussed. These services could potentially be integrated within the middleware system to improve the functionality.

8.6.1 Social tagging technologies

Social networking initiatives such as social tagging and folksonomies are sometimes widely-agreed as an approach to reducing the burden of generating and managing KOS. Social taggings can play an important role in expressing users' perceived subject requirements. Some controlled vocabularies, which are based on the use of notations to represent concepts, are not very meaningful for users. In other words, the captions within some KOS do not represent the real meaning of DDC concepts. It is worth noting that a number of services may wish to map the tags added by the users to existing controlled vocabularies. The taggings can further complement explanation of a concept that may be not easy to understand. In parallel, in some sites, tagging is taken very seriously based on well-defined guidance and rules. Subsequently, tags can be analysed and linked into a folksonomy. It is possible that these folksonomies could be further developed to become conventional controlled vocabularies.

In addition, it is worth noting that in social tagging research, "tag clouds" allow tags to be displayed in different formats for different purposes. In order to browse various KOS, it might be possible to use this type of display to highlight some important KOS concepts. For example, there are a number of concepts (e.g., DDC, UDC, LCC, BC2, etc) at the same level of granularity within a KOS, in which the concept "DDC" is more widely-used within a subject service than any other concept in this level. In this case, tag cloud technology could be used to highlight the concept "DDC", and allow the concept "DDC" to be displayed differently from other concepts.

8.6.2 KOS registry

As discussed in Section 7.2.11, instead of using a subject cross-browsing interface to search information, a KOS registry can be used as a starting point to discover a variety of terminology resources. An individual KOS may consist of a variety of characteristics. It is important to develop a number of appropriate metadata elements to record what a KOS is likely to contain. In this context, a KOS registry can store "the information allowing the selection of schemes suitable for different purposes, address information for contacting owners and maintainers, hypertext-links to connect to the vocabularies or maintainer sites, information to differentiate between versions and identifiers, names and labels to unambiguously refer to a given scheme" (UKOLN 2008).

In addition, a KOS registry could also provide metadata about the characteristics of different terminology services. Because different terminology services may be developed based on different methods (such as the ways to create mappings, the encoding formats used to exchange terminological information, access protocols, terminology database structures, etc), it is important for a KOS registry to record all these characteristics of different terminology services. A variety of scenarios can developed based on using a KOS registry. These are listed and discussed as follows:

1. The mapping staff could use a KOS registry to access different KOS, and understand the basic structures of the vocabularies that they wish to map, which is helpful for their mapping work.

2. The end-users could use a KOS registry to link to the relevant vocabulary web sites, and the vocabulary web sites can directly link to the relevant information resources. Through the vocabulary structures, therefore, users can be navigated to find relevant metadata results.
3. The service providers: When a service wishes to develop a KOS to index their information, they could go to the KOS registry, and select to reuse some existing KOS within their own service.
4. The collection finders: In a collection registry, it is useful to record which KOS each of its collections is using. In this situation, the KOS used by the collections could be linked to the KOS registry. Thus, it is possible that users select appropriate KOS from the KOS registry to identify relevant collections.
5. The subject cross-browsing and searching service providers: Through a KOS registry, a subject cross-browsing and searching service provider could get a basic understanding of basic mappings established between different KOS within terminology services, and consider ways to re-use the mappings for the development of subject cross-browsing and searching services.

8.6.3 Collection registry

As discussed in Section 5.3, there are two functions that the collection registry can provide. The first function is to work with the meta-search engine to distribute different returned mapped terms to cross-search appropriate collections that are using these mapped terms. In this function, the collection registry needs to record which KOS different collections are using. When the mapped terms are gained from a KOS, through the collection registry, these mapped terms will be become the query to cross-search against the collections using this KOS.

In the second function, different information resources recorded within a collection registry, such as online databases, subject portals, e-print archives, institutional repositorie, etc., can be indexed according to the subject browsing structure used. In this sense, users can be guided by the subject browsing structure to find relevant collections.

It is desirable that a collection registry supports an API to let people use the local taxonomy or DDC to index different collections. When the users click a concept from the taxonomy, the concept will be the query to retrieve the collections indexed by using this concept. See Figure 8.7.

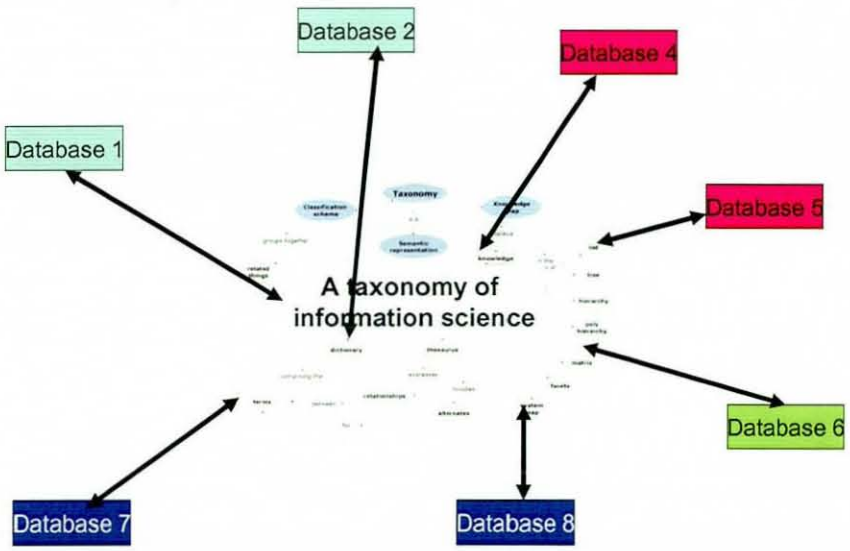


Figure 8.7: Using a local taxonomy to index different collections

Based on the discussion of different aspects for the development of this middleware framework between terminologies, the next chapter will mainly describe the conclusion of this research, and present the further development for this research.

Chapter Nine: Conclusions

This concluding chapter will give an overview of this research, describe how the research findings match the research aim and objectives given in Chapter One, and present some conclusions and recommendations for the further development of this research. The limitations of the research will be identified and discussed.

9.1 Research overview

This research was aimed to develop a middleware framework between different terminologies to facilitate subject cross-browsing service for library portal products. A number of library portal systems have been developed and applied in libraries, some of which can offer search boxes to cross-search different collections. However, most of these portal systems cannot provide a single subject structure or navigator to cross-browse different items from different collections. The main barrier is the heterogeneity between different knowledge organisation systems used by different collections, which can be defined as different subject areas, granularities, degrees of coordination, the use of languages, etc. This makes it difficult to subject cross-browse different collections. The methods to solve this problem point to establishing mappings between these vocabularies, and integrating different vocabularies into a semantic network through these mappings.

This research began with an in-depth literature review to investigate and analyse the methods of using various KOS to improve subject access, and the methods of improving the interoperability between them. A number of essential elements to develop the framework were identified, but two important points were still not clear. First, it was necessary to investigate a number of KOS used by different collections before using appropriate technologies, standards, and semantic methods to develop the framework. Without a clear insight into the different characteristics of various KOS, it is impossible to make the right decision on using appropriate methods to establish the mappings between these KOS. For this reason, an investigation was conducted to review a variety of KOS and their characteristics, and the findings formed a basis to establish the mappings. Second, based on the methods summarised from the literature review, it was

still not clear how these technologies, standards, and semantic methods could be combined to make up this framework. Most of these elements were context-dependent, and may not be suitable to be used for this research scope. It was therefore important to collect in-depth ideas from a number of experts, who were involved in different terminology mapping projects, and match these ideas to this research situation. For this reason, nine expert interviews were conducted to introduce the research context to the experts, and encourage them to match their knowledge to the development of the framework. Much valuable data were gained based on the interviews. These data were analysed to form a basis to develop a theoretical framework.

In the theoretical framework, based on the findings from expert interviews, different components were combined, and a number of technologies were chosen. This framework offered a number of guidelines and theories to help develop a real middleware system between different terminologies. Subsequently, a design research approach was undertaken, and a prototype system was developed based on the guidelines and theories. The prototype was used as a basic platform for evaluating the theories developed. In this context, six experts (not those who were previously interviewed) were asked to interact with the prototype system, and provide the suggestions to improve the framework. The evaluation findings were obtained from these experts, which formed a foundation to develop the final theories.

9.2 Achievements of research objectives

There were five research objectives with their sub-objectives identified. This section will discuss how the research findings match these objectives.

Objective 1: To understand the basic principles to develop controlled vocabulary-based information retrieval systems for library portal products, and improve the interoperability between different controlled vocabularies.

Objective 1.1: To investigate the methods to develop library portal products, such as cross-searching technologies, data conversion, etc., and understand how the various terminology resources could be embedded into library portals:

Based on the literature review, it was found that one of the most important services that a library portal system can provide is subject cross-searching. Methods to develop cross-searching focus on establishing mappings between different metadata elements used by different collections, setting up appropriate protocols and queries to access various collections, and converting distributed results into a consistent format. However, because different collections may use different controlled vocabularies to organise their information, the development of library portal cross-searching services has been greatly impeded by the heterogeneity of different KOS used by different information resources. In addition, due to the lack of interoperability between different KOS, it is also impossible to develop a subject cross-browsing interface, through which a user can be navigated to find relevant information across different collections. Therefore, Objective 1.1 was partly accomplished by the literature review, and it is important to further look into different KOS and the semantic heterogeneity between them, and explore methods to improve the interoperability between these KOS.

Objective 1.2: To investigate how various controlled vocabularies support information retrieval systems in both traditional and innovative ways:

Objective 1.1 was fully accomplished by the literature review. It was found that various controlled vocabularies, such as classification schemes, subject headings, thesauri, ontologies, etc., greatly differ in their subject areas, degrees of coordination, granularity, concept synthesis capability, use of languages, data formats, etc.

Three basic functions, which a controlled vocabulary may be able to provide, were identified. The first function is a term disambiguation function, which allows the users to see the contextual information of a given concept, such as the scope note, definition, non-preferred terms, hierarchical information that the term is related to, etc. In this case, users can understand the real meaning of a given term, and avoid errors caused by misunderstanding the concepts. The second function is a semantic expansion function,

which employs computing algorithms to conduct the expansion from a given term over the thesaurus-based network to produce a neighbourhood of subject terms semantically close for retrieval purposes. In most cases, this function is treated as an effective method to improve the recall for information retrieval. The third function is subject browsing (also called knowledge navigator), which employs the subject structure of a KOS as a browsing interface to navigate the users to find the most relevant subject terms. This function can enhance the interactivity between the users and KOS-based information retrieval systems, and encourage users to input their own intelligence to improve the search.

Different types of vocabularies were designed for different retrieval purposes, and have their own strengths and weaknesses. For example, a pre-coordinated classification scheme (e.g. DDC) may have a good subject structure to support subject browsing, but in many cases, its captions cannot represent real meanings. In this case, it is not sensible to apply captions as search terms to conduct a subject keyword search. A post-coordinated thesaurus may be more suitable to support Boolean-based search algorithms, but its subject terms may be too specific to form a good browsing structure. With this in mind, it is necessary for different vocabularies to co-exist in the medium to long term, and it is not practical to apply a universal KOS to all collections.

Objective 1.3: To review a variety of widely-used controlled vocabularies, and understand their subject areas, levels of specificity of concepts, degrees of pre/post-coordination, semantic relationships, the use of languages, representation formats, services using them, types of information that these controlled vocabularies index: An investigation into different characteristics of KOS used by different information resources was undertaken to achieve Objective 1.3. Based on reviewing a variety of KOS, it was found that many databases may be indexed according to more than one KOS for different purposes. For example, a database may use a classification scheme as a subject browsing interface, and use a thesaurus to support text searching. A great number of in-house vocabularies are also being used by these databases, but among thirty-five selected databases, none of them employ a fully-fledged ontology to describe their information.

In addition, it was found that two knowledge organisation systems from different domains may differ greatly in terms of their structure, terminology, and granularity, but there are much more similarities between the KOS in the same domain. In fact, a number of vocabularies were established based on merging a number of controlled vocabularies in the same domain into a semantic network, such as UMSL. It is relatively easy to establish the mappings between the vocabularies in the same domain. For this reason, before conducting a large amount of mapping work between different KOS from different domains, it is expected to begin with merging different KOS in the same subject area into a semantic network.

Objective 1.4: To investigate the methods to improve the interoperability between different controlled vocabularies:

Based on the literature, the findings point to applying three different approaches. The first approach focuses on establishing mappings between different KOS. A combination of automated algorithms, statistical approaches, and human effort could be applied to facilitate the mapping work. With the increase in the number of KOS that need to be mapped, it is important to use a switch language to exchange terminological information between these KOS. In most cases, DDC was widely-agreed as an appropriate switch language with some reservations. Other issues related to the mapping work might consist of other structural models for mapping, mapping direction, the degree of equivalence between terms, the mapping logics, the collaboration in mapping work between different participants, the methods to handle compound concepts during mapping, the top-level metadata elements to describe characteristics of the mapping data, etc. Because it is difficult to estimate the efforts to address these issues for the development of a terminology mapping system, further research was required to collect a more realistic sense of the technical and mapping overheads from other similar projects.

The second approach focuses on harvesting heterogeneous metadata records into a centralised database, and using a single vocabulary to index these collected metadata records. This centralised database needs to hold a huge number of metadata records, and

it would be very labour-intensive to index these records based on the human efforts. Therefore, in many cases, it is preferable that this approach could employ some automated indexing technologies to automatically create the subject metadata.

The third approach is based on establishing an integrated environment among different terminology resources, using different protocols and queries to cross-search different terminology resources. In this approach, these terminology resources can be treated as databases, and a number of access protocols can be applied to cross-search these databases, such as SRW/U, Z39.50, SPARQL, etc.

According to these three methods, it would be possible to develop a terminology mapping middleware framework to combine these methods in comprehensive ways. Thus, Objective 1.4 was partly accomplished by the literature review, and further research was required to gain a basic understanding of how to use different methods to complement each other. See next objective.

Objective 2: To explore the methods to combine a variety of terminology resources, and integrate the combined terminologies into the library portal services.

Objective 2.1: To investigate a number of terminology service projects, and explore how they work within a library portal system, facilitating subject cross-browsing: According to the interviews with a number of experts who were involved in different terminology service-related projects, such as Renardus, OCLC Terminology Service, HILT Service, etc., Objective 2.1 was fully accomplished. In the research, it was found that some mapping data had been established by these projects, and many mappings were based on the use of a DDC spine. Some of these services can offer their mappings to different library portal systems. It is important to develop programmatic interfaces to access these mapping sets from different terminology services, and reuse the established mappings. However, different terminology services may use different mapping strategies, such as provenance (source), methods (intellectual, co-occurrence, other automatic, etc),

concept indicators, and so on. This might cause the inconsistency between different mapping sets. Some of this mapping data might not be suitable for the proposed use scenarios of this research framework. Thus, it is suggested to develop a number of metadata elements to characterise various mapping sets. Based on the metadata elements, some mapping staff with strong expertise could provide intellectual effort to re-generate the mapping data, which could be derived from the established mappings, and guarantee the consistency and quality of the mapping work.

Furthermore, it was found that most terminology services, such as HILT, OCLC Terminology Service, STAR Service, etc., had developed and provided their own API functions to different clients. Different library services can use these developed API functions from different terminology services through the web to facilitate their subject access. However, these functions are being developed using different protocols and query languages. For example, the OCLC Terminology service was developed based on the use of SRW and CQL, but the STAR Terminology Service was based on the use of SKOS APIs. In this context, it is not practical to use a single standard method to access all these terminology services.

Apart from collecting the KOS data in different terminology services, there are a variety of KOS located in distributed information environments, and encoded in different data formats, but these KOS are not included in the existing terminology services. It is important to integrate these KOS into this middleware framework. In this context, the main findings focus on employing XSL technology to convert these KOS into a consistent format, such as SKOS, MARC21, etc., and storing the converted KOS data into a local database. Also, the loss of data precision during the conversion was pointed out. For this reason, it was indicated that this middleware framework can store these KOS in their original formats, and develop or apply different ways to present these KOS data in different formats.

Objective 2.2: To identify the appropriate technologies and standards, and propose a technological architecture supporting the middleware framework that exchanges *the terminological information between different terminology services:*

As mentioned in the last paragraph, no single standard method can be employed to access all these terminology services. Thus, a number of interviews were conducted to collect ideas from the experts who have strong expertise in developing distributed access services. Objective 2.2 was achieved through these interviews.

According to the interview findings, a knowledge base can be developed to cross-access different terminology resources using different protocols, query languages, formats, etc. This knowledge base can store the connectivity details of different terminology resources. Three types of connectivity details were found:

1. Linkings between different functions from different terminology services: Although the functions were designed through using various technologies, protocols, and standards, it was found that a lot of functions from different terminology services were similar to each other. It is possible to summarise a number of common functions from different terminology services. For example, most terminology services include functions to get broader, narrower, or alternative terms for a given concept, and the functions to generate and present a subject structure from a given KOS on the fly. For this reason, it is important that the knowledge base should record information about the common functions from different terminology services.
2. Information about the rules to translate a user's query to different forms that different terminology services can accept: Because different terminology services use different functions based on different protocols, query languages, and formats, it is important to develop a transmission programme within the knowledge base to translate a users' query to different forms of queries that different terminology services can accept. These forms may include different functions, protocol-dependent query languages, etc, to query against the terminology services.
3. Information about the rules to convert the heterogeneous results into a consistent format: Based on using different protocols and servers, the returned results might

be represented in different encoding formats, such as Zthes XML, SKOS, MARC21 XML, etc. For this reason, it is important to develop data conversion programmes to convert these results to a consistent format, and use XSLT programmes to present the results to the users.

Based on these three types of connectivity details within the knowledge base, this middleware framework could cross-access different terminology resources, and could get consistent data from them.

Objective 2.3: To develop a mechanism to be able to exchange terminological information between different terminology services and the local controlled vocabularies:

This objective was accomplished through interviewing four experts who are responsible for developing or managing library portal systems. It was found that in most cases, the general classification schemes are not suitable for most institutions, and institutions may prefer to use local taxonomies, departmental structures, or subject-specific vocabularies as their subject browsing structure. In order to exchange the terminological information between different terminology services and the local taxonomy, the findings focus on using DDC as a switch language. However, because it is a two-step journey from the local taxonomy through DDC to the other KOS, this indirection problem, which might cause loss of precision, was emphasised by the interviewees. It was necessary to explore a more advanced approach to solve this problem.

An expert-based evaluation was conducted to explore the solutions to this indirection problem. The findings were based on using a more faceted classification with great concept synthesis ability to be a switch language, establishing the many-to-many mappings derived from the existing mappings between DDC and other KOS, and employing semantic expansion algorithms to expand the mapped terms and show the users more terms considered semantically close to the mapped terms. Furthermore, as mentioned in Section 7.2.8, a specific algorithm was indicated to make the machine automatically decide the direction of expanding the mapped concepts in a KOS depending on analysing different mapping relationships used to establish the mappings

between the local taxonomy and DDC, and mappings between DDC and other vocabularies.

When establishing the mappings between a pre-coordinated vocabulary and a post-coordinated vocabulary, it is required to consider methods to handle compound concepts for the mappings. The findings focus on employing Boolean operators, defining clear mapping relationships, and developing appropriate mapping logic. From the users' perspective, however, it was found that when users click a compound term to get a number of single terms, they might not be concerned with what mapping relationship is used to establish mapping, or which vocabulary the mapped data comes from. Most users may prefer to get a number of mapped terms without knowing how the terms were mapped. For this reason, it was concluded that using a combination of all the mapped single terms to a given compound concept might be appropriate to suit the users' subject requirements.

Objective 2.4: To define a workflow to distribute the manual mapping work to create and maintain the middleware:

Objective 2.4 was partly achieved through interviewing library portal service providers, and it was found that most database providers were over confident in the KOS that they are using, and it is difficult to encourage them to establish mappings between their own KOS and DDC. Thus, based on the findings, it is helpful that local librarians could establish the mappings between their local KOS and DDC, and the middleware provider could create the mappings between DDC and other KOS, or collect the existing mappings from different terminology services.

However, based on the evaluation findings, it was found that different communities, such as terminology services, local librarians, and the developers of the middleware itself, need to put effort into creating the mappings. These mappings could be created in different ways, such as direct mapping, co-occurrence mapping, on-the-fly mapping, etc. This might lead to the inconsistency of the mappings. For this reason, a centralised team with great expertise should be set up to create, maintain, or enhance the mappings. This

would consolidate a long term commitment to create, collect, enhance, and maintain the mapping, and guarantee the consistency and quality of mapping work.

Objective 2.5: To identify a number of relevant user scenarios to facilitate user-friendly subject cross-browsing:

This objective was accomplished through a combination of interviewing experts and expert evaluation. Based on the findings, the importance of creating a correct balance between the user engagement and automated applications was emphasised. From the users' perspective, it was found that most users are not concerned with which vocabulary a mapped concept comes from, and it is preferable to let the users select mapped terms without knowing the information about the vocabulary. For this reason, using "Do you mean" plus the mapped terms from different vocabularies would be helpful. However, when the mapped terms are not suitable for the users' subject needs, the users might be frustrated by the returned results. In this case, it is important to use a query expansion algorithm to expand the mapped terms to include a number of terms considered semantically close to the mapped terms. In this context, the users could use their intelligence to re-formulate their subject queries.

Objective 3: To formulate a theoretical framework to facilitate subject cross-browsing service for library portal products:

A theoretical framework was formulated in Chapter Five, in which a number of methods and technologies were employed. This objective was achieved based on the collected findings of nine interviews.

In this framework, DDC was selected as the switch language, through which a number of other vocabularies could be mapped together. Different mapping elements from various vocabularies were identified. A concept called "bag" was introduced in this framework to combine a number of post-coordinated concepts. The bag could then be used to establish a mapping between a compound concept and a number of individual concepts. In this context, when users click a compound concept, a number of mapped post-coordinated

concepts within a bag could be returned so that the users could select some of these mapped concepts as search terms, or use Boolean operators to combine some of them together to conduct a search. Five types of mapping relationships were defined. These consisted of exact match, broad match, narrow match, major match, and minor match. Among these mapping relationships, the "major match" was used to establish the mapping between a compound concept and a bag of several relevant post-coordinated concepts.

From a technical perspective, it was proposed in this framework to develop a knowledge base, which aimed to record the linkings between the common functions provided by different terminology services, translate the users' query into different forms that different terminology resources can accept, and convert the returned results from different terminology resources into a consistent format. In this context, this framework would be able to access different terminology resources using different formats, protocols, and query languages.

In order to finally retrieve item-level metadata records across different resources, the middleware framework between terminologies was proposed to interact with the meta-search engine provided by a library portal system and the collection registry. In this context, the collection registry is responsible for recognising the provenance of the mapped terms returned, and then distributing the mapped terms to search against the collections using these mapped terms through the meta-search engine. A basic set of metadata elements were defined for the collection registry in this framework.

Objective 4: To develop a simplified programmatic prototype system for the middleware framework to demonstrate the methods used in the theoretical framework:

Based on the description of the theoretical framework, a simplified programmatic prototype system was designed using JAVA to accomplish this objective. In this prototype system, three vocabularies were mapped together through DDC spine. These

consist of ACM Computing Classification, UKAT, and an Information Science Taxonomy developed by Hawkins. In these three vocabularies, the Information Science Taxonomy was imagined as a local taxonomy that could be accepted by most local users.

The mapping data was encoded by the SKOS Mapping Vocabulary. An RDF bag, which was based on the idea of combining a number of post-coordinated terms together to a DDC compound concept, was applied. The SPARQL-compatible ARQ API was employed to manipulate the mapping data. In order to access various terminology resources, a knowledge base was implemented to use SPARQL to search the UKAT data in SKOS, and a DOM API was used to search the ACM data in XML format. A collection registry was also developed based on using a RDF database to record the details of the databases using UKAT, ACM or DDC.

Based on this prototype design, it was found that:

1. It is easy to use the defined mapping methods to establish the semantic mappings between the three vocabularies through DDC, although the indirection problem of using the switch language was not solved;
2. Through using a subject browsing interface, users do not need to input text-based queries, and users only need to click a number of pre-arranged term labels, and then find the relevant information. Thus, the use of SPARQL is appropriate to facilitate subject cross-browsing services, and query expansions;
3. It was proven that the developed knowledge base was able to send different queries to different terminology resources, and convert different results into a consistent format.

However, a number of issues were identified during the development of this prototype. They needed to be further addressed, as follows:

1. It is important to further investigate ways to present a returned concept to the users. A good presentation of concepts could help users disambiguate the concept.
2. The indirection problem was still not solved. When the users interact with the local taxonomy, there would be a long process that requires a fairly high degree of interaction from users to progress from the stage where they browse the hierarchy

to getting the relevant mapped terms from various KOS through DDC spine. It is important to explore the methods to enhance the interactivity between the users and the middleware system.

With this mind, it was necessary to conduct an evaluation to explore the methods to solve the existing problem, and further test the effectiveness, applicability, and viability of the framework.

Objective 5: To evaluate the prototype for improving the effectiveness and usability of the framework:

This objective was accomplished by asking a number of experts to walk through the prototype system with the assistance of a framework introduction document, and collecting the feedback information from the experts. According to the feedback, a number of problems with their solutions were discovered, and some unsolved problems have been further addressed. These problems and solutions mainly include:

1. There are various advantages and disadvantages of mapping to a DDC spine. Because of the poor notational synthesis capability, in future, it would be useful to develop more advanced faceted vocabularies as switch languages instead of DDC. However, in the current state, DDC is still suitable to be a common switch language.
2. Poor presentation of different returned concepts were identified during the evaluation. It was suggested to use some standard ways to present the concepts from different types of vocabularies. For example, it was suggested to present sibling terms, broader term, narrower terms, related terms, and alternative terms of the mapped thesaurus concept returned, and present the different notes, cross-reference, notation, and caption of a given classification concept. In addition, because most users may not be concerned with the relationships between different concepts, or which vocabulary a mapped concept is from, it is important to present "Do you mean" plus the mapped terms from different vocabularies to the users.

3. The indirection problem introduced by the use of switch language was further addressed. One solution focused on creating direct mappings between the local taxonomy and other KOS based on the use of the existing mappings between DDC and local taxonomy, and between DDC and other KOS. Another solution focused on developing the semantic expansion algorithms to expand the mapped terms and show the users a neighbourhood of concepts considered semantically close for retrieval purposes. Importantly, it was found that the semantic expansion algorithms could not only be used for expanding the returned terms for presenting more subject results, but also could be applied to assist in creating the mappings between the local taxonomy and other KOS by presenting the users with more terms semantically close.
4. More use scenarios were identified based on the findings of the evaluation. The findings included using the local taxonomy as an information discovery tool to index different databases on the collection level, so that through the local taxonomy, users could not only find the relevant item-level metadata records, but also find the collection-level metadata to describe different online databases. Likewise, it was found that the local taxonomy could be used within the KOS registry to index different KOS on the top level, and the shallow mappings could be established to make the local taxonomy a starting point to help users begin their subject navigation, and guide the users to jump to another concept scheme in another user interface.
5. A number of mapping errors were identified and corrected based on the findings of evaluation.

Based on the findings of the evaluation, a number of functions with relevant use scenarios for the development of this middleware framework between terminologies were formulated and discussed in Section 8.4.2 and 8.4.3. These functions could be used by different library portal systems to comprehensively collect the terminological information from distributed terminology resources using different functions, protocols, encoding formats, and query languages. The core part of this middleware is the knowledge base, which is used to map these defined functions to different functions or query languages

compatible with other terminology resources. An example to map the HILT functions to the proposed functions of this middleware framework was provided to indicate the applicability of the defined functions of this middleware.

9.3 Recommendations

Based on the objectives of this research, several recommendations are presented as follows:

Recommendation 1—The structural model

There are various advantages and disadvantages to mapping different vocabularies to a DDC spine. A number of terminology mapping projects (HILT, Renardus, OCLC TS, etc.) through their histories are committed to this approach. Given the nature of many of other academic databases which include DDC data, this still makes sense in the medium term. However, with the further development of faceted classification schemes with great notational synthesis capability, such as BLISS, BSO, etc., these faceted classifications might be a better option than DDC. Thus, it is recommended to further explore a switch language with great notational synthesis capability by employing advanced faceted classification theories, and explore methods to encode these classifications in semantic web-enabled formats to improve the reusability of these classifications.

Recommendation 2—An approach to improving the consistency of the mappings *from different terminology resources*

A variety of mapping initiatives have been proposed and developed based on different mapping strategies. The mappings from different initiatives may vary from different features, such as their provenances (source), methods (intellectual, co-occurrence, other automatic, etc), subject indicators, encoding formats, and the services that the mappings are located. For this reason, it is recommended to develop a metadata application profile to characterise these features, and ideally a centralised team with mapping expertise should be formed to investigate these different characteristics of the mappings. They should focus their intellectual effort on enhancing the consistency and quality of the mapping data from different sources.

Recommendation 3—The use of local taxonomies and query expansion

Based on the findings, it is recommended that a library portal developer should apply a locally-used KOS structure as a subject cross-browsing structure. This local taxonomy should be tailored to the local subject needs. Also, it is recommended that library portal developers should be responsible for establishing mappings between this local KOS and the DDC spine. Because there are a number of mappings that have been established between DDC and other KOS, the terminological information could then be exchanged between the local KOS and different other KOS through DDC. In order to solve the indirection problem introduced by the use of DDC switch language, it is recommended to employ a query expansion algorithm, because the query expansion could be helpful in improving the recall and user interaction. When establishing the mappings, it is recommended to use the query expansion to expand the mapped terms to help the mapping staff find more relevant terms as the mapping options. When users interact with the subject cross-browsing interface, it is recommended to use query expansion to provide the users more subject terms considered semantically close.

Recommendation 4—The presentation of various concepts returned to the users

There are a variety of standard elements used to characterise concepts from different types of vocabularies. For example, broader term, preferred term, narrower terms, related terms, non-preferred terms, and so on, are usually used to describe a thesaurus concept, and different notes, cross-reference, notation, caption, and so on, are usually applied to describe a classification concept. For this reason, it is recommended to employ these standard elements to present a concept with its neighboroughhood of concepts for the users. In this way, the users could understand the context of a given concept. In parallel, with the dominance of Google as a search medium, students and staff now have increasingly impoverished search skills. Thus, it is recommended to present Goolge-based "Do you mean" plus the mapped terms from different vocabularies to the users. Finally, it is recommended that these two ways to present concepts (using standard KOS elements and Goolge-based "Do you mean" plus the mapped terms) should be simultaneously used. Thus, different users can get different aspects of a mapped concept.

Recommendation 5—Technical architecture

Because there have been a number of existing terminology mapping services that use different representation formats, access protocols, API functions, and query languages, it is recommended to establish linkings between the functions provided by different terminology services, and develop programmatic interfaces to access many distributed terminology resources. Furthermore, it is recommended to develop a knowledge base, which is able to record the linkings between the functions provided by different terminology services, translate the users' query into different forms of the queries that different terminology services can accept, and convert different results into a consistent format.

Recommendation 6—mapping relationships

It is recommended to establish mappings based on the use of four types of basic mapping relationships. These include exact match, narrow match, broad match, and related match. Other types of mapping relationships could be adapted from these four types. Mapping relationships, which are used to represent co-occurrence mappings, need to be developed in future research. Using a bag to map a number of individual concepts against a relevant compound concept should be considered in other terminology mapping projects.

Recommendation 7—Using the subject browsing in different ways

Based on the findings, it is recommended to use the subject cross-browsing structure to index different databases at the collection level. Thus, the browsing structure could be used as an information resource discovery tool. In another case, it is recommended to use the subject cross-browsing structure to link to the top level concepts of different KOS. In this situation, the subject cross-browsing structure can be a starting point to help users begin their subject navigation, and help users to jump to another concept scheme in another user interface.

9.4 Contribution to knowledge

This thesis has made an original contribution to knowledge in a number of areas. These are discussed below.

9.4.1 Decentralised model to access distributed terminology resources

A contribution of this research, which is different from most terminology service projects, focuses on developing programmatic interfaces to cross-search different established terminology resources, and then re-using the terminological information within these terminology resources.

In order to facilitate subject access in heterogeneously indexed collections, a number of terminology services, such as HILT Project, OCLC TS Project, and STAR Project, have been developed. These projects focus on using appropriate methods to hold a number of controlled vocabularies in a centralised database, create mappings between these vocabularies, and provide relevant functions to let other services access the terminological data within the centralised database. However, it was found that in fact, only a limited number of controlled vocabularies were stored in a single terminology service, and a subject cross-browsing service may require more controlled vocabularies and their mappings to facilitate subject access in heterogeneously indexed collections.

Thus, the novelty of this research is based on developing a decentralised model to provide programmatic interfaces to cross-access distributed terminology services on the Web. In addition, in many cases, a subject cross-browsing service may require integrating other vocabularies and mappings excluded by these terminology services, because these vocabularies and mappings may be widely-used to index the collections covered by the subject cross-browsing service. For this reason, instead of developing a new centralised terminology service to provide the terminology data centrally, the intention of this research was to explore appropriate methods to develop a decentralised model that provides programmatic interfaces to access a variety of distributed terminology data resources on the Web.

These terminology resources, which should be integrated into a subject cross-browsing service, include:

1. Terminology services, such as OCLC Terminology Service, HILT Terminology Service, etc.—which were developed as shared services that allow other services to access their terminological data;
2. Controlled vocabularies, which were used to index important collections, and were represented in well-defined encoding formats, and published on the Web;
3. Mapping sets between different controlled vocabularies which were represented in well-defined encoding formats, and published on the Web;
4. Local vocabularies which were used by a library portal system for local subject indexing and cataloguing.

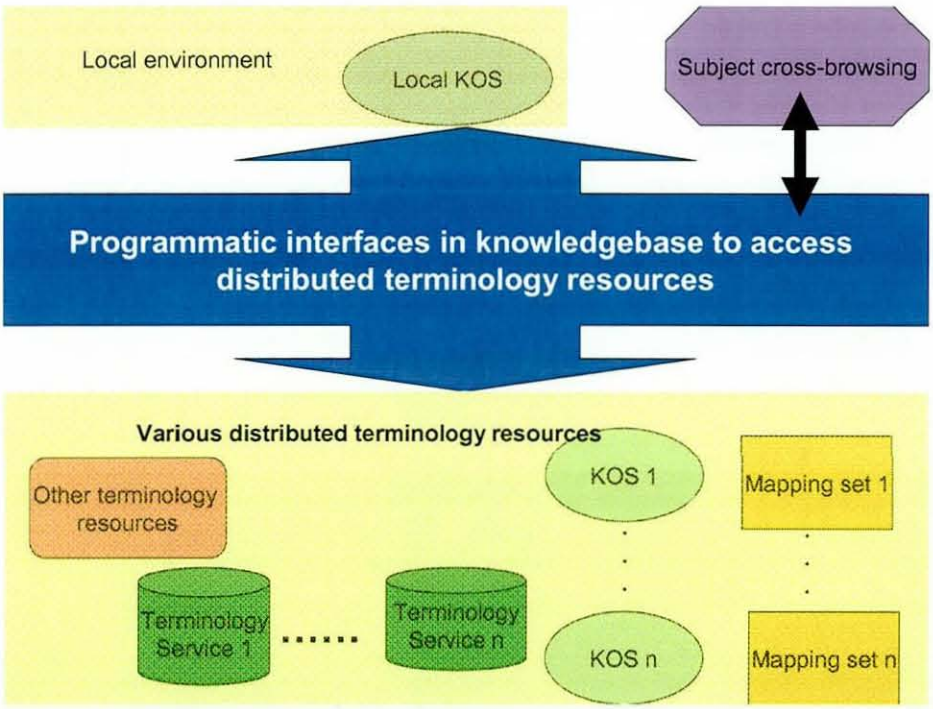


Figure 9.1: The decentralised model of this framework

Figure 9.1 shows the basic principle of this research framework. The middleware framework could behave not only as a bridge between different terminology services and different KOS, but also as a bridge between the local vocabulary environment and external terminology environments. Different terminologies could be linked through this middleware to each other in different ways. From technical perspectives, the novelty of this research is based on developing a core knowledge base to translate a user's query

into different forms that different terminology resources can recognise, and convert the returned results into a consistent format. For this reason, this knowledge base solved the technical difficulty in cross-accessing different terminology resources using different formats, access protocols, database structures, etc.

From the perspective of the semantic web, this research framework employed a DDC-based switch language, to which other vocabularies could be mapped. In other words, DDC is a key component for linking different KOS from different terminology resources. Because there is much DDC-based mapping work that has been done by different terminology service projects (e.g. OCLC TS, HILT TS, etc.), it is possible that this middleware framework can query these terminology services through using the relevant DDC concept identifiers, such as DDC notations, URIs, etc., as the basis for exchanging the terminological information between terminology services. In this sense, different established DDC-based mappings could be re-used by this middleware framework.

9.4.2 Local vocabulary for subject cross-browsing

Unlike other research projects using DDC as a switch language and browsing structure (e.g. Renardus, HILT, etc.), the novelty of this research framework focuses on using DDC as a switch language, but using other local vocabularies as browsing structures. This is because most library portal users might prefer to use their own local vocabularies as a starting point for subject cross-browsing.

It was found that using DDC as a switch language to exchange terminological information between different KOS and a local vocabulary might lead to loss of precision, and this problem has not been addressed by most terminology projects. Therefore, another main contribution of this research is based on developing a number of methods to solve this problem. These methods are listed below:

1. Creating shallow direct mappings (also called one-to-one mappings) between a local vocabulary and other vocabularies. In this context, a user could begin to browse his/her local vocabulary, and then directly jump to top levels of other vocabularies.

2. Applying or developing relevant query expansion algorithms to semantically expand mapped terms to return a range of terms considered semantically close. It was found that query expansion algorithms could improve the recall of information retrieval.
3. Designing appropriate term disambiguation applications to present the mapped terms to end-users. A term disambiguation application was intended to enable different concepts in different types of vocabularies to be presented in different ways. Users could understand the real meaning of a mapped concept, and browse the semantically close concept of the mapped concept. In this context, users may be able to re-formulate the query by selecting other terms semantically surrounding the mapped terms returned.
4. Developing Google-styled "do you mean" plus mapped terms to present all possible combinations of mapped concepts in a vocabulary to the user, and letting users select appropriate terms for further searching. Furthermore, because query expansion algorithms can also produce a number of terms considered semantically close, it is possible to add these terms to the Google-styled "do you mean" guidance. In this sense, an end-user could have more options to re-formulate his/her subject search.

Although in themselves these methods are not novel, this research has shown that they can be used in different ways to reduce the potential loss of precision caused by the use of a switch language for mapping, and be combined in comprehensive ways using machine intelligence (query expansion) to provide the users with more subject options, and subsequently encourage users to use their own human intelligence to re-formulate their subject searches. In other words, the use of query expansion algorithm to expand the mapped terms returned offers the possibility that the mappings can be improved by users' self-reflection and engagement.

Many terminology mapping projects, such as HILT, Renardus, OCLC TS, etc., depend too much on generating manually established terminology mappings to harmonise the heterogeneity between different KOS. Users' engagement to improve precision was

neglected by these projects. In this research framework, a number of user-centred interactive processes with the assistance of query expansion were designed to stimulate users to continually consider the most accurate term from a number of optional subject terms for their own subject searches.

9.4.3 Machine to Machine interaction with meta-search engines

Unlike most subject cross-browsing services that can only guide users to find relevant collections and mapped terms, one of the novel features in this research is based on developing a subject cross-browsing service to guide users to find item-level metadata results from distributed information collections. This requires this terminology mapping middleware to interact with a federated search service and a collection registry.

In many cases, users may not be willing to browse more than one vocabulary. They may only prefer to browse only their own local vocabulary and then get a range of item-level metadata records returned from distributed information resources indexed according to different vocabularies. In other projects (e.g. Renardus Project, HILT, etc.), the established mappings cannot guide the users to find relevant item-level metadata records, instead they direct the users to leave the user interface of this middleware system, and jump to browse another interface of a local collection. In this context, users may have to separately interact with many different systems that use different browsing structures, and switch their mental models between these browsing structures.

Thus, in this research, the middleware system is intended to interact with a meta-search engine in a machine-to-machine fashion, and a number of specific scenarios were designed to facilitate the interaction between the meta-search engine and this framework. It was proposed to use a meta-search engine to send the mapped terms to different information resources, get the results from different information resources, convert all the results into a consistent format, and present the results to the users. When results are retrieved by the meta-search engine from distributed information resources, it is possible

to merge all these distributed results into a single result set, and then to use ranking and filtering algorithms to improve the precision.

In addition, a collection registry was proposed to record the usage of different vocabularies in different databases. Through the collection registry, it was therefore possible to broadcast the mapped conceptual term from different KOS to be queries to cross-search against the specific databases that were indexed by these KOS. See Figure 9.2 as an example. In this figure, the collection registry is responsible for splitting these mapped terms from different KOS to become different query terms for searching against the databases indexed by the relevant KOS.

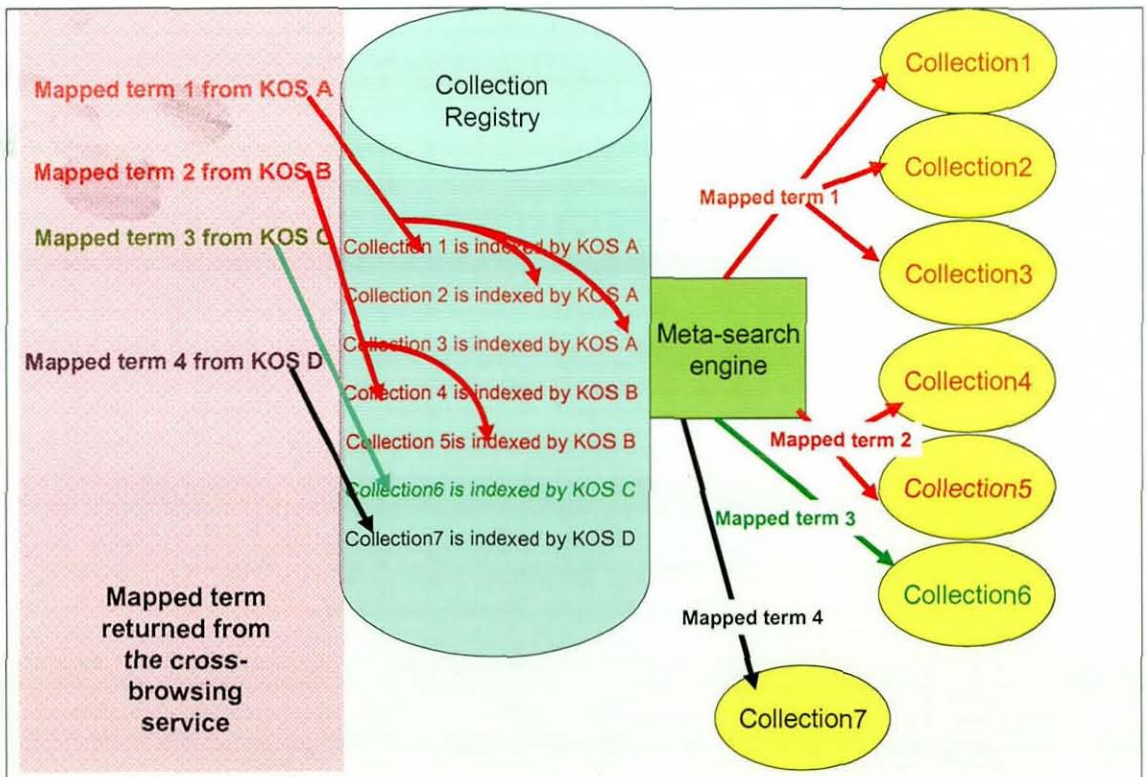


Figure 9.2: M2M interactions between different components

Based on making this research framework interact with the meta-search engine and collection registry, a user can simply click a subject term from his/her local browsing structure and then get the item-level metadata results from distributed information collections.

9.5 Research limitations

9.5.1 Research methods limitation

This research was intended to develop a middleware framework between terminologies to reuse the mappings within different terminology services, and provide subject cross-browsing interface for library portal systems to access different item-level metadata records across distributed information resources. Based on this initiative, the middleware framework needs to interact with a number of services, which include library portal services with the meta-search engines, distributed information resources, terminology services, collection registries, etc. The method to get a basic understanding of how to make the middleware system interacts with other services was based on interviewing a variety of experts, who are responsible for managing, or developing different services. Two main issues arise from this method.

First, most findings of this research were based on the experience from different experts interviewed. These experts may have their own strong biases on some approaches, or technologies for the development of this middleware framework. Thus, when analysing the collected data from these interviews, it is necessary for the researcher to minimise these bias.

Second, although a simplified prototype system was developed in this research, it is difficult for this prototype to really interact with other services to get data. This is because it is impractical in this early stage of the research to get in touch with all these different services across geographical and technical barriers, and many services may not be willing to collaborate with this immature system.

Further, the lack of end-user involvement is another issue in this research. This is because the prototype is only a middleware system between different information services in a very pilot stage. It is impossible in this early stage to use this system to get the metadata records from distributed information resources. Most users may only be concerned with

the metadata records finally retrieved, and they may not pay any attention on the subject-based interface provided by a library portal product. They may not be able to offer valuable feedback information to the researcher. Thus, in the current state, the framework is quite theoretical, and lacks user-based enhancements.

9.5.2 Limitations in the use of different vocabularies

Different KOS owners may use different policies related to their copyright management. Take DDC as an example. Based on the evaluation, it was found that if an organisation wants to map their KOS to DDC, the OCLC will allow this organisation to do that, but they do not allow this organisation to use the mappings in their service. They need to give the mappings to the OCLC, and the OCLC will sell the mappings back to them. When developing a research prototype, it is difficult to use these KOS without permission from their owners. For this reason, only four vocabularies were applied in this research, and even the applied DDC version was a very simplified version, which was adapted from a paper-based version into SKOS. This is a major developmental barrier when assessing the effectiveness and applicability of the developed theoretical framework.

9.5.3 Limitations for mapping data within different terminology services

This research framework was intended to access the mapping data from different terminology services, such as HILT, OCLC Terminology Service, STAR, etc. However, the mappings within these terminology services are incomplete. Most terminology services are also in the pilot stage and investigating different mapping strategies. No complete mapping data exists in most of these mapping services. For example, to ensure the usefulness of the mappings, the HILT Project are exploring two different strategies. In the first approach, several vocabularies were mapping to the top three levels of DDC. This is known as high level mapping. Here, a user can begin to interact with the top three levels of DDC concepts to find the mapped concepts from other vocabularies (HASSET, IPSV, and UNESCO), and then the user can go into the hierarchies of other vocabularies (HASSET, IPSV, and UNESCO), and navigate through those hierarchies to find relevant

information. The second approach is to establish deep mappings. In this approach, mappings occur at deeper levels in the hierarchy of the classification schemes. Instead of using manual mappings, in another case, OCLC applied the statistical methods to create the mappings for their terminology service.

It is difficult for this framework to access the incomplete and inconsistent mapping data from different terminology services, and use the data to provide appropriate subject cross-browsing services. For this reason, the prototype system was not designed to really interact with terminology servers. Instead, it was only used to query against a number of controlled vocabularies that have been represented in different encoding formats.

9.5.4 Limitations of the representation formats

This framework was intended to use the SKOS vocabulary to represent the mapping data between vocabularies. However, during the three-year research, the SKOS RDF vocabulary was continuing to be up-dated. The developed prototype system is not based on the latest version of the SKOS vocabulary. For example, in the old version of SKOS that was used for the development of this prototype, major match, minor match, exact match, broad match, and narrow match were utilised as the fundamental mapping relationships, but the current version of SKOS uses exact match, related match, broad match, and narrow match as the basic mapping relationships.

Furthermore, because different technologies are continuing to be developed and updated, some of the selected technologies within this framework may be not suitable for the current development. This is an issue existing in different semantic web-based services. How to integrate new ideas or technologies into an existing system is important.

9.6 New development in the field and related discussion

It is becoming more and more widely-agreed that the roadmap for the further development of semantic web services is based on using RDF data to create a semantic layer above heterogeneous information resources (Miles 2008, Tuominen *et al.* 2008, and Geser 2008). This RDF-based semantic layer can facilitate subject cross-searching and

browsing heterogeneous databases. In order to create this semantic layer, Geser 2009 indicated four directions to develop SKOS-based web services. See Figure 9.3.

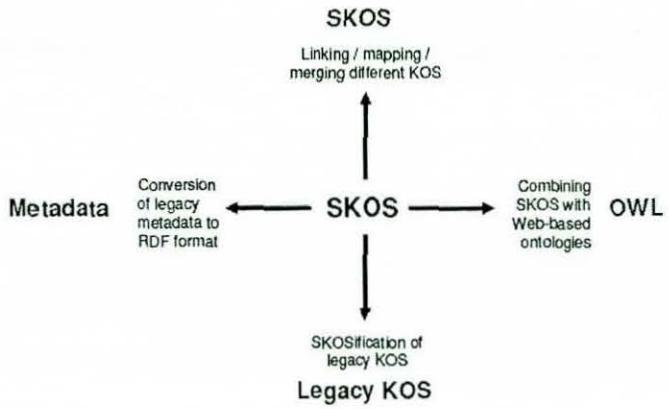


Figure 9.3: Four directions to develop SKOS-based web services (Source: Geser 2008)

In this figure, the four directions are listed below:

1. The first direction focuses on converting the legacy metadata into RDF format;
2. The second direction is based on converting various controlled vocabularies into SKOS format (Summers *et al.* 2008, and Tudhope and Binding 2008). The term "SKOSification" refers to the conversion of KOS to SKOS representation (Miles 2008).
3. Establishing semantic mappings between different KOS is the third direction. Because different methods to create mappings between different KOS were deeply discussed through this thesis, this section does not mention this issue.
4. The final direction focuses on migrating existing KOS into ontologies in OWL format.

Among these development directions, Omelayenko (2008) and Tordai *et al.* (2007) emphasised the importance of combining the RDF metadata of converted datasets with relevant "SKOSified" controlled vocabularies. Based on the combination, it is possible to

apply a number of value-added APIs above the converted RDF data. With the development of semantic technologies, more and more semantic web-based APIs are being developed to improve the functionality of RDF-based web services. Likewise, Tuominen *et al.* (2009, p.2) argues, "many common, sharable tasks in such vocabulary-aware applications related to e.g. term/concept finding, browsing, selecting, fetching, and query expansion have been developed. Lots of work and costs can be saved by implementing such functionalities in standard ways and by providing them for production use as ready-to-use services without having to reimplement the functionalities separately for each local application case".

In STERNA (Semantic Web-based Thematic European Reference Network Application) Project[6], for example, when distributed metadata records were converted into RDF, and a number of KOS were "SKOSified", a number of well-developed APIs were applied to manipulate the converted data and KOS, and enrich the service functionality (Nederbragt 2008). These APIs were shown in Figure 9.4.

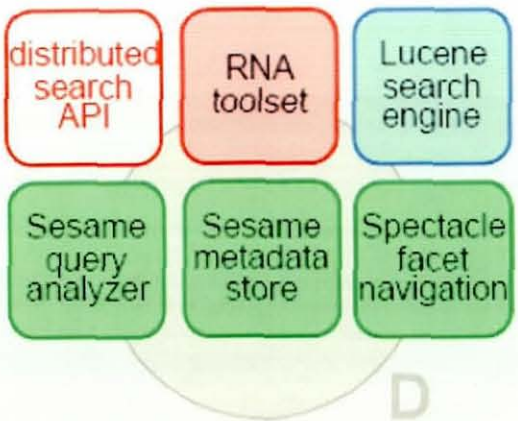


Figure 9.4: Applications to enrich the use of the converted RDF data and "SKOSified" KOS (Source: Geser 2008)

[6] Sterna Project is aimed to create a dispersed and networked information space, supported and sustained by a member network of autonomous content organisations which serves users with a special interest in nature and wildlife worldwide

In Figure 9.4, the specific explanations of different applications are listed below:

- Sesame query analyser is used for intelligent distribution of queries to different converted RDF resources.
- Spectacle facet navigation provides advanced search and navigation facilities on top of large RDF information sources, using Sesame's metadata storage and querying facilities as the main middleware component[7].
- Lucene search engine supports high-performance, full-featured text search and semantic web search[8].
- Distributed search API is used to support incorporating intelligent search functionalities in websites.
- RNA toolset is used to edit and maintain the metadata and reference structure.
- Sesame metadata store is used to store and query RDF metadata.

In addition, the methods to migrating SKOS data with OWL-based ontologies was explored (Vatant 2008, Nußbaumer and Haslhofer 2007, and Tudhope and Binding 2008). Two barriers that impede the integration of SKOS with OWL ontologies were realised by Nußbaumer and Haslhofer (2007). The first barrier is the ambiguity of the concepts within some ontologies. It was found that most ontology concepts were very abstract, and it is difficult for people to disambiguate these concepts. The second barrier is that most upper ontologies are lack of formal instructions of how to conduct migrating from SKOS data. For this reason, it is possible that different organisations may have heterogeneous interpretations on a standardised global ontology. These heterogeneous interpretations may further cause the semantic interoperability problems between different information services. Also, Tudhope and Binding (2008) argues that the purpose of migrating SKOS data into an OWL-based ontology is to facilitate automatic inferencing, and in many cases, the migrating process is quite resource-intensive and costly. It is important to consider the rationale. Tudhope and Binding (2008a) pointed out that "SKOS representation offers a cost-effective approach for annotation, search and browsing

[7] http://www.aduna-software.com

[8] http://lucene.apache.org/solr

oriented applications that don't require first order logic". Thus, before planning to migrate SKOS data into an OWL-based ontology, it is important to clarify the real purpose of the migration, and consider the cost-effectiveness of the migration.

It is true that RDF data would become more and more widely-used in different communities. However, in some particular projects, other data formats are still needed. For example, in library communities, MARC21 XML has been used for a long history. It can represent most library data in a very accurate and cost-effective way. It is still necessary to convert the returned terminological records into different formats, such as MARC21 XML, Zthes, SKOS, etc. In this research, as mentioned in Section 5.2, different data conversion programmes were proposed to be set up in the knowledge base. The knowledge base can potentially become a data converter that converts different terminological resources into the formats that different clients need, and converted data can be aggregated and stored into a new database system as a new terminology service. Figure 9.5, for example, indicates that through this framework, different terminological resources are "SKOSified", and then the "SKOSified" data are stored in a database as a terminology service for various web clients. In this example, the middleware system can be used as a "SKOSification" tool to convert different vocabularies into SKOS format.

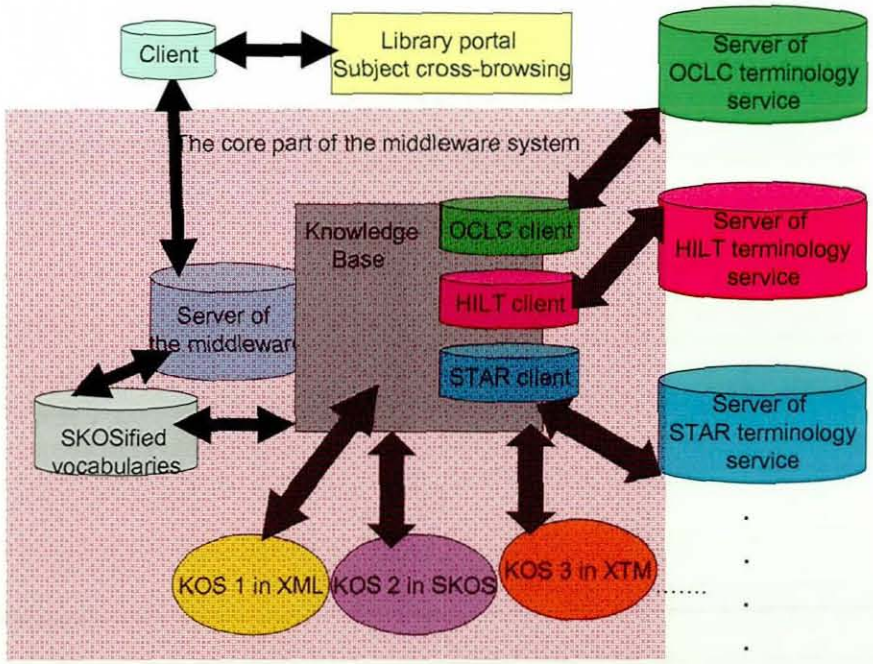


Figure 9.5: The middleware as "SKOSification" tool

In another example, it is possible that through this middleware system, different terminological resources are converted into MARC21 XML format, and then different clients use MARC21 XML data to support their subject cross-browsing. See Figure 9.6.

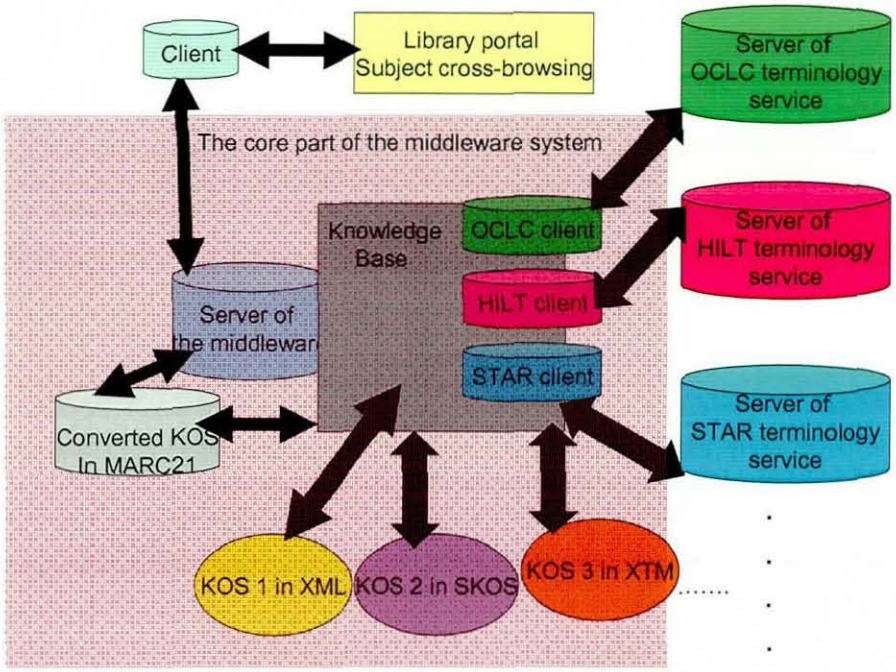


Figure 9.6: The middleware as a tool to convert data into MARC21 XML

9.7 Future research

The purpose of this section is to outline ideas and hypotheses for future research. It is suggested that the future research should be undertaken with the respect of the following aspects:

1. Due to a lack of notational synthesis capability in DDC, it is not easy to create mappings between DDC and other pre-coordinated vocabularies. Future research should focus on attempting to adopt some faceted vocabularies, such as UDC, BLISS, BSO, etc., to be the switch language instead of using DDC, and establish the mappings between different vocabularies through faceted vocabularies. It is important to explore and identify the advantages and disadvantages of using these

vocabularies. This effort will give a more realistic sense of the mapping overheads based on other switch languages.

2. Based on the findings of the evaluation, it is important to integrate a number of user-centred elements and functions into this framework. These might include developing the query expansion algorithm to improve the user interaction, using the Google-like "Do you mean" plus mapped terms, and employing clouding technology to highlight the most widely-used terms. Therefore, future research should focus on implementing all these user-centred elements and technologies into an improved prototype system, letting the real end-users interact with the system, and gaining feedback information from the end-users.
3. New technologies and initiative are always being developed. It is important to keep looking into a number of W3C initiatives for Semantic Web, such as RDF triple store technologies, the new versions of SKOS vocabulary, SPARQL Query and Protocol, SKOS API, etc, which may provide better facilitates to improve the interoperability between different services. Future research should focus on implementing appropriate new technologies into the prototype system.
4. Because the established mappings may vary from one terminology-based project to another, and the licenses for the use of different vocabularies also are very different, it is suggested in future to develop a registry service to record these basic characteristics of different terminology resources. In this way, it would be easier for a terminology services to gain a basic understanding of the differences between the characteristics of other terminology resources.
5. Because mappings could be established based on a variety of strategies, a comparison study of the strategies between different mapping projects should be undertaken to explore the ways to improve the consistency between the mappings from different terminology resources.

Bibliography

Ahmed, K., 2003, *Techquila's published subject identifiers*, http://www.techquila.com/psi/

Aitchison, J. and Bawden, D., 2000, *Thesaurus construction and use: a practical manual*, London: Aslib IMI.

Alonso, S. and Barriocanal, E., 2006, Making use of upper level ontologies to foster interoperability between SKOS concept schemes, *Online information review*, 30(3), 263-277. <http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2640300305.html > [accessed 02.02.08]

Arms, W., *et al.*. 2002, A spectrum of interoperability, *D-LIB magazine*, 8(1), <http://www.dlib.org/dlib/january02/arms/01arms.html> [accessed 10.10.08].

Arroyo,S., Bruijn, J. and Ding, Y., 2001, Ontology mapping and aligning. < http://www.deri.at/fileadmin/documents/deliverables/Esperonto/Del14-v01.1.doc. > [accessed 10.09.07]

Audience Research Basics [n.d.], http://www.health.gov/communication/primer/aud_res_prim.asp#qlro

Bacon, F., 1640, *The advance of learning,* Oxford : Young and Forrest.

Baeza-Yates, R., and Berthier, A., 1999, *Modern information retrieval*, Harlow: Addison-Wesley.

Bailey, K.D., 1982, *Methods of social research*, London: The Free Press.

Bailey, B., 2005, *Paper prototypes work as well as software prototypes*, <http://www.usability.gov/pubs/062005news.html> [accessed 09.11.08]

Bater, B., 2003, Topic maps, ***In***: Gilchrist, A. and Mahon, B., eds. *Information,* 2nd. London: Facet Publishing, pp. 132-145

Beghtol, C., 2004, James Duff Brown's subject classification and evaluation methods for classification systems, *Proceedings 15th workshop of the American society for information science and technology special interest group in classification research*, < http://goliath.ecnext.com/coms2/gi_0199-3408941/Exploring-new-approaches-to-the.html > [accessed 10.08.08]

Bergman, M.K., 2001, The deep web: surfacing hidden value. *Journal of electronic publishing,* 17(8), <www.press.umich.edu/jep/07-01/bergman.html> [accessed 19.05.05].

Berners-Lee, T., *et al.*., 2001, *The semantic web,* <www.sciam.com/2001/0501issue/0501berners-lee.html> [accessed 10.03.05]

Berners-Lee, T., 2003, Foreword, ***In***: Fensel, D. *et al.*, **eds**. *Spinning the semantic web*, London: The MIT Press, pp.xi-xvi.

Binding, C., 2008, Semantic technology for archaeology resources, *ISKO UK meeting*, 21st July, London, <http://www.iskouk.org/presentations/STAR_UCL_20080721a.pdf> [accessed 02.11.08]

Binding C., and Tudhope D., 2004, Integrating faceted structure into the search process, *Proceedings 8th international society of knowledge organization conference (ISKO 2004)*, 9, 67–72.

Binding C., and Tudhope D., 20 04a, KOS at your Service: Programmatic access to knowledge organisation systems, *Journal of digital information*, 4, <http://journals.tdl.org/jodi/article/view/jodi-124/109> [accessed 10.09.07]

Blocks, D., 2004, *A qualitative analysis of thesaurus integration for end-user searching. Ph.D. thesis,* Hypermedia research unit, school of computing, University of Glamorgan, Pontypridd, UK,
< http://www.comp.glam.ac.uk/~FACET/dblocks/DBlocks_ThesisOnline_Main.html > [accessed 09.09.08]

Blois, M., Escobar, M., and Choren, R., 2007, Using agents and ontologies for application development on the semantic web, *Journal of the Brazilian computer society*, 13(2), 35-44.

Borgan, M., 2003, A survey of digital library aggregation services,
< http://www.diglib.org/pubs/dlf101/dlf101.htm > [accessed 09.09.08]

Boss, R.W. 2003, Library portal solutions, *Aslib proceedings*, 55(3), 155-165,
<http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2760550304.html > [accessed 10.08.08]

Boye, J. 2006, *The enterprise portals report*, CMS watch, Olney,
<http://www.cmswatch.com/Portal/Report/> [accessed 09.09.08]

Brickley, D., and Guha, R., 2004, RDF vocabulary description language 1.0: RDF schema,
<http://www.w3.org/TR/rdf-schema/> [accessed 10.07.06]

Broekstra, J. *et al.*, 2003, Enabling knowledge representation on the web by extending RDF schema, *Proceedings of the tenth international World Wide Web conference*

(WWW10), Hong Kong, <http://www.cs.man.ac.uk/~horrocks/Publications/download/2001/extending-RDF.pdf> [accessed 09.10.08].

Broughton, V., 2001, Faceted classification as a basis for knowledge organization in a digital environment; the Bliss bibliographic classification and the creation of multi-dimensional knowledge structures, *New review of hypermedia and multimedia*, Vol. 7, pp. 67-102.

Broughton, V., 2006, *Essential thesaurus construction*, London: Facet.

Broughton, V. and Slavic, A., 2006, Building a faceted classification for the humanities: principles and procedures, *Journal of documentation*, 63(5), 727-754, <http://dlist.sir.arizona.edu/1976/> [accessed 10.09.08]

Brown, J.,J. D.,1939, Subject classification for the arrangement of libraries and the organization of information, with tables, indexes, *et al.*., for the subdivision of subjects, London: Grafton.

BS 8723-1; BS 8723-2, 2005, *Structured vocabularies for information retrieval, Part 1: definitions, symbols and abbreviation; Part 2: thesauri*, London: British Standards Institution.

BS 8723-3, 2008, *Structured vocabularies for information retrieval, Part 3:Vocabularies other than thesauri*, London: British Standards Institution.

BS8723-Part4, 2008, *Structured vocabularies for information retrieval, Part 4:Interoperability between vocablaries*, London: British Standards Institution.

Buckland, M., 1999, Mapping entry vocabulary to unfamiliar metadata vocabularies. *D-LIB Magazine*, 5(1),

<http://www.dlib.org/dlib/january99/buckland/01buckland.htm > [accessed 11.01.05].

Butters, G. 2003, What features in a portal, *Ariadne,* 35,
<http://www.ariadne.ac.uk/issue35/butters/> [accessed 02.02.07]

Chan, L. M., 1999, *A guide to the Library of Congress Classification,* Colo: Libraries Unlimited.

Chan, L., 2000, Exploiting LCSH, LCC, and DDC to retrieve networked resources issues and challenges, *Conference on bibliographic control in the new millennium,*
<http://www.loc.gov/catdir/bibcontrol/chan_paper.html > [accessed 12.12. 04] .

Chan, L. and Zeng, M., 2002, Ensuring interoperability among subject vocabularies and knowledge organization schemes: a methodological analysis, *Proceedings 68th IFLA Council and General Conference,*
< http://dlist.sir.arizona.edu/475/ > [accessed 12.12.08]

Chan, L.M. and Zeng, M., 2004, Trends and issues in establishing interoperability among knowledge organization systems, *Journal of the American society for information science and technology,* 55(5), 377-395,
< http://portal.acm.org/citation.cfm?id=986354.986356 > [accessed 09.07.06]

Chan, L. and Zeng, M., 2006a, Metadata interoperability and standardization – A study of methodology Part I, *D-LIB magazine*, 12(6),
<http://www.dlib.org/dlib/june06/chan/06chan.html> [accessed 10.03.08]

Chan, L. and Zeng, M., 2006b, Metadata interoperability and standardization – A study of methodology Part II, *D-LIB magazine*, 12(6).
<http://www.dlib.org/dlib/june06/zeng/06zeng.html> [accessed 10.03.08]

Chaplan, M.A., 1995, Mapping laborline thesaurus terms to Library of Congress Subject

Headings: implications for vocabulary switching, *Library Quarterly,* 65(1), 39-61.

Chebotko, A., Lu, S., Jamil, H., and Fotouhi, F., 2006, *Semantics preserving SPARQL-to-SQL query translation for optional graph patterns,* <http://www.cs.wayne.edu/~artem/main/research/TR-DB-052006-CLJF.pdf> [accessed 09.09.08]

Chen, X. 2006, MetaLib, WebFeat, and Google, *Online information review,* 30(4), 413-427,
<http://www.emeraldinsight.com/Insight/viewContentItem.do?contentType=Articleandcontentld=1570033 > [accessed 20.11.08]

Clark, K., 2005, *SPARQL protocol for RDF,*
<http://www.w3.org/TR/2005/WD-rdf-sparql-protocol-20050527/> [accessed 20.09.07]

Clarke, S.D., Gilchrist, A., Davies, R. and Will, L., 2005, Glossary of terms relating to thesauri and other forms of structured vocabulary for information retrieval, *Willpower information,*
< www.willpowerinfo.co.uk/glossary.htm> [accessed 09.09.07]

Cliff, P., 2001. Building resource finder. *Ariadne,* 30.
<http://www.ariadne.ac.uk/issue30/rdn-oai/intro.html > [accessed 22.02.05].

Coates, E., Lloyd, G., and Simandl, D., 1978, *BSO: Broad System of Ordering: schedule and index*, Paris:Unesco.

Cordeiro, M., 2003, Knowledge organization from libraries to the web: Strong demands on the weakest side of international librarianship, *Cataloging and classification quarterly,* 37(1/2), 65-79.

Coombes, H., 2001, *Research using IT*, New York: Palgrave.

Cox, A. and Yeates, R. 2003, Library portal solution, *Aslib Proceedings,* vol. 55, no. 3, pp. 155-165.

Creswell, J.W., 2007, *Qualitative inquiry and research design: cho osing among five traditions*, London: Sage.

Cunliffe, D., Taylor, C., and Tudhope, D., 1997, Query-based navigation in semantically indexed hypermedia, *Proceedings 8th ACM conference on hypertext (Hypertext'97)*, Southampton, 87–95, <http://portal.acm.org.ergo.glam.ac.uk/citation.cfm?doid=267437.267447 > [accessed 09.05.07]

Dahlberg, I., 1980, The Broad System of Ordering (BSO) as a basis for an integrated social science thesaurus, *International classification*, 7, pp.66–72

Damljanovic, D., Tablan, V., and Bontcheva, K.,2008, *A text-based query interface to OWL ontologies,*
<http://gate.ac.uk/sale/lrec2008/clone-ql/clone-ql-paper.pdf > [accessed 10.09.08]

Davies, J. *et al.,* 2005. Next generation knowledge access. *Journal of Knowledge Management,* 9(5), 68-84.

Davies, R. 2007, Library and institutional portals: a case study, *The electronic library*, 25(6), 641-647.
<http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2630250601.html > [accessed 20.09.08]

Dawson, A. 2004, Creating metadata that work for digital libraries and Google, *Library*

review, 53(7), 347-350.

Day, M., 2001, *Renardus DDC classification mapping - a summary of work in progress*, <http://hilt.cdlr.strath.ac.uk/Dissemination/Workshop%20documents/Renardus%20Report.doc.> [accessed 11.11.05].

Day, M., 2003, Prospects for institutional e-print repositories in the United Kingdom, *ePrints UK supporting study*, 1, <online: http://www.rdn.ac.uk/projects/eprints-uk/docs/studies/impact/ > [accessed 02.08.04]

Dean, R., 2004, FAST: Development of simplified headings for metadata, *Cataloguing and classification quarterly,* 39(1/2), 331-352.

Dempsey, L. 2000, The subject gateway: experiences and issues based on the emergence of the resource discovery network, *Online information review,*24(1), 8-23.

Delphi Group, 2002, Taxonomy and content classification: market milestone report, <http://www.verity.com/pdf/Delphi_class_mktstudy.pdf > [accessed 28.05.05]

Ding, Y., 2001, A review of ontologies with semantic web in view, *Journal of information science*, 27(6), 377-384.

Doerr, M. 2001, Semantic problems of thesaurus mapping, *Journal of digital information,* 1(8), < http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr> [accessed 10.09.06].

Doerr, M., Hunter, J., and Lagoze, C., 2003, Towards a core ontology for information integration, *Journal of digital information*, 4(1), <http://www.cs.cornell.edu/lagoze/papers/core_ontology.pdf.> [accessed 10.01.08]

Efthimiadis, E., 1996, Query expansion, *ARIST*, 31, 121-187. <http://faculty.washington.edu/efthimis/pubs/Pubs/qe-arist/QE-arist.html> [accessed 10.10.08]

European Library Automation Group, 2002, *Report of the portal workshop of the European library automation group meeting,* <www.ifnet.it/elag2002/ws_paper/ws2_post.html> [accessed 09.08.04].

Evmorfopoulou, K., 2000, *Focus group methodology for the MADAME Project,* <http://www.shef.ac.uk/~scgisa/MADAMENew/Deliverables/FGEnd1.htm> [accessed 10.09.07]

Fang, L. 2004, A developing search service: Heterogeneous resources integration and retrieval system, *D-Lib magazine,* 10(2), < http://www.dlib.org/dlib/march04/fang/03fang.html > [acccessed 10.10.08]

Fensel, D., 2001, *Ontologies: silver bullet for knowledge management and electronic commerce,* <http://www.sigmod.org/sigmod/record/issues/0212/B1.review.pdf>[accessed 07.01.06]

Fensel, D.. 2003, Introduction, ***In:*** Fensel, D. *et al.*, **eds**. *Spinning the semantic web*, London: The MIT Press, pp.1-8.

Foskett, A. C., 1996, *The subject approach to information,* London: Bingley.

Garshol, L.M., 2005, *Metadata? thesauri? taxonomy? topic map!,* <http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html>[12.12.05].

Geser, G., 2008, *Sterna Project: Technology watch report,* < http://www.tdwg.org/homepage-news-item/article/sterna-technology-watch-report/ > [accessed 01.03.09]

Glazier, J., 1992, Qualitative research methodologies for library and information science: An introduction, ***In***: Glazier, J. and Powell, R., eds. *Qualitative research in information management*, Englewood: Library Unlimited, INC, pp.1-11

Glaser, G. and Strauss, A., 1976, *The discovery of grounded theory,* Chicago: Aldine.

Gnoli, C., 2004, Is there a role for traditional knowledge organization systems in the Digital Age?, *The Barrington report on advanced knowledge organization and retrieval (BRAKOR),* 1(1),
<http://eprints.rclis.org/archive/00001415/01/kos-role.htm.> [accessed 10.09.08]

Gnoli, C., 2005, BC2 classes for phenomena: an application of the theory of integrative levels, *The Bliss classification bulletin,* 47, 17-21,
<http://dlist.sir.arizona.edu/920/> [accessed 10.08.08].

Goldhor, H., 1972, *An introduction to scientific research in librarianship.* Urbana, IL: University of Illinios, Graduate Library School.

Goldman, A. E. and McDonald, S. S., 1987, *The group depth interview: principles and practice.* NJ: Prentice Hall

Golub, K., 2007, Controlled vocabulary based approach to automated subject classification of textual web pages, *Application for doctoral forum at 1st IIiX symposium,*
<http://www.it.lth.se/knowlib/publ/WebGolubPhDForum06.pdf> [accessed 09.09.08]

Golub, K. and Tudhope, D., 2008, *Terminology registry scoping study (TRSS): Excerpt on metadata,*
< http://www.ukoln.ac.uk/projects/trss/dissemination/metadata.pdf > [accessed 09. 12.08].

Golub, K., and Tudhope, D., 2008a, TRSS: Terminology registry scoping study, *ECDL*

NKOS workshop,
<http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2008/presentations/GolubTudhope-NKOS.ppt > [accessed 09.11.08]

Gomez-Perez, A., and Benjamins, V., 1999, Overview of knowledge sharing and reuse components: ontologies and problem-solving methods, *Proceedings of IJCAI-99 workshop on ontologies and problem-solving methods: Lessons learned and future trends*, <http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-18/1-gomez.pdf> [accessed 10.09.08].

Gorman, G.E., and Clayton, P. R., 1997, *Qualitative research for the information professional: a practical handbook*, London: Library Association.

Gray, W.D., and Salzman, M. C., 1998, Damaged merchandise? A review of experiments that compare usability evaluation methods, *Human-computer Interaction*, 13, 203-261. <http://pdfserve.informaworld.com/738631_775646432_784767073.pdf> [accessed 01.01.08]

Green, D., 2000, The evolution of Web searching, *Online information review,* 24(2), 121-137.

Greenberg, J., 2001, Optimal query expansion (QE) processing methods with semantically encoded structured thesauri terminology, *Journal of the American society for information science and technology*, 52(6):487–498.

Greenberg, J., 2003, Metadata and the World Wide Web, *Encyclopaedia of library and information science*, Dekker, M. (ed.).

Greenberg, J., 2005, Understanding metadata and metadata schemes, *Cataloging and classification quarterly,* 40(3/4), 17-36.

Grover, R., and Glazier, J. D., 1985, Implications for application of qualitative methods to library and information science research, *Library and information science research*, 7(3),
<http://cat.inist.fr/?aModele=afficheNandcpsidt=8502553> [accessed 10.09.08]

Gruber, T. R., 1993, A translation approach to portable ontology specifications, *knowledge acquisition,* 5, pp.199–220.

Guarino, N., 1997, Understanding, building and using ontologies: a commentary to 'using explicit ontologies in KBS development', *International journal of human and computer studies,*46(2/3), 293–310

Guba, E., and Lincoln, Y., 1994, Competing paradigms in qualitative research, ***In:*** Denzin, N., and Linciln, Y., eds. *Handbook of qualitative research.* London: Sage, 105-118

Guest, G., Bunce, A., and Johnson, L., 2006, How many interviews are enough? An experiment with data saturation and variability, *Field methods*, 18(1), 59-82.

Guy, M. 2005, Lessons and outcomes from the subject portals project, *The journal of information and knowledge management systems,* 36(1/2), 58-63,
<http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2870350111.html > [accessed 10.09.08]

Gysenns, M., Paredaens, J., Van den Bussche, J., and Van Gucht, D., 1994, A graph-oriented object database model. *IEEE Trans.* on KDE, 6(4), pp. 572-586.

Haya, G., Nygren, E. and Widmark, W. 2007, Metalib and Google scholar: a user study, *Online information review,* 31(3), 365-375,
<http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2640310308.html >[accessed 09.08.08]

Heery, R., Carpenter, L. and Day, M. 2001, Renardus project developments and the wider digital library context, *D-LIB magazine,* 7(4),
< http://www.dlib.org/dlib/april01/heery/04heery.html.> [accessed 09.07.03]

Heery, R. and Wagner, H., 2002, A metadata registry for semantic web, *D-LIB magazine,* 8(5).
< http://www.dlib.org/dlib/may02/wagner/05wagner.html > [accessed 31.10.05]

Heery, R. and Patel, M., 2004, App lication profiles: mixing and matching metadata schemas, *Ariadne,* 25.
< http://www.ariadne.ac.uk/issue25/app-profiles/ >[accessed 16.07.05]

Hernon, P., 1999, The elusive nature of research in LIS, ***In***: McClure, C. and Hernon P.eds. *Library and information science research: Perspectives and strategies for improvement*, United States: Ablex Publishing Corporation, pp.3-10.

Hevner, A., March, S., Park, J. and Ram, S., 2004, Design science in information systems research, *MIS quarterly*, 28(1), 75-105,
<http://www.idi.ntnu.no/emner/empse/papers/hevner_etal_2004,pdf> [accessed 09.09,08]

Hickey, T., 2000, CORC: a system for gateway creation. *Online information review,* 24(1), 49-56,
<http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2640240106.html> [accessed 09.09.08]

HILT Phase II final report, 2003,
< http://hilt.cdlr.strath.ac.uk/hilt2web/finalreport.htm> [accessed 09.08.08]

HILT Phase III final report, 2007,

<http://hilt.cdlr.strath.ac.uk/hilt3web/Reports/hiltIIIfinalreport.pdf> [accessed 09.09.08]

Hodge, G., 2000, Systems of knowledge organization for digital libraries: Beyond traditional authority files, *CLIR Pub9*,
<www.clir.org/pubs/abstract/pub91abst.html> [accessed 09.08.07]

Howarth, L., 2003, Metadata schemas for subject gateways, *69th IFLA General Conference and Council,*
< http://www.ifla.org/IV/ifla69/papers/053e-Howarth.pdf > [accessed 29.03.05]

Hunter, J., 2001, MetaNet - A metadata term thesaurus to enable semantic interoperability between metadata domains, *Journal of digital information,* 1(8),
<http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Hunter/ > [accessed 18.08.05]

Hunter, P. and Guy, M., 2004, Metadata for harvesting: the Open Archives Initiative, and how to find things on the Web, *The electronic library,* 22(2), 168-174.

Hunter, J. and Lagoze, C., 2001, *Combining RDF and XML Schemas to enhance interoperability between metadata application profiles,*
<http://archive.dstc.edu.au/RDU/staff/jane-hunter/www10/paper.html> [accessed 08.10.05].

IEEE Learning Technology Standards Committee, 2002, *Draft standard for learning object metadata,*
<http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf > [12.12.04].

Jacsó, P. 2005, Google Scholar: the pros and the cons, *Online information review,* 29(2), 208-214,
<http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2640290206.html> [accessed 09.09.08]

Johnson, E., 2004, Distributed thesaurus web services. *Cataloguing and classification quarterly,* 37(3/4), 121-153.

Joyce, A., Wickham, J., Cross, P. and Stephens, C. 2008, Intute integration, *Ariadne,*55, <http://www.ariadne.ac.uk/issue55/joyce-et-al/> [accessed 09.09.08]

Klee, M., 2000, *Five paper prototyping tips*, <http://www.uie.com/articles/prototyping_tips/> [accessed 09.08.08]

Klein, M., *et al.*, 2003, Ontologies and schema languages on the Web, ***In***: Fensel, D. *et al.*, **eds**. *Spinning the semantic web*, London: The MIT Press, pp.95-141.

Kline, V., 2002, Missing links: the quest for better search tools, *Online information review,* 26(4), 252-255, <http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2640260403.html > [accessed 09.09.08]

Koch, T. 1997, *The role of classification schemes in Internet resource description and discovery*, < http://www.ukoln.ac.uk/metadata/desire/classification/ >[accessed 10.09.08]

Koch, T. 2000, Quality-controlled subject gateways: definitions, typologies, empirical overview, *Online information review,* 24(1), 24-34, <http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2640240102.html > [accessed 09.09.08]

Koch, T., Neuroth, H., and Day, M., 2001, Renardus: Cross-browsing European subject gateways via a common classification system (DDC), *IFLA satellite meeting: subject retrieval in a networked environment,*

<http://www.ukoln.ac.uk/metadata/renardus/papers/ifla-satellite/ifla-satellite.pdf>
[accessed 09.09.08]

Koch, T., Neuroth, H. and Day, M., 2001a, *DDC mapping report*, Renardus D7.4.
<http://renardus.sub.uni-goettingen.de/wp7/d7.4> [accessed 09.09.06]

Koepsell, D., 2000, *The ontology of Cyberspace: law, philosophy, and the future of intellectual property,* Chicago: Open Court.

Krueger, R. A., and Casey, M. A., 2000, *Focus group*, 3rd edition, California: Sage Publications, Inc.

Kumar, K., 1990, Post implementation evaluation of computer-based information systems: current practices, *Communication of the ACM*, 33(2), 203-212,
<http://portal.acm.org/citation.cfm?id=75577.75585> [accessed 09.08.08]

Laborda, C., and Conrad, S., 2006, Bringing relational data into the semantic web using SPARQL and relational OWL, *22nd International conference on data engineering workshops,*
<http://csdl2.computer.org/persagen/DLAbsToc.jsp?resourcePath=/dl/proceedings/andtoc=comp/proceedings/icdew/2006/2571/00/2571toc.xmlandDOI=10.1109/ICDEW.2006,37> [accessed 09.08.08].

Lagoze, C. and Hunter, J., 2001, ABC ontology and model, *Journal of digital information*, 2(2),
<http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze/> [accessed 09.08.08]

Lancaster, F.W. and Smith, L. 1983, Compatibility issues affecting information systems and services, *General information programme and UNISIST,* Paris:UNESCO.

Legard, R. and J. Keegan, *et al.*., 2003, In-depth interview, *In:* J. Ritchie and J. Lewis, eds., *Qualitative research practice*. London, SAGE, 138-169.

LevelTen Knowledgebase, [n.d.]
<http://www.leveltendesign.com/category/tags/focus-group> [accessed 09.07.07]

Lewis, N. 2008, Implementing Ex Libris's Primo at the University of East Anglia, *Ariadne, 55,*
<http://www.ariadne.ac.uk/issue55/lewis/>[accessed 09.08.08].

Mangan, J., Lalwani, C., and Gardner, B., 2004, Combining quantitative and qualitative methodologies in logistics research, *International journal of physical distribution and logistics management*, 34(7), 563-578

Mah, C., and Stranack, K., 2005, dbWiz : open source federated searching for academic libraries, *Library Hi Tech*, 23(4), 490-503,
<http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2380230404.html > [accessed 09.08.08]

Mai, J., 2003, The future of general classification, *Cataloging and classification quarterly,* 37(1/2), 3-12

Marchionini G. 1995, Information seeking in electronic environments, *UK:CUP*,
<http://ils.unc.edu/~march/isee_book/web_page.html> [accessed 09.08.08]

Martin, R., 2003, ePrints UK: Developing a national e-prints archive, *Ariadne*, 35,
< http://www.ariadne.ac.uk/issue35/martin/>[accessed 09.09.08]

Matthews, B., 2005, Semantic web technologies,
<http://www.jisc.ac.uk/uploaded_documents/jisctsw_05_02bpdf.pdf> [accessed 09.10.05]

McCulloch, E. 2004, Multiple terminologies: an obstacle to information retrieval, *Library review,* 53(6), 297-300,
<http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/0350530601.html >[accessed 09.09.08]

McCulloch, E. and Shiri, A. and Nicholson, D. 2005, Challenges and issues in terminology mapping: a digital library perspective, *Electronic library*, 23 (6), 671-677,
<http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2630230606.html > [accessed 09.09.08]

McGuinness, D., and van Harmelen, F., 2004, *OWL web ontology language overview*,
<http://www.w3.org/TR/owl-features> [accessed 09.09.08]

Mertens, D.M., 1998, *Research methods in education and psychology: integrating diversity with quantitative and qualitative approaches*, CA: Sage.

Miles, A., 2008, SKOS in the context of semantic web deployment, *ISKO UK meeting*, London,
< http://www.iskouk.org/SKOS_July2008.htm > [accessed 09.02.09]

Miles, A., and Brickley, D., 2005, *SKOS Core vocabulary specification*,
<http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/> [accessed 09.09.07].

Miles, A., and Brickley, D., 2004, *SKOS mapping vocabulary specification*,
<http://www.w3.org/2004/02/skos/mapping/spec/>[accessed 09.08.08].

Miles, A., and Pérez-Agüera, J., 2004, *SKOS: Simple knowledge organisation for the web*,
<http://epubs.cclrc.ac.uk/bitstream/1621/skos-ksw.pdf> [accessed 09.09.08]

Mili, H., 1988. Merging thesauri: principles and evaluation, *IEEE transactions on pattern*

analysis and machine intelligence archive, 10(2), 204-220.

Miller, P., 1999, Z39.50 for all, *Ariadne*, 21,
<http://www.ariadne.ac.uk/issue21/z3950/> [accessed 09.09.08]

Miller, P. 2002, The concept of the portal, *Ariadne,* 30,
<http://www.ariadne.ac.uk/issue30/portal/> [accessed 09.09.08].

Morgan, E., 2004, An introduction to Search/Retrieve URL Service, *Ariadne*, 40,
<http://www.ariadne.ac.uk/issue40/morgan/> [accessed 09.06.08]

Nederbragt, H., 2008, *Introduction to the STERNA architecture*,
<http://www.sternanet. eu/images/stories/documents/sterna_architecture_01.pdf>
[accessed 09.02.09]

Needleman, M., 2000, Z39.50 – a review, analysis and some thoughts on the future,
Library Hi Tech, 18(2), 158-165,
<http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2380180206.html> [accessed 09.08.08]

Neuroth, H. and Koch, T., 2001, *Cross-browsing and cross-searching in a distributed network of subject gateways: Architecture, data model, and classification,*
<http://www.stk.cz/elag2001/Papers/HeikeNeuroth/HeikeNeuroth.html>[accessed 09.09.08]

Neuroth, H. and Koch, T., 2001a. Metadata mapping and application profiles.
Approaches to providing the cross-searching of heterogeneous resources in the EU
Project Renardus, ***In:*** *DC-2001*NII, Japan,
<http://www.nii.ac.jp/dc2001/proceedings/product/paper-21.pdf> [accessed 16.07.05].

Nielsen, J., 1993, *Usability engineering*, London: Academic Press,
<http://www.usabilityhome.com/FramedLi.htm?Heuristi.htm> [accessed 19.09.08]

NISO Z39.19: Guidelines for the construction, format, and management of monolingual controlled vocabularies, 2005, USA: NISO Press.

NUA Internet Surveys, 1998,
<http://www.nua.com/surveys/how_many_online/index.html> [accessed 12.04.05].

Nunamaker, J., and Chen, M., 1990, System development in information systems research, *Proceedings of the twenty-third annual Hawaii international conference,*
<http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=205401> [accessed 09.09.08]

Nußbaumer, P. and Haslhofer, B., 2007, CIDOC CRM in action – Experiences and challenges,
Poster for the 11th European conference on research and advanced technology for digital libraries (ECDL07), Budapest,
<http://www.cs.univie.ac.at/upload//550/papers/cidoc_crm_poster_ecdl2007.pdf>
[accessed 09.09.08]

OCLC Research, 2008, *Terminology services: experimental services for controlled vocabularies*,
<http://tspilot.oclc.org/resources/overview.pdf> [accessed 09.07.08]

O'Neill, E. and Chan, L., 2003, FAST (facted application of subject terminology): a simplified LCSH vocabulary, ***In***: 69th IFLA general conference and council, Berlin.
<http://www.ifla.org/IV/ifla69/papers/010e-ONeill_Mai-Chan.pdf> [accessed 21. 03.05]

Omelayenko, B., 2008, Porting cultural repositories to the semantic web. ***In:*** Kollias, S., and Cousins, J., eds., *Semantic interoperability in the european digital library, Proceedings of the first international workshop*, SIEDL 2008, Tenerife, June 2, 2008, 14-35,

<http://image.ntua.gr/swamm2006/SIEDLproceedings.pdf> [accessed 09.03.09]

Palmer, S.B., 2001, *The Semantic Web: an introduction.*
<http://infomesh.net/2001/swintro/> [accessed 11.11.05]

Paskin, N., 1999, The digital object identifier system: digital technology meets content management, *Interlending and document supply,* 27(1), 13-16.

Patel, M., Koch, T., Doerr, M., and Tsinaraki, C. 2005, *DELOS2 WP5 Task 5.3 Deliverable D5.3.1: Semantic interoperability in digital library systems,*
<http://delos-wp5.ukoln.ac.uk/project-outcomes/SI-in-DLs/> [accessed 09.08.08]

Patton, M. Q., 1987, *How to use qualitative methods in evaluation*, London: Sage.

Peig, E., Delgado, J. and Pérez, I., 2001, Metadata interoperability and meta-search on the Web, ***In:*** *Proc. Int'l. Conf. on Dublin Core and metadata applications 2001,*
<http://www.nii.ac.jp/dc2001/proceedings/product/paper-37.pdf> [accessed 09.09.08]

Pepper, S., and Moore, G., 2001, *XML Topic Maps 1.0,*
<http://www.topicmaps.org/xtm/>[accessed 10.10.08]

Perrault, J.M., 1969, Towards a theory for UDC: essays aimed at structural understanding and operational improvement, London: Clive Bingley.

Powell, R., 1986, *Basic research methods for librarians,* United States: Ablex Publishing Corporation.

Powell, R., 1991. Guides to conducting research in library and information science, ***In***: McClure, C. and Hernon P.eds., *Library and information science research: Perspectives and strategies for improvement*, United States: Ablex Publishing Corporation, pp.15-25.

Powell, A. 2003, Mapping the JISC IE service landscape, *Ariadne,* 36, <http://www.ariadne.ac.uk/issue36/powell/> [accessed 09.09.08]

Powell, A. and Apps,A., 2001, Encoding OpenURLs in Dublin Core metadata, *Ariadne,* 27, <http://www.ariadne.ac.uk/issue27/metadata/intro.html> [accessed 22.07.05]

Purao, S., 2002, Design Research in the Technology of Information Systems: Truth or Dare, *GSU Department of CIS working paper*, <http://iris.nyit.edu/~kkhoo/Spring2008/Topics/DS/000DesignSc_TechISResearch-2002,pdf> [accessed 09.01.08].

Qin, J., and Parling, S., 2001, Converting a controlled vocabulary into an ontology: the case of GEM, *Information research,* 6(2), <http://informationr.net/ir/6-2/paper94.html>[accessed 10.12.08].

Ramsden, A. 2003, The library portal marketplace, *VINE,* 33(1), 17-24, <http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2870330103.html > [accessed 09.08.08]

Research Centre for SRM [n.d.], <http://www.socialresearchmethods.net/kb/sampnon.php> [accessed 09.08.07]

Reed, S. and Lenat, D., 2002, *Mapping ontologies into Cyc*, <http://www.cyc.com/doc/white_papers/mapping-ontologies-into-cyc_v31.pdf#search=%22Mapping%20ontologies%20into%20Cyc%22> [accessed 09.09.08].

Reese, T., 2006, Metasearch: Building a shared, metadata-driven knowledge base system, *Ariadne*, 47, <http://www.ariadne.ac.uk/issue47/reese/intro.html> [accessed 09.09.08].

Ritchie,J and J. Lewis, *et al.*., 2003a. Designing and selecting samples. ***In:*** J. Ritchie and J. Lewis, eds., *Qualitative research practice*, London, SAGE, 77-108.

Robson, C., 1993, *Real world research*, Oxford: Blackwell Publishers.

Rosson, M., and Carroll, J., 2002, *Usability engineering: Scenario-based development of human-computer interaction*, USA: Morgan Kaufmann, <http://books.google.co.uk/books?hl=enandlr=andid=RRC9IODz4VsCandoi=fndandpg=PR9anddq=Rosson+and+Carroll+2002+usability+evaluationandots=593oSKGO2Aandsig=HB_qFV6tLPy33Ob0R4o8Nlo84wk#PPP1,M1> [accessed 09.09.08]

Rowland, G. 1998, Getting more information from Internet, *New Library World,* 99(6), 222-229.

Rowley, J. 2001, "Knowledge organisation in a Web-based environment", *Management decision,* 39(5), 355-361.

Russell, R., and Day, M., 2001, *Automated and manual approaches to the provision of thesauri and subject vocabularies*, <http://www.ukoln.ac.uk/metadata/hilt/interfaces/> [accessed 09.09.08]

Sadeh, T. 2004, The challenge of metasearching, *New library world,* 105(3/4), 104-112, <http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/0721050301.html >[accessed 09.09.08]

Sadeh, T. 2006, Google Scholar Versus metasearch systems, *HEP Libraries webzine,* 12, < http://library.cern.ch/HEPLW/12/papers/1/ > [accessed 09.08.08]

Sadeh, T. 2007, Time for a change: new approaches for a new generation of library users, *New library world,* 108(7/8), 307-316,

<http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/0721080701.html > [accessed 09.09.08]

Satija, M. P., 2007, *The theory and practice of the dewey decimal classification system*, Oxford: Chandos.

Scholtz, J., 2004, Usability evaluation, AD <http://www.itl.nist.gov/iad/IADpapers/2004/Usability%20Evaluation_rev1.pdf> [accessed 09.09.08]

Scott, M. L., 2005, *Dewey decimal classification, 22nd edition: a study manual and number building guide,* Englewood, Colo: Libraries Unlimited.

Scriven, M., 1967, The methodology of evaluation, ***In***: R. W. Tyler, R. M. Gagne, and M. Scriven, Eds., *Perspectives of curriculum and evaluation evatuation*, Chicago, IL: Rand McNally, 39-83.

Sefelin, R., Tscheligi, M., and Giller, V., 2003, Paper prototyping – what is it good for? A comparison of paper-and computer-based prototyping, *Proceedings of CHI 2003*, 778-779, <http://portal.acm.org/citation.cfm?id=765986>[accessed 09.09.08]

Seorgel, D., *et al.,* 2004, Reengineering thesauri for new applications: the AGROVOC Example, *Journal of digital information,* 4(4), <http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/ > [accessed 15.10.05].

Shiri, A. and Revie, C., 2000, Thesauri on the Web: current developments and trends, *Online information review,* 24(4), 273-280, <http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2640240401.html >[accessed 09.09.08]

Shiri, A., *et al.*, 2004, User evaluation of a pilot terminologies server for a distributed multi-scheme environment, *Online information review*, 28(4), 273-283, <http://www.emeraldinsight.com/Insight/ViewContentServlet?Filename=Published/EmeraldFullTextArticle/Articles/2640280403.html >[accessed 09.09.07]

Si, L., 2007, *Encoding formats and consideration of requirements for mapping*, NKOS 2007, Budapest. <http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2007/papers/abstracts.html#si> [accessed 09.09.08]

Sicilia, M., 2006, Metadata and semantic research, *Online information review*, 30(3), 213-216.

Simon, H., 1996, *The Sciences of the Artificial*, Cambridge: MIT Press.

SKOS API, 2004, <http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html> [accessed 10.11.07]

Sowa, J.F., 2008, *Ontology*, J.F. *Sowa's Homepage,* <http://www.jfsowa.com/ontology/index.htm> [accessed 09.08.08]

Speziale, H. and Carpenter, D. R., 2003, *Qualitative research in nursing: Advancing the humanistic imperative,* New York.

St. Pierre, M. and LaPlant, W., 1998, *Issues in crosswalk content metadata standards*, < http://www.niso.org/press/whitepapers/crsswalk.html> [accessed 09.09.08].

Summers, E., Isaac, A., Redding, C., Krech, D., 2008, LCSH, SKOS and linked data, *Proc. Int' l Conf. on Dublin Core and metadata applications,* < http://dc2008,de/wp-content/uploads/2008/09/summers-isaac-redding-krech.pdf >[accessed 09.12.08]

Svenonius, E. 2000, *The intellectual foundation of information organization,* The MIT Press, London, England.

Swan, A., *et al.*, 2005, Developing a model for e-prints and open access journal content in UK further and higher education, *Learned publishing,* 18(1). < http://dx.doi.org/10.1087/0953151052801479> [accessed 18.11.05].

Takeda, H., Veerkamp, P., Tomiyama, T., and Yoshikawam, H., 1990, Modeling design processes, *AI magazine_winter*, 11(4), 37-48, <http://portal.acm.org/citation.cfm?id=95788.95795anddl=GUIDEanddl=GUIDE> [accessed 09.09.07]

Taylor, A., 2000, *Wynar's introduction to cataloging and classification*, London: Englewood.

Taylor, M., 2001, *Zthes: a Z39.50 profile for thesaurus navigation*, <http://zthes.z3950.org/profile/zthes-05.html#2 > [accessed 09.09.06]

Taylor, A., 2003, *The organization of information,* Englewood: Libraries Unlimited

The Unabridged Oxford English Dictionary, 1989, Oxford: Oxford University Press.

Tordai, A., Omelayenko, B. and Schreiber, G., 2007, Thesaurus and metadata alignment for a semantic e-culture application, ***In:*** *Proceedings of the 4th international conference on knowledge capture (KCAP-2007)*, October 28–31, 2007, Whistler, British Columbia, Canada, 199–200, <http://www.cs.vu.nl/~guus/papers/Tordai07a.pdf> [accessed 09.02.09]

Tough, A. and Moss, M., 2003, Metadata, controlled vocabulary and directories:

electronic document management and standards for records management, *Records management journal,* 13(1), 24-31.

Trippe, B., 2001, Taxonomies and topic maps: categorization step forward, *EContent magazine*,
<http://www.econtentmag.net/Magazine/Features/trippe8_01c.htm> [accessed 09.09.08]

Tudhope D., Binding C., Blocks D., and Cunliffe D., 2002, Representation and retrieval in faceted systems. (Ed: M. Lpez-Huertas), *Proceedings 7th international society of knowledge organization conference (ISKO 2002)*, Granada. *Advanced in knowledge Organization*, 8, Ergon Verlag, 191–197.

Tudhope D., and Binding C., 2005, Towards terminology services: experiences with a pilot web service thesaurus browser, *DC-2005: Proc. Int. Conf. on Dublin Core and metadata applications.*
< http://dcpapers.dublincore.org/ojs/pubs/article/view/831/827> [accessed 09.09.08]

Tudhope, D., Koch, T., and Heery, R. 2006, *JISC Terminology service review*, <http://www.ukoln.ac.uk/terminology/TSreview-jisc-final-Sept.html.> [accessed 09.09.08]

Tudhope D., Binding C., Blocks D., and Cunliffe D., 2006, Query expansion via conceptual distance in thesaurus indexed collections, *Journal of documentation*, 62 (4), 509–533.

Tudhope, D., Binding, C., and May, K., 2008, Semantic interoperability issues from a case study in archaeology, ***In:*** Stefanos Kollias and Jill Cousins, Eds, *Semantic interoperability in the European digital library, Proceedings of the first international workshop SIEDL 2008*, 88–99.

Tudhope, D., and Binding, C., 2008, *Machine understandable knowledge organisation system*,
<http://hypermedia.research.glam.ac.uk/media/files/documents/2008-07-05/SIEDL08-Tudhope-v3.pdf> [accessed 09.11.08]

Tuominen, Jouni *et al.*, 2008, *ONKI-SKOS – Publishing and utilizing thesauri in the semanticweb*,
<http://www.seco.tkk.fi/publications/2008/tuominen-et-al-onki-skos-2008.pdf> [accessed 01.03.09]

User guide for Sesame 2.2, 2008,
<http://www.openrdf.org/doc/sesame2/2.2.1/users/userguide.html> [accessed 09.08.08]

Vaishnavi, V., Buchanan, G. and Kuechler, W., 1997, A data/knowledge paradigm for the modeling and design of operations support systems, *IEEE transactions on knowledge and data engineering*, 9(2), pp. 275 – 291.
<http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=00591452> [accessed 09.09.08]

Van Assem A., Menken M., Schreiber G., Wielemaker J., and Wielinga B., 2004, A method for converting thesauri to RDF/OWL, *Proceedings international semantic Web conference*, 17-31,
< http://www.cs.vu.nl/~mark/papers/Assem04a.pdf > [accessed 09.09.08]

Van de Sompel, H., 1999, Reference linking in a hybrid library environment Part 2: SFX, a generic linking solution, *D-LIB magazine.*
<http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt2.html>[accessed 03.07.05].

Van der Ham, J.J., Dijkstra, F., Travostino, F., Andree, H., and T.A.M. de Laat, C., 2006, Using RDF to describe networks, *Future generation computer systems*, 22(8), 862-867.

Vatant, B., 2003, *OWL and Topic Map pudding.*
<http://www.mondeca.com/owl/owltm.htm> [accessed 08.08.05]

Vatant, B., 2008, Wondering about either SKOS or Web Ontology Language? Use both!, *ISKO UK meeting,*
< http://www.iskouk.org/SKOS_July2008.htm > [accessed 09.09.08]

Veltman, K. H., 2004, Towards a semantic web for culture, *Journal of Digital Information*, 4 (4),
<http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Veltman/> [accessed 09.09.07]

Vickery, B., 1997, Ontologies, *Journal of information science*, 23(4), 277-286.

Vishnevsky, T., 2008, Qualitative research, *Nephrology nursing journal*,
<http://findarticles.com/p/articles/mi_m0ICF/is_2_31/ai_n17206974> [accessed 09.09.08]

Vizine-Goetz, D., *et al.*., 2004, Vocabulary mapping for terminology services, *Journal of digital information,* 4(4),
<http://jodi.tamu.edu/Articles/v04/i04/Vizine-Goetz/> [accessed 12.06.05]

Wake, S., and Nicholson, D., 2001, Building consensus for interoperable subject access across communities. *D-Lib*, 7(9),
<www.dlib.or/dlib/september01/wake/09wake > [accessed 09.09.08]

Walker, J., 2001, Open linking for libraries: The OpenURL framework, *New library world,* 102(1163/1164), 127-133.

Warner, A., 2004, Information architecture and vocabularies for browse and search, ***In:*** Gilchrist, A. and Mahon, B., eds., *Information,* 2nd. London: Facet Publishing, pp.177-191.

Warren, P. and Alsmeyer, D., 2005, Applying semantic technology to a digital library: a

case study, *Library management,* 26(4/5), 196-205.

Weibel, S., 1995, Metadata: the foundations for resource description, *D-Magazine.* <http://www.dlib.org/dlib/July95/07weibel.html> [accessed 16.06.05].

Welty, C., 1999, Formal ontology for subject. *Knowledge and Data Engineering,* 31(2), 155-182.

Wester, J., 2007, AutoFocus: An open source facet-driven enterprise search solution, UCL, London, *Ranganathan revisited: facets for the future,* <http://www.iskouk.org/kokonov2007,htm> [accessed 09.09.08].

Wharton, C., Rieman, J., Lewis, C., and Polson, P., The cognitive walkthrough method: a practitioner's guide, ***In:*** J. Nielsen and R. Mack, Eds, *Usability Inspection Methods*, pp. 105-140.

Zeng, M., and Salaba, A., 2005, Keynote speech: Toward an international sharing and use of subject authority data, *FRBR in 21st Century catalogues: An invitational workshop,* OCLC, Dublin, Ohio,
<http://www.oclc.org/research/events/frbr-workshop/program.htm> [accessed 09.08.08]

Zeng, L., 2008, Registry requirements and issues, *NKOS 2008*, Aarhus, Denmark,
<http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2008/programme.html> [accessed 09.12.08]

Zthes, 2006,
<http://zthes.z3950.org/> [accessed 06.06.2006]

Appendix

Appendix 1: Investigation of different KOS

In this appendix, descriptive information about the characteristics of different vocabularies was listed. It is worth noting that all the information was copied from different information resources, and there is no author's comment. These information resources include:

1. Controlled vocabularies, thesauri and classification systems available in the WWW. DC Subject http://www.lub.lu.se/metadata/subject-help.html;
2. Services survey result in HILT Project Phase II http://hilt.cdlr.strath.ac.uk/hilt2web/finalreport.htm;
3. SPECTRUM Terminology Bank. http://www.mda.org.uk/spectrum-terminology/termbank.htm;
4. Cendi Terminology Locator. http://www.cendi.gov/projects/proj_terminology.html
5. Synapse (now: Factiva) Taxonomy Warehouse http://www.taxonomywarehouse.com/ ;
6. Leonard Will's web site include a number of widely-used vocabularies with the detailed information http://www.willpowerinfo.co.uk/thesbibl.htm#lists;
7. OCLC Terminology Service http://www.oclc.org/terminologies/default.htm;

Subject Headings:

1. **Library of Congress Subject Headings (LCSH)**: LCSH consists of terms with references that have been established over the years since 1898 for use in the Library of Congress's subject catalogues. It was developed to give subject access to the vast collections of one particular library. Currently, there have been already 265,000 headings developed. Different headings could be linked through the thesaurus-based relationships, such as BT, NT, and RT. The services using LCSH:

BUBL; CORC; Electronic Journal Miner (Colorado Alliance of Research Libraries); INFOMINE; OCLC's WorldCat.

2. **Sears List of Subject Headings**: For 80 years, Sears List of Subject Headings has served the needs of small and medium-sized libraries, delivering a basic list of essential headings, together with patterns and examples to guide the cataloguer in creating further headings as needed. Practical features include a thesaurus-like structure, an accompanying list of cancelled and replacement headings, and legends within the list that identify earlier forms of headings. Today, it has been developed with the new need of changing information environments. Some new and updated heading are added according to the new disciplines, such as computer technology, psychology, and new culture.
3. **Aerospace & High Technology Database Index Terms**: The index supports access to the Aerospace & High Technology Database, which covers such topics as aeronautics, astronautics, and space sciences, as well as technology development and applications in complementary and supporting fields such as chemistry, geosciences, physics, communications, and electronics. The basis was the NASA Thesaurus, which was expanded to include High Technology. 17,000 terms are developed in the Descriptors field in each record, and this vocabulary also provides more and relevant words to describe the subject of the original source material.
4. **Emerald Subject Headings**: It covers subjects, which include management, human resource management, marketing, library study, mechanical engineering, electronic and electrical engineering, etc. This subject heading system is based on a very flat structure (one or two-level of subject headings), and different journal titles are listed under each subject heading. It is only used for Emerald database.
5. **GCL (Government Category List)** is a classified list of 450 headings for use with the subject category element of the e-Government Metadata Standard (e-GMS). GCL terms in the metadata of all UK government resources drive automatic categorisation by the portal directories, and make it much easier for citizens to browse or navigate through the services and information listed on the

sites. All UK public sector resources published on external networks or entered into internal systems must carry metadata complying with the e-GMS.

6. **GEM Subject Vocabulary**: In this vocabulary, only two levels of headings are used to describe the broad topic of the resource. It can be characterized as describing the subject area of a course of study--e.g., "physics". To describe the topicality of a resource in greater detail, additional controlled vocabularies should be used. 21st Century Skills database is using this vocabulary to index the information.
7. **Medical Subject Headings (MeSH)**: MeSH is a controlled vocabulary of medical subject terms. It is used by the National Library of Medicine for indexing articles for the MEDLINE ® database and other databases that include cataloging. It is also the source of the descriptors used in Index Medicus ®. MeSH Browser provides online access to the vocabulary by allowing both alphabetical searching and hierarchical browsing. MeSH currently contains nine levels of specificity, additional levels are being planned.
8. **Proquest controlled terms**: A controlled vocabulary of 15,305 general business terms. The vocabulary can be used for searching ProQuest ® databases such as ABI/INFORM ®, Banking Information Source, Business Dateline ®, Business Periodicals, BusinessLINK ®, General Periodicals, Newspaper Abstracts, Periodical Abstracts, and Resource/One. This controlled vocabulary is based on a flat structure, and only two levels of subject terms are used to organise this controlled vocabulary.

Classifications:

1. **Dewey Decimal Classification (DDC)**: There is a specific introduction of the DDC in Section **2.2.1.4.** It is worth noting that DDC is a very widely-used vocabulary, and a big number of information services are using this vocabulary as a basic subject indexing tool. These services include Art, Design, Architecture and Media Information Gateway, BUBL Information Service, OCLC Cooperative Online Resource Catalogue, CyberDewey, IESR, HILT, Renardus, CAIRNS, a big number of University Catalogues, etc.

2. **Universal Decimal Classification (UDC)**: There is a specific introduction of the DDC in Section 2.2.1.5. The services using UDC include NISS Information Gateway, WWW Subject Tree of WAIS Databases, a number of European University Catalogue, etc.
3. **Library of Congress Classification (LCC)**: There is a specific introduction of the LCC in Section 2.2.1.3. The Information services using LCC include OCLC Cooperative Online Resource Catalogue, Cyberstacks, Internet Collegiate Reference Collection, Ready Reference Using the Internet, etc.
4. **Bliss Classification 2 (BC2)**: There is a specific introduction of the BC2 in Section 2.2.1.6. Because this classification is not completed, only few of library catalogues are using this classification. These library catalogues include Library of King's Collection within Cambridge University, Barnardo's Library, etc.
5. **National Library of Medicine Classification**: This classification covers the field of medicine and related sciences, utilizing schedules QS-QZ and W-WZ, permanently excluded from the Library of Congress (LC) Classification schedules. The various schedules of the LC Classification supplement the NLM Classification for subjects bordering on medicine and for general reference materials. The services using this classification include OMNI, NLM Classification Subject Index, etc.
6. **Ei Classification Codes**: The *Ei Classification Codes* are a classification scheme developed by Engineering Information, Inc. *EELS* is arranged according to the Ei subject classification for most of subjects. The following subject fields are included: civil engineering; mechanical engineering; electrical engineering; computing; chemical engineering/chemistry; mathematics; physics; environmental engineering/science; engineering management. The services using this classification include Compendex, Inspec, etc.
7. **Mathematics Subject Classification**: The Mathematics Subject Classification (MSC) is used to categorize items covered by the two reviewing databases, Mathematical Reviews (MR) and Zentralblatt MATH (Zbl). The MSC is broken down into over 5,000 two-, three-, and five-digit classfications, each corresponding to a discipline of mathematics (e.g., 11 = Number theory; 11B =

Sequences and sets; 11B05 = Density, gaps, topology). The service using this classification is American Mathematical Society.

8. **ACM Computing Classification**: This classification is a standard for identifying and categorising computing literature, as well as areas of computing interest and/or expertise. The Classification has two main parts: a numbered tree containing unnumbered subject descriptors, and a General Terms. The services using this classification are ACM online dagtabase, and UCT Computer Science Research Document Repository.
9. **Physics and Astronomy Classification Scheme**: It is a hierarchical numbering classification scheme. It is developed by the American Institute of Physics (AIP). It is used in a variety of ways, for example, in the online journals as a tool in searching for articles by subject. PACS is arranged hierarchically, by subdivision of the whole spectrum of subject matter in physics- and astronomy-related sciences into segments and then repeating the process of subdivision down to four levels. The services using this classification is Physical Review.
10. **AGRICOLA Subject Category Code**: The AGRICOLA Code field contains the National Agricultural Library (NAL) subject category code. Each four-character category code indicates a subject area. AGRICOLA Database is using this classification.

11. **Economic Literature Classification**: It is a subject-based alphanumeric classification scheme, and mainly covers the economic-related subject areas. It is used by Journal Economic Literature.
12. **OceanBase Taxonomy**: It is based on the Integrated Taxonomic Information System (ITIS Project, US). Limitations of the system on plankton data are as follows: up to 7 levels of taxonomic classification up to 1000 species names in standard version (can be adjusted according to individual database needs). OceanBase is using this taxonomy.
13. **NCBI Taxonomy**: This attempts to incorporate phylogenetic and taxonomic knowledge from a variety of sources, including the published literature, web-based databases, and the advice of sequence submitters and outside taxonomy expert. NCBI database is using this taxonomy.
14. **INSPEC Classification:** The Inspec Classification is an alphanumeric coding system. It covers the areas of physics, electrical & electronic engineering, computers & control and information technology. It is used for searching and retrieval of information on the Inspec database.
15. **Psycinfo Classification:** The classification categories and codes are designed to describe the content of the PsycINFO databases. Every record in PsycINFO databases receives a classification code, which is used to categorize the document according to the primary subject matter. The code also serves as the table of contents for Psychological Abstracts.
16. **JACS:** The Joint Academic Coding System is used by the Higher Education Statistics Agency (HESA) and the Universities and Colleges Admissions Service (UCAS) in the United Kingdom to classify academic subjects, especially for undergraduate degrees. A JACS code for a single subject consists of a letter and three numbers. The letter represents the broad subject classification and subsequent numbers represent further details, similar to the Dewey Decimal System. For example, F represents the Physical Sciences, F300 Physics, F330 Environmental Physics and F331 Atmospheric Physics.
17. **Defense Technical Information Center) Subject Categories:** This contains 25 broad subject fields and 251 groups to categorize the areas of scientific and technical interest. These fields and groups provide the structure for the subject grouping of technical reports in DTIC's collection and are used to define the

areas of need-to-know in distributing these reports. Published by the Defense Technical Information Center.

Thesaurus:

1. **UNESCO**: A general thesaurus. Its subjects cover education, science, culture, social and human sciences, information and communication, and politics, law and economics. The services using this thesaurus include UK National Data Archive, NDAD's on-line catalogues of datasets and documentation, UNESCO Integrated Documentation Network;
2. **Humanities and Social Science Electronic Thesaurus (HASSET)**: is a multidisciplinary thesaurus. Having been developed specifically for and by the UKDA, the subject coverage reflects the subject content of the UKDA holdings. Coverage is fuller in the core subject areas of social science disciplines: politics, sociology, economics, education, law, crime, demography, health, employment, and, increasingly, technology. These continue to be developed and are subject to addition and change as the holdings grow. It is based on the structure of UNESCO.
3. **SOSIG Thesaurus**: It is developed by social science information gateway, and mainly covers social science subject terms. It is based on HASSET. The information service using this thesaurus is SOSIG Internet Catalogue.
4. **NASA Thesaurus**: It contains authorized subject terms by which the documents in the NASA Aeronautics and Space Database are indexed and retrieved. The scope of this controlled vocabulary includes not only aerospace engineering, but all supporting areas of engineering and physics, the natural space sciences (astronomy, astrophysics, and planetary science), Earth science, and to some extent, the biological sciences. The Thesaurus contains over 18,270 terms, 4,287 definitions, and 4,470 USE references. Terms are organized within a hierarchical structure, and also include 'related terms' lists.
5. **ERIC Thesaurus**: The Thesaurus of ERIC Descriptors (Thesaurus) is a controlled vocabulary - a carefully selected list of education-related words and phrases assigned to ERIC records to organize them by subject and make them easier to retrieve through a search. Searching by Descriptors involves selecting relevant terms from this controlled vocabulary to locate information on your topic.

6. **ASIS Thesaurus of Information Science**: covers the fields of information science and librarianship. Related and peripheral fields, such as computer science linguistics, and behavioral and cognitive sciences, are examined as warranted by the strength of their relationship to information science and librarianship. The thesaurus is intended primarily as a resource to aid in indexing and searching within the fields of information science and librarianship. It has also been designed for students and researchers to serve as a guide to the terminology of the field of library science.
7. **Art and Architecture Thesaurus (AAT)**: The AAT is a structured vocabulary, arranged in 7 facets and 33 hierarchies. It is designed to provide terminology for indexing and retrieval of art information. The conceptual scope includes art, architecture, decorative arts, archaeology, material culture, and archival materials. The coverage of the AAT ranges from antiquity to the present, and the geographic scope is global. It grows through contributions by selected museums, libraries, and other cultural institutions. The focus of each AAT "record" is a concept (not a term). There are currently around 33,000 records. Although every record isn't translated into all the same languages, the thesaurus is multilingual in the sense that each record can have multiple language terms within it.
8. **GeoRef Thesaurus**: The thesaurus is a guide to the 21000 index terms used in the GeoRef database. It includes usage notes, dates of addition, and coordinates for selected place names. The interactive thesaurus is available for use with a subscription to the GeoRef database which is provided by a number of sources.
9. **Embase Thesaurus**: It is a controlled subject vocabulary arranged in a hierarchically-structured cascading "tree". The printed EMTREE includes over 42,000 drug and medical terms, approximately 10,000 codes and almost 180,000 synonyms including alternative drug names and MeSH subject headings used by the National Library of Medicine. This three-volume set is an indispensable tool for both novice and experienced searchers. The 2001 edition of the *EMTREE Thesaurus* incorporates several enhancements, including the addition of 47 new drug links focusing on "Route of Drug Administration. The services using this thesaurus are EMBASE and MEDLINE.

10. **IEEE Web Thesaurus**: IEEE Web Thesaurus Keywords contains vocabulary associated with electrical and electronic engineering. The terminology can be used to assist in indexing and retrieval of information. There are 2800 subject terms included in this thesaurus. IEEE Electronic Library is using this thesaurus.
11. **Library and Information Science Abstract Thesaurus**: This is intended as a practical tool for those searching the LISA database. It comprises some 6000 descriptors from the LISA hard-copy indexes.
12. **AGROVOC Thesaurus**: This is a multilingual, structured and controlled vocabulary designed to cover the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains (e.g. environment). Different languages are used to establish this thesaurus, which include Arabic, Czech, Chinese, English, French, Portuguese and Spanish.
13. **The NAL Agricultural Thesaurus (NALT)**: This is annually updated and the 2007 edition contains over 65,800 terms organized into 17 subject categories. NALT is searchable online and is available in several formats (PDF, ASCII text, XML, SKOS) for download from the web site. NALT has standard hierarchical, equivalence and associative relationships and provides scope notes and over 2,400 definitions of terms for clarity.
14. **The CSA/NBII Biocomplexity Thesaurus** was developed in 2002-2003 through a partnership between the NBII and CSA. It is a merger of the CSA Aquatic Sciences and Fisheries Thesaurus, the CSA Life Sciences Thesaurus. the CSA Pollution Thesaurus, the CSA Sociological Thesaurus and the CERES/NBII Thesaurus. The services using this thesaurus include CSA Biological Science; CSA Biotechnology & BioEngineering, Life Science Collection, etc.
15. **UKAT Thesaurus**: UKAT is a subject thesaurus which has been created to support indexing and searching in the UK archive sector. The backbone of UKAT is the UNESCO Thesaurus (UNESCO), a high-level thesaurus with terminology covering education, science, culture, the social and human sciences, information and communication, politics, law and economics. UNESCO was used as the basis for UKAT because of its adoption for indexing purposes by a number of archives and archive projects, including the National Archives, Archives in London and M25 Area (AIM25), Access to

Archives (A2A), CASBAH (Caribbean, Black and Asian Studies), Gateway to Archives of Scottish Higher Education (GASHE), MUNDUS (missionary archives), and the Archives Hub (which also uses Library of Congress Subject Headings).

16. **USAID Thesaurus**: It is a tool for the indexing and retrieval of technical and project document information processed by the Center for Development Information and Evaluation (CDIE) of the U.S. Agency for International Development. Among the subjects covered are: agriculture, communications, culture, demography, economics, education, energy, government and law, health, housing, industry, labor, management, natural resources, science and technology, sociology and psychology, trade, and transportation.
17. **UNBIS Thesaurus**: It contains the terminology used in subject analysis of documents and other materials relevant to United Nations Programmes and Activities. The UNBIS Thesaurus is multidisciplinary in scope, reflecting the wide-ranging concerns of the Organization. The terms included in the UNBIS Thesaurus are meant to reflect accurately, clearly, concisely and with a sufficient degree of specificity. This thesaurus is used by United Nations Bibliographic Information System
18. **OECD Macrothesaurus**: It is a social and economic science thesaurus, and structured both hierarchically and alphabetically. It is used by Organisation for Economic Cooperation and Development.
19. **The Defense Technical Information Center (DTIC) Thesaurus:** This provides a basic multidisciplinary subject term vocabulary that aids in information search and retrieval. The Thesaurus contains broader terms (BT), narrower terms (NT), scope notes, and "used for" (UF) references. Published by the Defense Technical Information Center.

Ontologies and other vocabularies:

1. **ABC Ontology**: It provides the notional basis for developing domain, role, or community specific ontologies, and it incorporates a number of basic entities and relationships common across other metadata ontologies including time and object modification, agency, places, concepts, and tangible objects. Communities wishing to build their own metadata ontologies and models may

then extend the ABC entities and relationships as needed. It is used to describe the objects within the CIMI museum and Library.

2. **Gene Ontology**: This provides a controlled vocabulary to describe gene and gene product attributes in any organism. The three organizing principles of GO are molecular function, biological process and cellular component. It is used by Gene Ontology Database.
3. **Upper Cyc Ontology**: It stores 3000 terms capturing the most general concepts of human consensus reality. It also represents a vast structure of more specific concepts descending below this upper level: logical axioms (rules and other assertions) which specify constraints on the individual objects and classes found in the real world. The service using this ontology is Public Domain Knowledge Bank.
4. **WordNet**: This is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser.
5. **CIDOC CRM**: CIDOC Conceptual Reference Model is core ontology for cultural heritage information. Expressed as an object-oriented semantic network, it defines the key implicit and explicit concepts behind data structures used for museum and cultural documentation. It aims at semantic interoperability between heterogeneous data structures and guidance for the design of good data structures, and it can be used as top level for thesauri in the domain. It has been submitted to ISO as a Draft International Standard. The CIDOC CRM forms a complete Knowledge Representation Schema. It does not intend to define terminology of any natural language, but the global concepts (interlingua) and relationships that domain experts have in mind when designing data structures (schemata, DTD etc.). It serves as a common language of normative character to compare the intended meaning of schemata, to transform queries or data and to transport data in an application neutral form.
6. **Unified Medical Language System (UMLS):** This is called a meta-thesaurus, which merges concepts from about fifty medical controlled vocabularies in to a meta-thesaurus. It is a very large, multi-purpose, and multi-lingual

vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them. The services using this meta-thesaurus are OMNI, Medical World Search, NLM Classification Subject Index, etc.

Appendix 2: Interview questions

Section One: Federated search and library portal

1. Please describe the current federated search services provided by MetaLib/SirSi/CAIRNS?
2. Please describe what library resources are cross-searchable for MetaLib/SirSi/CAIRNS product? And potentially, what materials could be searchable for this kind of federated search service in the future? (i.e. articles in some online databases, library catalogues, some local institutional repositories, e-learning objects, images, or some external e-print archives).
3. Can you describe the ways to map different metadata schemes used by different collections to facilitate subject cross-searching?
4. In Ex Libris MetaLib, a knowledge base was developed for cross-searching different collections using different protocols. Please describe the main purpose of the knowledge base, and what kind of information is stored in this knowledge base?
5. What subject-related access services can a library portal system offer?
6. Do you know if there is any controlled vocabulary used for any library portal system?
7. When a federated search service cross-searches a number of collections, and gets the results, how does this federated search service convert the heterogeneous returned results into a consistent format?
8. Because most online databases, such as LISA, ABI/INFORM, etc., have their own controlled vocabularies, do you think it is possible that a library portal system could reuse these subject terminologies to improve the subject accessibility? Can you describe potential ways to reuse these vocabularies used by different collections?
9. Please describe the future of subject-related services within different library portal systems?

Section Two: Semantic interoperability between different controlled vocabularies

1. Please describe the ways in which your project attempts to improve the interoperability between different KOS?

2. In some projects (RENARDUS, HILT, etc), Dewey Decimal Classification (DDC) has been applied as a switch language, into which different knowledge organization systems can be mapped. Can you identify any difficulties in applying a switch language in these projects?
3. Instead of using DDC as a single spine, do you think there are any other options?
4. In the project in which you are involved, is there a subject cross-browsing interface? How would you create a subject cross-browsing interface to navigate users to find the distributed item-level metadata records?
5. What mapping relationships can be used to create the mappings?
6. Who should create the mappings within a mapping project (a centralised mapping team, different local KOS holders, machine, etc.)?
7. How should the mappings be created (automatic, intellectual, statistical, etc)?
8. How does a mapping project support distributed intellectual mapping work with different partners?

Section Three: Terminology service

1. Can you describe the main ways to create a terminology service in distributed information environments?
2. Can you describe the main functionalities that a terminology service can provide?
3. To evaluate a terminology service, what standards and criteria can be used?
4. What vocabularies should be stored in a widely-used UK terminology service?
5. Who should create and maintain the mappings within a terminology service?

Section Three: Ontologies

1. Could you describe the possibility of the use of ontologies in terminology systems?
2. Is a lexical reference database (i.e. WordNet) useful for helping develop a terminology service?
3. Which ontology is the most practical to be used for terminology services (e.g. CIDOC-CRM, ABC, OpenCyc, WordNET, SUMO, etc)? And what is the reason?

4. What functionality could an ontology offer for a terminology service?

Section Four: Technical standards and semantic web technologies

1. What are the advantages or disadvantages of different protocols used for accessing a terminology service? (e.g. ADL, Zthes, SRW, SPARQL, SKOS API, etc.)
2. What are the advantages of disadvantages of different protocols for encoding terminology information? (e.g. SKOS, Zthes XML Schema, DD-8723, MARC21, etc.)
3. Can SKOS be extended to encode most widely-used controlled vocabularies, such as DDC, UDC, ERIC, etc?

Appendix 3: Established mappings

Appendix 3.1: Mappings between DDC and UKAT

000 computer science, information & general work	Major	**A bag of concepts:** • computer science • information science • information/library research • *information* • indexing languages • information exchange • information media • standards

001knowledge	Exact	knowledge
001.01 theory of knowledge	Broad	knowledge
001.012 classification	Exact	structure of knowledge

In this case, because "structure of knowledge" has an alternative label which is called as classification, so we use the exact match to establish the mapping.

001.3 humanities	Exact	humanities education

The term "humanities education" has an alternative label called Humanities, so use exact match.

001.1 intellectual life	Exact	intellectual life
001.2 scholarship and learning	Exact	learning

The term "learning" in UKAT has an altlabel called learning and scholarship.

001.4 Research; statistical methods	Exact	statistics

001.40684 research management	Exact	science administration

001.4092 researcher	Narrow	scientific researchers

Scientific researchers is narrower than 001.4092.

001.42 research methods	Exact	5138 research methods

001.4202854678 Internet research	Major	**A bag of concepts:** • internet • research

The term "Internet research" in DDC is a single concept, but the terms "internet" and "research" are two concepts combined by a Boolean operator "and". In this case, we use the minormatch to establish mapping.

002 the book	Exact	book

003 systems	Narrow	information systems

003.0285 data processing computer applications	Major	**A bag of concepts:** • data processing • computer applications

004 data processing and computer science	Exact	data processing

004.0151 mathematical principles	Broad	mathematics

Mathematical principles is in the context of computer science.

004.015113 mathematical logic--computer science	Broad	mathematical logic

004.019 psychological principles	Broad	psychology

004.023 computer science--vocational guidance	Major	**A bag of concepts:** • vocational guidance • computer science
004.0287 testing and measurement	Major	**A bag of concepts:** • testing • measurement • computer science

In this case, in UKAT the term "testing" and the term "measurement" are two general terms in the field of scientific research, but 004.0287 refers to the context of computer science. In this case, three concepts can be combined together (testing and measurement and computer science) into a bag. This bag can link to 004.0287.

004.029 performance evaluation--computer science	Broader	information systems evaluation.
004.071 computer science—education	Exact	computer science education
004.076 computer science—examinations	Major	**A bag of concepts:** • examinations • computer science

004.082 computers and women	Major	**A bag of concepts:** • computers • women

004.0846 computers and older persons	Major	**A bag of concepts:** • elderly

		• computers

004.0871 computers and the visually-impaired persons	Major	**A bag of concepts:** • disabled persons • computers

004.088378198 computers and college students	Major	**A bag of concepts:** • University students • computers

004.11 supercomputers	Broader	computers
004.12 mainframe computers	Broader	computers
004.14 minicomputers	Exact	minicomputers
004.16 microcomputers	Exact	microcomputers

004.19 hybrid and analog computers	Exact	analogue computers
004.2 systems analysis and design, computer architecture, performance evaluation	Exact	systems design
004.3 processing modes	Broader	data processing
004.5 storage	Broader	data processing
004.6 interfacing and communications	Major match	**A bag of concepts:** • computer interfaces • computer networks
004.60218 standards	Broader	standards in the field of information science
004.603 computer communications—dictionaries	Major	**A bag of concepts:** • dictionaries • computer networks
004.6076 computer communications—examinations	Major	**A bag of concepts:** • examinations • computer networks
004.62 interfacing and communications protocols	Exact	communications protocols
004.64 kinds of hardware	Broader	hardware
004.65 communications network architecture	Broader	computer networks
004.66 data transmission modes and data switching methods	Major	**A bag of concepts:** • data transmission • data exchange

004.67 wide-area networks	Broader	computer networks
004.678 Internet (World Wide Web)	Broader	computer networks
004.68 local network 004.682 intranet	Broader	computer networks

- Alternative label of computer networks is :
- computer communications
- data networks
- electronic networking
- internet
- LANs
- local area networks
- WANs

004.692 electronic mail	Exact	electronic mail
004.693 discussion groups	Exact	discussion group
004.696 videotex	Exact	videotex
004.7 peripherals	Exact	computer peripheral equipment

- wide area networks

004.71 peripherals for digital computers	Major	**A bag of concepts:** • digital computers • computer peripheral equipment
• 004.75 peripherals combining input and output functions • 004.76 input peripherals • 004.77 output peripherals	Broader	computer peripheral equipment
004.9 nonelectronic data processing	Broader	data processing
005 computer programming, programs&data	Major	**A bag of concepts:** • computer programming • computer applications
005.019 psychological principles (in computer programming)	Broad	psychology (in general)

	Broad	computer programming

005.1 programming	Exact	computer programming
005.1015113 computer logic	Broader	logic

Computer logic is a kind of logic, but psychological principles for computer programming is not in the field of psychology, it is just relevant to psychology.

005.10287 testing and measurement	Major	**A bag of concepts:** • testing • measurement • computer programming
005.10288 maintenance and repair	Major	**A bag of concepts:** • maintenance • computer programming
005.1068 computer programming management	Major	**A bag of concepts:** computer programming management
005.1092 computer programmers	Broader	computer personnel
005.11 special programming techniques	Broader	computer programming
• 005.112 modular programming • 005.112modular programming • 005.113 structured programming • 005.114 functional programming • 005.115 logic programming • 005.116 constraint programming • 005.117 object-oriented programming • 005.118visual programming	Broader	computer programming

005.12 software systems analysis and design	Exact	2306 systems design

005.13 programming languages	Exact	4724 programming languages
005.16 program maintenance	Major	**A bag of concepts:** • maintenance • computer programming

005.18 microprogramming and microprograms	Broader	computer programming
005.2 programming for specific types of computers, for specific operating systems, for specific user interfaces	Broader	computer programming
005.3 programs	Exact	computer software
005.4 systems programming and programs	Broader	information systems
005.5 general purpose application programs	Exact	computer applications
005.52 word processing	Exact	word processing
005.54 electronic spreadsheets	Broader	office automation
005.55 statistical programs	Broader	office automation
005.57 personal information management programs	Broader	office automation
005.58 presentation software	Major	**A bag of concepts:** • office automation • presentations

005.6 microprogramming and microprograms	Broader	computer programming

005.7 data in computer systems	Broader	2301 data processing
005.71 data communications	Exact	1461 data exchange
005.72 data preparation and representation, record formats	Broader	2301 data processing
005.73 data structures	Minor	databases
005.74 data files and databases	Exact	databases
005.75 specific types of data files and databases	Broader	databases
005.752 flat-file databases 005.754 network databases 005.755 hierarchical databases 005.756 relational databases 005.757 object-oriented databases 005.758 distributed data files and databases	Broader	databases
005.759 full-text database management systems	Exact	textual databases
005.8 data security	Exact	data protection

Data protection—Alternative label: data security

006 special computer method	Broad	computers
006.3 artificial intelligence	Exact	artificial intelligence
006.31 machine learning	Minor	robotics
006.312 data mining	No	
006.32 neural nets (Neural networks	Broader	artificial intelligence
006.33 knowledge-based systems	Exact match	expert systems

Expert systems-→ Alternative label: knowledge based systems

006.331 knowledge acquisition	Minor	data collection
006.332 knowledge representation	Minor	encoding
006.333 deduction, problem solving, reasoning	Major	**A bag of concepts:** • reasoning • problem solving • artificial intelligence
006.336 programming for knowledge-based systems	Major	**A bag of concepts:** • computer programming • expert systems
006.337 programming for knowledge-based systems for specific types of computers, for specific operating systems, for specific user interfaces	Major	**A bag of concepts:** • computer programming • expert systems
006.338 programs for knowledge-based systems	Major	**A bag of concepts:** • computer applications • expert systems

006.35 natural language processing	Broader	artificial intelligence
006.37 computer vision	Broader	artificial intelligence

006.4 computer pattern recognition	Exact match	pattern recognition
006.42 optical pattern recognition 006.45 acoustical pattern recognition	Broader	pattern recognition
006.5 digital audio	Narrower	digital radio broadcasting
006.6 computer graphics	Exact	computer graphics

006.7 multimedia systems	Exact	media resource centres
006.72 hardware	Broader	hardware

006.72 refers to hardware stuff in the context of multimedia system.

010 bibliography	Exact	bibliographies
010.92 bibliographers	Exact	bibliographers
011 bibliographies	Exact	bibliographies
020 library & information science	Exact match	information Sciences
020.2854678 internet—libraries	Major	**A bag of concepts:** • libraries • internet
020.601 international organizations	Narrower	international libraries
020.603-.609 national, state, provincial, local organizations	Narrower	national libraries
020.68 management	Exact	information/library management
020.682 plant management	Broader	information/library management
020.72 Library research	Exact match	Information/library research
020.7 education, research, related topics	Exact match	**A bag of concepts:** • information/library research • library education
020.7155 on-the-job training	Broader	library education
021 library relationships	Narrower	information/library cooperation
021.2 relationships with the community	Major	**A bag of concepts:** • communities • libraries
021.3 relationships with other educational institutions	Major	**A bag of concepts:** • educational institutions • libraries
021.6 cooperation and networks	major	**A bag of concepts:** • information/library cooperation information/library networks
021.64 cooperation	Exact	information/library cooperation
021.642 cooperation through union catalogs	Major	**A bag of concepts:** • information/library cooperation • union catalogs

021.82 commissions and governing boards	No	
022 administration of physical plant	Broader	information/library administration
022.1 location and site	Major	**A bag of concepts:** • sites • libraries
022.3 buildings	Exact	library buildings
022.314 public library buildings—planning	Major	**A bag of concepts:** • public library • library buildings • planning
022.317 academic library buildings—planning	Major	**A bag of concepts:** • academic library • library buildings • planning
022.4 stacks and shelving	no	
022.7 lighting for library buildings	Major	**A bag of concepts:** • lighting • library buildings
022.8 heating, ventilation, air conditioning	Major	**A bag of concepts:** • heating • air conditioning • library buildings
022.9 equipment, furniture, furnishings	Exact	information/library equipment
023 personnel management	Exact	information/library personnel
023.2 professional positions	Exact	information/library profession
021.65 networks	Exact	Information/library networks
021.7 promotion of libraries, archives, information centers	Narrower	library use promotion

Promotion of libraries, archives, information centers can cover library use promotion, information providers' promotion, etc.

021.8 relationships with government	Major	**A bag of concepts:** • government • libraries
021.83 financial support	Major	**A bag of concepts:** financial support libraries

In this case, 021.83 is actually referred to as "financial support for libraries".

023.3 technician positions	Exact	library technicians
023.4 administrative positions	Broader	information/library personnel
023.7 job description	Major	**A bag of concepts:** • job description • information/library personnel

023.8 management of in-service training	Major	**A bag of concepts:** • in-service training • personnel management • information/library personnel
023.9 elements of personnel management	Broader	information/library personnel

025 operations of libraries, archives, information centers	Major	**A bag of concepts:** • information processing • information retrieval • records management • acquisitions • information dissemination • library circulation

025.00285 libraries—automation	Exact	library automation
025.00285536 libraries--automation--microcomputer programs	Broader	library automation
025.00285536 libraries--automation--microcomputer programs	Major	**A bag of concepts:** • computer applications • microcomputers • library automation
025.02 technical services	Broader	information services
025.04 information storage and retrieval systems	Major	**A bag of concepts:** • information systems • information processing (alternative label:Information storage and retrieval)
025.04082 women--information systems use	Major	**A bag of concepts:** • information systems • information use • women
025.04087 disabled persons--information systems use	Major	**A bag of concepts:**

		• information systems • information use • disabled persons
025.06 information storage and retrieval systems devoted to specific disciplines and subjects	Broader	information systems
025.060012 scholarly Web sites	Broader	Web resources
025.060013 humanities--information systems	Major	**A bag of concepts:** • Humanities • information systems
025.0615 psychology--information systems	Major	**A bag of concepts:** • psychology • information systems
025.063 social sciences--information systems	major match	**A bag of concepts:** • social sciences • information systems
025.063067 computer sex--information systems	Broader	information systems
025.06324 elections--information systems	Major match	**A bag of concepts:** • elections • information systems
025.0633 economics--information systems	Major match	**A bag of concepts:** • economics • information systems
025.0634 law--information systems	Major match	**A bag of concepts:** • Law • information systems
025.0635 public administrations--information systems	Major match	**A bag of concepts:** • public administration • information systems
025.06364 criminal justice information systems	Major match	**A bag of concepts:** • criminal justice system • information systems
025.065 science--information systems	Major match	**A bag of concepts:** • science • information systems
025.06526 mathematical geography--information systems	Major match	**A bag of concepts:** • geography • mathematic • information systems
025.0654 chemistry--information systems	Major	**A bag of concepts:** • chemistry • information systems
025.0655 earth sciences--information systems	Major	**A bag of concepts:** • earth sciences • information systems
025.0661 medical sciences--	Major	**A bag of concepts:**

information systems	match	• medical sciences • information systems
025.0661073 nursing--medicine--information systems	Major	**A bag of concepts:** • nursing • information systems
025.067 art--information systems	Major	**A bag of concepts:** • art • information systems
025.0691 geography--information systems	Major	**A bag of concepts:** • geography • information systems
025.1 administration	Exact	information/library administration
025.11 finance	Exact	information/library finance
025.12 duplication services (Reprography)	Exact	reprography
025.17 administration of collections of special materials	Minor	library collections
025.1712 manuscripts--library treatment	Major	**A bag of concepts:** • manuscripts • information/library administration
025.1714 archival materials--library treatment	Major	**A bag of concepts:** • archival materials • information/library administration
025.1716 rare books--library treatment	Major	**A bag of concepts:** • rare books • information/library administration
025.172 broadsides--library treatment	Major	**A bag of concepts:** • broadsides • information/library administration
025.1732 serials--library treatment	Major	**A bag of concepts:** • serials • information/library administration
025.1734 government documents--library treatment	Major	**A bag of concepts:** • government information • information/library administration
025.1736 technical reports--library treatment	Major	**A bag of concepts:** • technical reports • information/library administration
025.174 CD-ROMs--library treatment	Major	**A bag of concepts:** • optical discs • information/library

		administration
025.176 atlases--library treatment	Major	**A bag of concepts:** • atlases • information/library administration
025.177 audiovisual materials--library treatment	Major	**A bag of concepts:** • audiovisual materials \ • information/library administration
025.1773 films (Photographic records)--library treatment	Major	**A bag of concepts:** • films • information/library administration
025.1782 sound recordings--library treatment	Major	**A bag of concepts:** • sound recordings • information/library administration
025.1788 scores (Music)--library treatment	Major	**A bag of concepts:** • Music • information/library administration
025.179 large-print publications--library treatment	Major	**A bag of concepts:** • Publications • information/library administration
025.1792 braille publications--library treatment	Major	**A bag of concepts:** • Braille • Publications • information/library administration
025.1794 microforms--library treatment	Major	**A bag of concepts:** • microforms • information/library administration
025.1796 flashcards--library treatment	Broader	information/library administration
025.19 administration of specific types of institutions	Major	**A bag of concepts:** • information/library administration • special libraries
025.197 research libraries—administration	Major	**A bag of concepts:** • research libraries • information/library administration
025.1974 public libraries—administration	Major	**A bag of concepts:** • public libraries • information/library administration

025.1974 public libraries—administration	Major	**A bag of concepts:** • public libraries • information/library administration
025.1977 academic libraries—administration	Major	**A bag of concepts:** • academic libraries • information/library administration
025.2 acquisitions and collection development	Major	**A bag of concepts:** • acquisitions • library collections
025.213 censorship	Exact	censorship
025.216 weeding	Minor	book selection
025.218 collection development in specific types of institutions	Major	**A bag of concepts:** • special libraries • library collections
025.21874 public libraries--collection development	Major	**A bag of concepts:** • public libraries • library collections
025.2187625 children's libraries--collection development	Major	**A bag of concepts:** • childrens libraries • library collections
025.23 acquisition through purchase	Broader	acquisitions
025.233 relations with vendors	No	
025.236 clerical operations	No	
025.26 acquisition through exchange, gift, deposit	Broader	acquisitions
025.27 acquisition of and collection development for materials on specific disciplines and subjects 025.28 acquisition of and collection development for materials in special forms 025.29 acquisition of and collection development for materials from geographic areas	Broader	acquisitions
025.3 bibliographic analysis and control	Exact	bibliographic control
025.30218 standards	Exact	bibliographic standards
025.30285 data processing Computer applications	Major	**A bag of concepts:** • data processing • computer applications • bibliographic control
025.30285574 applications of	Major	**A bag of concepts:**

data files and databases		• database applications • bibliographic control
025.302855741 applications of file organization and access methods	Major	**A bag of concepts:** • computer applications • bibliographic control • file organisation • access to information
025.31 the catalog	Exact	catalogues
025.313 form	no	
025.3132 online catalogs	Broader	catalogues
025.315 structure	Broader	bibliographic control
025.316 machine-readable record formats	Broader	machine readable materials
025.317 conversion and maintenance	Minor	data exchange

The purpose of 025.317 is to do data exchange.

025.3173 retrospective conversion	Minor	data exchange
025.3177 filing	Minor	file organisation
025.32 descriptive cataloguing	Broader	cataloguing
025.322 choice of entry and form of heading	Minor	subject headings

025.3222 authority files	Minor	indexing languages
025.324 bibliographic description	Exact	cataloguing Alternative label: Bibliographic description
025.34 cataloging, classification, indexing of special materials	Major	**A bag of concepts:** • cataloguing • classification systems • indexing
025.3412 manuscripts	Major	**A bag of concepts:** • cataloguing • classification systems • indexing • manuscripts
025.3414 archival materials	Major	**A bag of concepts:** • cataloguing • classification systems • indexing • archival materials
025.3416 rarities	Major	**A bag of concepts:** • cataloguing • classification systems • indexing

		• rare books
025.342 clippings, broadsides, pamphlets	Major	**A bag of concepts:** • cataloguing • classification systems • indexing • broadsides • pamphlets • clippings
025.343 serials, government publications, report literature	Major	**A bag of concepts:** • cataloguing • classification systems • indexing • serials • government information • reports
025.344 electronic resources	Major	**A bag of concepts:** • cataloguing • classification systems • indexing • electronic publishing
025.346 maps, atlases, globes	Major	**A bag of concepts:** • cataloguing • classification systems • indexing • atlases • globes
025.347 pictures and materials for projection	No	
025.3471 pictures and prints	Major	**A bag of concepts:** • cataloguing • classification systems • indexing • illustrations • prints
025.3473 motion pictures, slides, video recordings	Major	**A bag of concepts:** • cataloguing • classification systems • indexing • visual materials
025.348 sound recordings and music scores	major match	**A bag of concepts:** • cataloguing • classification systems • indexing • sound recordings • music
025.3492 publications in raised characters	Broader	**A bag of concepts:** • cataloguing

		• classification systems • indexing • publications
025.3494 microforms		**A bag of concepts:** • cataloguing • classification systems • indexing • microforms publications
025.3496 games, media kits, models, toys	Major	**A bag of concepts:** • cataloguing • classification systems • indexing • games
025.35 cooperative cataloging, classification, indexing	Major	**A bag of concepts:** • cataloguing • classification systems • indexing
025.39 recataloging, reclassification, re-indexing	Major	**A bag of concepts:** • cataloguing • classification systems • indexing
025.4 subject analysis and control	No	
025.40218 standards	Broader	information/library standards
025.4028 Abstracting techniques	Minor	automatic text analysis
025.42 Classification and shelflisting	Minor	classification systems
025.43 general classification systems	Broader	classification systems
025.46 classification of specific disciplines and subjects	Broader	classification systems
025.47 subject cataloguing	Broader	cataloguing
025.48 subject indexing	Broader	indexing
025.482 precoordinate indexing 025.484 coordinate and postcoordinate indexing	Broader	indexing
025.486 title manipulation	Broader	indexing
025.49 controlled subject vocabularies	Exact	controlled vocabularies
025.5 services for users	Major	**A bag of concepts:** • information services • information users
025.5019 library services--psychological aspects	Major	**A bag of concepts:** • information services • information users

		• psychology
025.50285 Public services--libraries--computer applications	Major	**A bag of concepts:** • computer applications information services • public services • information users
025.5092 Library users	Exact	information users
025.52 reference and information services	Exact	reference services
025.523 Cooperative information services	Broader	Information services
025.524 information search and retrieval	Exact	Information retrieval
025.5240218 Information retrieval--information science—standards	Major	**A bag of concepts:** • information retrieval information sciences • information/library standards
025.525 selective dissemination of information (SDI)	Exact	Selective dissemination of information
025.527 reference and information services in specific types of institutions	Broader	Reference services
025.54 reader advisory services to individuals and groups	Major	**A bag of concepts**: • readers • information services

025.56 orientation and bibliographic instruction for users	Exact	information user instruction
025.5677 academic libraries--bibliographic instruction	Major	**A bag of concepts**: • information user instruction • academic libraries
025.58 library use studies	Exact	information user studies
025.6 circulation services	Exact	library circulation
025.62 interlibrary loans	exact	interlibrary loans
025.7 physical preparation for storage and use	Minor	information use

The purpose of 025.7 is to do information use.

025.8 maintenance and preservation of collections	Major	**A bag of concepts**: • library collections

		• document preservation
025.81 physical arrangement and access to collections	Major	**A bag of concepts**: • library collections • access to information

.

025.82 security against theft and other hazards	Minor	data protection
025.84 preservation	Exact	document preservation
026 libraries for specific subjects	Exact	special libraries
027 general libraries	Broader	libraries
028 reading & use of other information media	Major	**A bag of concepts**: • information media • information use • reading
028.1 reviews	Exact	reviews
028.108 reviews with respect to kinds of persons	Broader	reviews
028.12 reviews of reference works	Major	**A bag of concepts:** • reference materials • reviews
028.13 reviews of works published in specific forms	Broader	reviews
028.132 paperbacks—reviews	Broader	reviews
028.137 audiovisual materials—reviews	Major	**A bag of concepts:** • audiovisual materials • reviews
028.138 sound recordings—reviews	Major	**A bag of concepts:** • sound recordings • reviews
028.16 reviews of works for specific kinds of users	Broader	Reviews
028.162 children--publications for—reviews	Major	**A bag of concepts:** • publications • reviews • children (age group)
028.163 disabled persons--publications for—reviews	Major	**A bag of concepts:** • publications • reviews • disabled persons
028.167 libraries--publications for—reviews	Major	**A bag of concepts:** • publications • reviews • libraries
028.5 reading and use of other information media by young	Major	**A bag of concepts:** • information media

people		• information use • Youth • reading
028.7 use of books and other information media as sources of information	Major	**A bag of concepts:** • information media • information use • books
028.8 use of books and other information media as sources of recreation and self-development	Major	**A bag of concepts:** • information media • information use • books
028.9 reading interests and habits	Exact	reading habit
030 general encyclopaedic works	Exact	Encyclopedias
050 general serial publications	Exact	serials
060 general organisation & museum science	Exact	museums
069 museum science	Exact	museology
069.1 museum services to patrons	Broader	museum services
069.2 management and use of physical plant	Broader	museum management
069.3 equipment, furniture, furnishings	Major	museum equipment
070 news media, journalism & publishing	Major	**A bag of concepts:** • journalism • publishing • newspapers
070.1 documentary media, educational media, news media	Minor	teaching materials
070.1079 documentary media--journalism—awards	Major	**A bag of concepts:** • documents • journalism • awards
070.17 print media	Exact	print media
070.172 newspapers	Exact	newspapers
070.175 serial publications (see above)	Exact	serials
070.18 motion pictures	Exact	films
070.19 broadcast media	Exact	broadcasting
070.194 radio	Exact	radio
070.195 television	Exact	television
070.4 journalism	Exact	journalism
070.408 journalism with respect to kinds of persons	Broader	journalism
070.41 editing	Exact	editing

070.41092 editors—journalism	Exact	editors
070.43 reporting and news gathering	Minor	reports
070.435 news agencies	Exact	news agencies
070.43092 reporters	Exact	reporters
070.431 news sources	Exact	news flow
070.433 reporting local, international, war news	No	
070.48 journalism directed to special groups	Broader	journalism
070.49 pictorial journalism	Broader	journalism
070.5 publishing	Exact	publishing
070.50285467 network publishing (computer networks)	Major	**A bag of concepts:** • computer networks • publishing
070.5079 federal aid—publishing	Broader	publishing
070.5083 children's books—publishing	Major	**A bag of concepts:** • childrens books • publishing
070.50871 blind persons—publishing	Major	**A bag of concepts:** • blind • publishing
070.5092 publishers—biography	Major	**A bag of concepts:** • publishers • biography
070.51 selection and editing of manuscripts	Major	**A bag of concepts:** • manuscripts • editing • acquisitions
070.51092 book editors, . . .	Broader	Editors
070.52 relations with authors	Major	**A bag of concepts:** • authors • external relations
070.57 kinds of publications	Broader	publications
070.572 serial publications	Major	**A bag of concepts:** • serials • publications
070.5722 newspapers	Major	**A bag of concepts:** • newspapers • publications
070.573 specific kinds of books	Major	**A bag of concepts:** • books • publications
070.579 special kinds of publications	broader	publications
070.5792 braille and other raised characters	Major	**A bag of concepts:** • braille

		• publications
070.5793 maps	Major	**A bag of concepts:** • maps • publications
070.5794 music	Major	**A bag of concepts:** • music • publications
070.5795 microforms	Major	**A bag of concepts:** • microforms • publications
070.5797 electronic publications (Digital publications)	Exact	electronic publishing
070.57973 web publications	Broader	electronic publishing
070.59 kinds of publishers	Broader	publishers
070.592 commercial publishers 070.593 private publishers 070.594 institutional publishers 070.595 governmental and intergovernmental publishers	Broader	publishers
080 general collections	Exact	collections
090 manuscripts &rare books	Major	**A bag of concepts:** • manuscripts • rare books
091 manuscripts	Exact	manuscripts
092 block books	Broader	rare books
093 incunabula	Broader	rare books
094 printed books	Narrower	early printed books
095 books notable for binding	Minor Match	bookbinding
095 books notable for binding	Broader	rare books
096 books notable for illustrations	Exact match	illustrated books
097 books notable for ownership or origin	Exact match	Private press books
097 books notable for ownership or origin	Broader	rare books
098 prohibited works, forgeries & hoaxes	Broader	rare books
099 books notable for format	broader	books

Appendix 3.2: Mappings between DDC and ACM

003 systems	Narrower	H. information systems
003.0285 data processing computer applications	Exact	J. computer applications

003.1 system identification	No	
003.092 systems analysts	No	
003.2 forecasting and forecasts		
003.3 computer modelling and simulation	Exact	I.6 simulation and modelling
003.5 theory of communication and control (theory)	Narrower	B.4 input/output and data communications (handware)
003.52 perception theory		
003.54 information theory (theory)	Broader	E.4 coding and information theory
003.56 decision theory	Exact	H.4.2.0 decision support (e.g., MIS)
003.7 kinds of systems	Exact	H.4.2 types of Systems
003.71 large-scale systems	Narrower	In E 1.3----sparse, structured, and very large systems From data-structure point of view
003.74 linear systems	Exact	In G 1.3----linear systems (direct and iterative methods)
	Minor	In G1.6----Linear programming
003.75 nonlinear systems	Minor	In G1.6----nonLinear programming
003.76 stochastic systems	Minor	In G1.6----- stochastic programming (NEW)
	Minor	In G.3----stochastic processes (NEW)
003.78 distributed-parameter systems	Minor	D2.12----- distributed objects (NEW)
	Minor	E.1-----distributed data structures (NEW)
003.8 systems distinguished in relation to time	No	
003.83 discrete-time systems	Minor	G.2 discrete matchematics
	Minor	I6.8-----discrete event
003.85 dynamic systems	Minor	B3.1---dynamic memory (DRAM) (NEW Or D3.3 dynamic storage management

		Or I2.8 dynamic programming
004 data processing and computer science	Narrower Broader	J.1 adminstrative processing E.data
004.0151 mathematical principles	Exact	G. mathematics of computing
004.015113 mathematical logic--computer science	exact	F.4.1 mathematical Logic (F.1.1, I.2.2-4)
004.019 psychological principles	Exact	D.m software psychology**
004.023 computer science--vocational guidance	No	
004.028 auxiliary techniques and procedures	No	
004.0287 testing and measurement In computer science	Narrower Minor	D.2.5 testing and Debugging K.7.3 testing, Certification, and Licensing
004.029 performance evaluation--computer science	Exact	H3.4 performance evaluation (efficiency and effectiveness) (NEW)
004.03 computer science—dictionaries	Broader	A.2 reference (e.g., dictionaries, encyclopedias, glossaries)
004.0688 computers--marketing management	Exact	K1----markets
004.071 computer science—education	Exact	K3 computers and education
004.076 computer science—examinations	Broader	K3 computers and education
004.082 computers and women	No	
004.0846 computers and older persons	No	
004.0871 computers and the visually-impaired persons	No	
004.088378198 computers and college students	No	
004.092 computer scientists	Broader	K.7.1 occupations
004.11 supercomputers	Exact	c.5.1 super (very large) computers
004.12 mainframe computers	Broader	C.5.1 large and medium (``Mainframe'') computer

004.14 minicomputers	Exact	C.5.2 minicomputers**
004.16 microcomputers	Exact	C.5.3 microcomputers
004.19 hybrid and analog computers	Exact	C1.3----heterogeneous (hybrid) systems (NEW) and c1.3---Analog computers (NEW)
004.2 systems analysis and design, computer architecture, performance evaluation	Major	**A bag of concepts:** • K6.1---systems analysis and design • c0--- modeling of computer architecture (NEW) • H3.4 Pperformance evaluation (efficiency and effectiveness) (NEW)
004.3 Processing modes	Major	**A bag of concepts:** • F3.2----Process models (NEW) • D2.9---Software process models (e.g., CMM, ISO, PSP)
004.5 storage	Exact	E.2 data storage representations
004.6 interfacing and communications	Major	**A bag of concepts**: • C.2 computer communication NETWORKS • C.0----Hardware/software interfaces
004.60218 standards	Exact	C2.6---- standards (e.g., TCP/IP) (NEW)
004.603 computer communications—dictionaries	Broader	A.2 reference (e.g., dictionaries, encyclopedias, glossaries
004.6076 computer communications—examinations	No	
004.61 interfacing and communications for specific types of electronic computers	Major	**A bag of concepts:** • C.2 computer communication network • C.0----Hardware/software interfaces
004.62 interfacing and communications protocols	Exact	C.2.2 network Protocols
004.64 kinds of hardware	Broader	B. hardware

004.65 communications network architecture	Exact	C.2.1 network architecture and design
004.66 data transmission modes and data switching methods	Major	A bag of concepts: • C.2 computer communication networks • C.0----Hardware/software interfaces
004.67 wide-area networks	Broader	C.2.5 local and wide-Area networks (REvaluator ISED)
004.678 Internet (World Wide Web)	Exact	C.2.5-----Internet (e.g., TCP/IP) (NEW)
004.68 local network 004.682 intranet	Broader Narrower	C.2.5 local and wide-Area networks (REvaluator ISED c.2.5----ethernet (e.g., CSMA/CD) (NEW
004.69 specific kinds of computer communications	Exact	H.4.3 communications applications
004.692 electronic mail	Exact	In H.4.3------- electronic mails
004.693 discussion groups	Broader	H.4.3 communications applications
004.696 videotex	Exact	In h.4.3----videotex
004.7 peripherals	Narrower	In B1.5----peripheral control**
004.75 peripherals combining input and output functions	Exact	B.4.2 input/Output Devices
004.76 input peripherals 004.77 output peripherals	Broader	B.4.2 input/Output Devices
004.9 nonelectronic data processing	No	
005 computer programming, programs&data	Exact	**A bag of concepts:** • D.software • E.data
005.019 psychological principles	Exact	In D.m----Software psychology**
005.0684 software consultants	No	
005.1 programming	Exact	D.1 programming techniques
005.101 philosophy and theory	Broader	D.1.0 general
005.1015113 computer logic	Exact	D.1.6 logic Programming

005.10285 computer-aided software engineering	Exact	D.2.2 computer-aided software engineering (CASE)
005.10287 testing and measurement	Broader	D.2.5 testing and debugging
005.10288 maintenance and repair	Minor	D.2.7 distribution, maintenance, and enhancement (REvaluator ISED)
005.1068 computer programming management	Exact	D.2.9 management (K.6.3, K.6.4
005.11 special programming techniques	Broader	D.1 programming techniques
005.113 structured programming	Exact	In D2.2---structured programming
005.114 functional programming	Exact	D.1.1 applicative (Functional) programming
005.115 logic programming	exact	D.1.6 logic Programming
005.117 object-oriented programming	exact	D.1.5 object-oriented Programming
005.116 constraint programming	Exact	F4.1 logic and constraint programming (REvaluator ISED)
005.112 modular programming	Broader	D.1 programming techniques
005.118visual programming	exact	D.1.7 visual Programming
005.12 software systems analysis and design	Exact	D.2 software engineering
005.13 programming languages	exact	D.3 programming languages
005.14 verification, testing, measurement, debugging	major	A bag of concepts: • D.2.5 testing and debugging • D.2.4 software/program verification (F.3.1) (REvaluator ISED)
005.15 preparation of program documentation	Broader	D2.7 documentation
005.16 program maintenance	Broader	D.2.7 distribution, maintenance, and enhancement (REvaluator ISED
005.18 microprogramming and microprograms	Major	A bag of concepts: • B.1 control structures and microprogramming • D.3.2---- microprogramming languages** • B.1.4 microprogram design aids (D.2.2, D.2.4,

		D.3.2, D.3.4)
005.2 programming for specific types of computers, for specific operating systems, for specific user interfaces	No	
005.3 programs	Major Narrower	J. computer applications D.4.9 systems Programs and Utilities
005.4 systems programming and programs	Narrower	D.4.9 systems Programs and Utilities
005.5 general purpose application programs	Broader	J. computer applications
005.52 word processing	Exact	h.4.1----word processing
005.54 electronic spreadsheets	Exact	In H.4.1---spreadsheets
005.55 statistical programs	Exact	In H.4.1---statistical software
005.57 personal information management programs	Broader	H.4.1 office Automation (I.7)
005.58 presentation software	Broader	H.4.1 office Automation (I.7)
005.6 microprogramming and microprograms	Major	**A bag of concepts:** • B.1 control structures and microprogramming • D.3.2---- Microprogramming languages** • B.1.4 Microprogram design aids (D.2.2, D.2.4, D.3.2, D.3.4)
005.7 data in computer systems	Broader	E. data
005.71 data communications	EXACT	C.2.0 Data communications
005.72 data preparation and representation, record formats	Major	E.2 data storage representation
005.73 data structures	Exact	E.1 data structures
005.74 data files and databases	Major	**A bag of concepts:** • E.5 files • H.2 database management

005.75 Specific types of data files and databases	Broader	**A bag of concepts:** • E.5 files • H.2 database management
005.752 flat-file databases	Broader	H.2.4 systems
005.754 network databases	Broader	H.2.4 systems
005.755 hierarchical databases	Broader	H.2.4 systems
005.756 relational databases	Exact	H.2.4----relational databases (NEW)
005.757 object-oriented databases	Exact	H.2.4----object-oriented databases (NEW)
005.758 distributed data files and databases	Exact	H.2.4---distributed databases (REvaluator ISED)
005.759 full-text database management systems	Exact	H.2.4 textual databases (NEW)
005.8 data security	Exact	K.6.5 security and protection (D.4.6, K.4.2)
006 special computer method	Exact	I. computer methodoliges
006.3 artificial intelligence	Exact	I.2 AI
006.31 machine learning	Exact	I.2.6 learning (K.3.2)
006.312 data mining	Exact	In H.2.8----data mining
006.32 neural nets (Neural networks	Exact	In I.5.1---- neural nets
006.33 knowledge-based systems	Exact	I.2.1 applications and expert systems
006.331 knowledge acquisition	Exact	I.2.6 knowledge acquisition
006.332 knowledge representation	Exact	I.2.4 knowledge representation formalisms and methods
006.333 deduction, problem solving, reasoning	Major	**A bag of concepts:** • I.2.3 Deduction and Theorem Proving • I.2.8 Problem Solving, Control Methods, and Search
006.336 programming for knowledge-based systems	Exact	I.2.1 applications and expert systems
006.338 programs for knowledge-based systems	Major	**A bag of concepts:** • I.2.3Logic programming • I.2.8 Dynamic programming

006.35 natural language processing	Exact	I.2.7 natural language processing
006.37 computer vision	Exact	I.5.4 computer vision
006.4 computer pattern recognition	Exact	I.5 pattern recognition
006.42 Ooptical pattern recognition 006.45 acoustical pattern recognition	Broader	I.5 pattern recognition
006.5 digital audio	No	
006.6 computer graphics	Exact	I.3 computer graphics
006.7 multimedia systems	Exact	H.5.1 multimedia information systems
006.7019 usability--web pages	No	
006.72 hardware		
006.74 markup language	Exact	I.7.2 markup languages
006.8 virtual reality	Exact	I.3.7 virtual reality

Appendix 4: Introduction of the framework

Note: This document was sent to all the selected evaluators before they interacted with the prototype system, and proposed to give the evaluators a basic understanding of the theories used to develop the framework and its prototype system.

1. Background

Currently, a number of meta-search engines have been developed to cross-search heterogeneous information resources via a single access point. The basic capability of these metadata search engines is to make the users enter queries and get results returned from heterogeneous resources. In order to develop these metadata search services, a wide range of mappings between different metadata schemes used by different services have been established. However, because different knowledge organisation systems (KOS) are used to describe these different metadata records, and these KOS differ greatly in their subject areas, degree of pre-coordination/post-coordination, level of granularity, languages, etc., the development of meta-search services has been greatly impeded by the heterogeneity of different KOS. In this context, end-users have to switch mental models between different KOS, and re-familiarise themselves with different terminologies. In addition, most of these meta-search services do not provide subject cross-browsing, because of the lack of established conceptual mapping between KOS. Subject cross-browsing is particularly helpful for inexperienced users or for users not familiar with a subject, its structure and terminology.

Prior work has been undertaken in exploring methods to establish conceptual mappings (BS8723-Part4, Zeng and Chan 2004, Koch, T., Renardus Project 2001, HILT 2003). Based on this work, this research aims to develop a framework to facilitate the subject cross-browsing functions for some library meta-search services, such as Ex Libris MetaLib, and MuseGlobal SingleSearch. The framework is based on developing different programmatic interfaces to access different established terminology resources, and establishing the semantic mappings between different KOS with these terminology resources to improve the interoperability. This report is organised as follows: Section 2 describes the strategy applied to establish mappings between different KOS used by different organisations; Section 3 presents different technical components of this framework system, and explain the relationships between these components in this framework; Section 4 indicates the relationships of this framework system with other services in a library portal, such as meta-search engines, service registry, and distributed information resources; Section 5 introduces the basic infrastructure and design principles of the software-based prototype system.

2. Mapping strategy

2. 1. Structural model for mapping: backbone structure

In this framework, it is necessary to exchange terminological information between a number of KOS. DDC was selected as a basic structure, to which different KOS used by different information resources are mapped.

2. 2. Elements to be mapped

In different types of KOS, concepts are represented in various ways. In this framework, the methods used to select the elements to be mapped are based on BS8723-Part4.This is shown in Table 1.

Types of KOS	Conceptual elements to be mapped
Thesaurus	Preferred terms
Classification scheme	Notations
Taxonomy	Category labels, notations or identifier.
Subject heading scheme	Terms or pre-coordinated strings
Authority lists	Terms or identifiers
Ontology	Terms or identifiers

Table 1: Elements to be mapped

2.3 Treatment of compound concepts

DDC is a pre-coordinated controlled vocabulary that includes a number of compound concepts. When a post-coordinated vocabulary, in which most of its concepts are individual terms, is mapped to DDC, it is important to combine several relevant concepts in the post-coordinated vocabulary to map to one concept in DDC. For example, DDC concept "020.2854678 Internet—libraries" can be mapped to the combination of the UKAT concepts "libraries" and "internet ". In theory, there are a variety of "connectors" that are able to combine different concepts from one vocabulary to another. These "connectors" mainly include Boolean Operators (and, or, not), facets (time, place, people, event, etc), and ontological relations. Currently, how to use these "connectors" for mapping is still an open issue, and some may make the mapping more complicated. With this in mind, these connectors are not used in this research. Instead, a number of relevant concepts are put into a "bag", and the bag is mapped to an equivalent DDC concept. The bag becomes a very abstract concept that may not have a clear meaning. In this case, it is not appropriate to use exact-match to represent the mappings between a bag of concepts and a DDC compound concept, because the relationship between concepts in this bag is not known. When the end-users get a list of mapped concepts through browsing, it is assumed that end-users can add appropriate Boolean operators to combine the concepts in a comprehensive way to conduct their search further. See Figure 1.

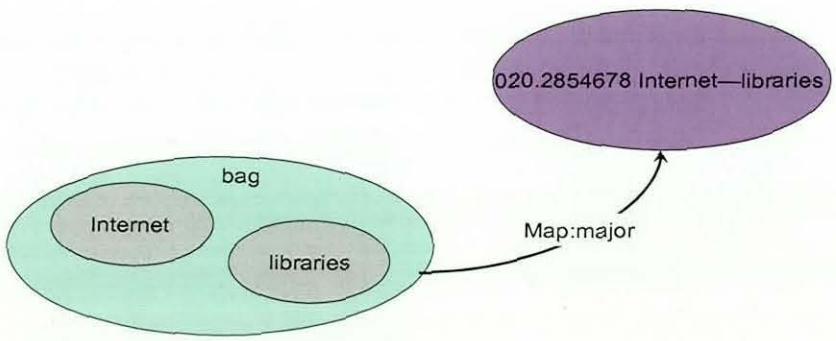


Figure 1: Treatment of a compound concept

2.4 Mapping relationships and logics

Five types of mapping relationships have been identified to represent the mapping between concepts from different KOS. These are "exact match", "major match",

"minor match", "broad match", and "narrow match". These five types are compatible with BS8723-Part4 and also form a basis to encode the mapping into SKOS-Map format. The specific definition of each mapping relationship is listed as follows:

1. Exact match: DDC contains a concept identical in scope to the concept in a specific KOS used by other resources (BS8723-part4);
2. Broad match: a concept in the target vocabulary is a superset of a concept in the source vocabulary (Renardus mapping report);
3. Narrow match: a concept in the target vocabulary is a subset of a concept in the source vocabulary (Renardus mapping resport);
4. Major match: This is used to map one compound concept against more than one post-coordinated concepts in another vocabulary, because it is not suitable to use exact match to establish the mappings between a bag of post-coordinated concepts and one compound concept in different KOS. In this situation, the compound concept should linguistically equal to the combination of post-coordinated concepts;
5. Minor match: This relationship is used to state an associative mapping link between two conceptual resources in different concept schemes. (Alistair and Matthew 2004).

However, because the actual mapping work is more complicated, and a wide variety of situations may emerge, it is necessary to establish some logic to deal with different situations, and ensure the mapping is consistent. Derived from Doerr's (2001) theory, the mapping logic is described as follows:

1. If a concept (C1) in the source vocabulary has no exact equivalence to any concept in the target vocabulary, then it is expected to find a number of concepts (Ca, Cb, Cc) in the target vocabulary that can be combined together. The combination of these concepts(Ca, Cb, Cc) in the target vocabulary can be linguistically equal to the concept (C1). In this case, we put these relevant combined concepts of the target vocabulary into a "bag", and make this bag major-map to the concept C1;
2. If a concept (C1) in the source vocabulary has no exact equivalence to any concept in the target vocabulary, and we cannot find a number of concepts in the target vocabulary that can be combined together to major map to C1, then for this concept C1, it is expected to find at least one broader equivalence or at least one narrower equivalence in the target vocabulary. The broader equivalence should be minimal, and the narrower equivalence should be maximal;
3. If a concept (C1) in the source vocabulary has no exact equivalence to any concept in the target vocabulary, we cannot find a number of concepts in the target vocabulary that can be combined together to major map to C1, and we cannot find at least one broader equivalence or at least one narrower equivalence in the target vocabulary, then it is expected to find some related concept(s) in the target vocabulary, and minor-map to the concept C1. The number of related concepts in target KOS can be one or more. If the number is more than one, we need to combine the concepts into a bag, and minor-map this bag to the concept C1.

Note: this mapping logic is used for establishing mapping from the UKAT (a post-coordinated thesaurus) to DDC (a pre-coordinated classification scheme), and for mapping from ACM (a pre-coordinated classification scheme) to DDC.

2.5 Mapping to local taxonomies

A subject cross-browsing service is an application that seamlessly combines the terminological resources from more than one KOS into a switch and browsing scheme. In many cases, it has been found that many users find it difficult to follow DDC structure to find relevant information. In many cases, people from a community may be familiar with a local taxonomy for a number of years. Thus, this framework plans to map DDC to fit with local directories, and present a local directory to the end-users. However, mapping the local taxonomies to different KOS via a DDC spine may cause one-step indirection problems and the specificity may be lost in this process. In order to minimise the indirection, only "exact-match" is allowed to be used when mapping local taxonomies to DDC. No combination of concepts is used in this stage. See Appendix 2 for an example of mappings between a local information taxonomy and DDC.

With this in mind, this framework has been developed as a vendor-hosted and a local-hosted terminology mapping service. In other words, the vendor will provide DDC-based cross-browsing API and detailed information of DDC concepts to each library. Each library has its own ability to map DDC to its own taxonomy As shown in Figure 2, the vendor will host a terminology mapping database, in which mappings established from different KOS to DDC are stored, and each organisation can also host another DDC-based terminology mapping database, in which mappings established from DDC to its local terminology are stored. In this way, DDC will become middleware between the local terminology and different KOS sources.

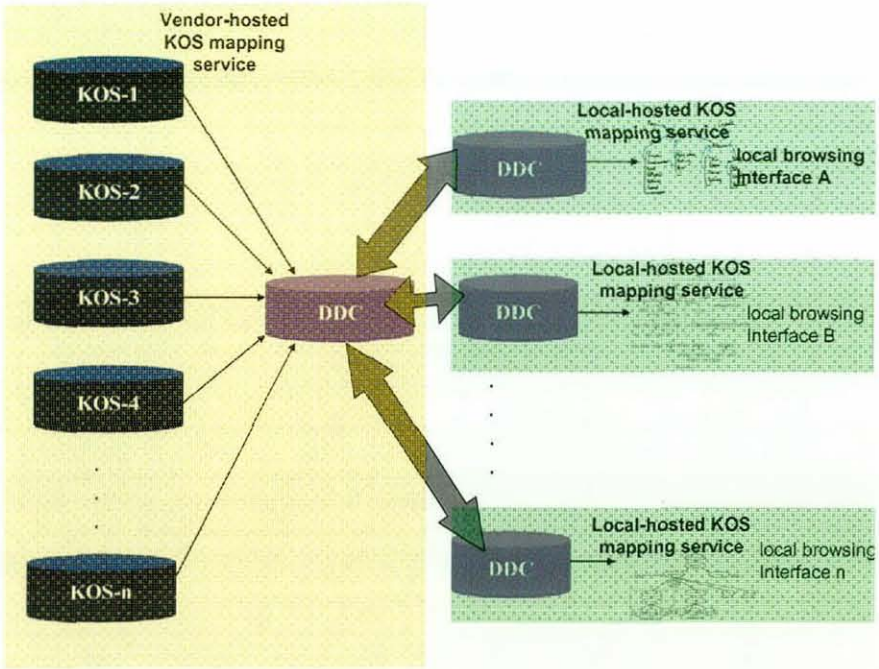


Figure 2: Mapping to the local taxonomies

2.6 Mapping work collaboration

Establishing the mappings is a tedious process. It is expected to distribute the work into different groups. All the groups can work together based on a range of consistent mapping guidelines. In this framework, it is suggested that different participants could be involved in developing a mapping service and establishing intellectual mapping work. These participants might include KOS owners, terminology mapping service

providers, and libraries. Thus, in this framework, it is proposed that the terminology mapping service providers and KOS owners can work together to generate the mappings between DDC and different KOS that form the basis of terminology mapping service. According to the mappings that have been done by the terminology mapping service provider and different KOS owners, a local librarian can establish its own mapping between its local taxonomy and DDC. The basic method is shown in Figure 3.

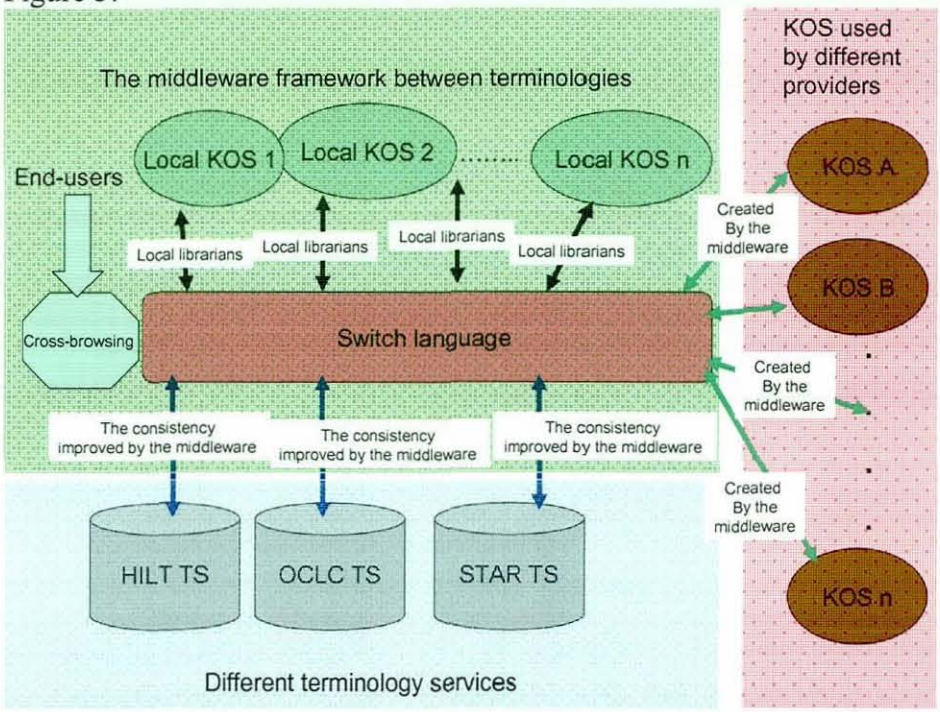


Figure 3: Mapping work collaboration

3. Framework description

As well as the semantic heterogeneity between different KOS, there are also a number of other factors that challenge interoperability between different KOS. These factors mainly include:

1. Multiple formats;
2. Multiple access protocols;
3. Multiple metadata schemes to describe different KOS;
4. Multiple terminology services where different KOS are located.

In many cases, KOS owners may not allow a third-party to hold their KOS, or convert it into a unified format. For this reason, the proposed terminology framework is expected to be able to access all the terminology data from different KOS that use different formats, access protocols, schemes and that are located in different servers. With this in mind, the basic principle of this framework is to employ a variety of agents to query various KOS. As shown in Figure 4, there are several steps that are described as follows:

1. Different KOS will intellectually be mapped into DDC;
2. The mapping data is put into SKOS-Mapping format;
3. An RDF Agent is adopted to establish a DDC-based cross-browsing API above the SKOS-mapping datasets;
4. Below the different SKOS-mapping datasets, a knowledge base is developed to store the interaction information about the type of protocols that distributed KOS support, the formats that the KOS use, the form of the queries to access

different KOS, and the formats of results that are retrieved based on the relevant protocols. In other words, the knowledge base will inform the system how to use different agents to access different KOS;

5. Based on the knowledge base, a number of appropriate agents are employed to process different KOS data in various formats;
6. When getting the results returned from different KOS, the knowledge base is also responsible for converting the results into a consistent format, and presenting this to the end-users;
7. Because different KOS owners have different licenses regarding the use of their KOS, some KOS can be set up in the local environment, but some must be accessed remotely.

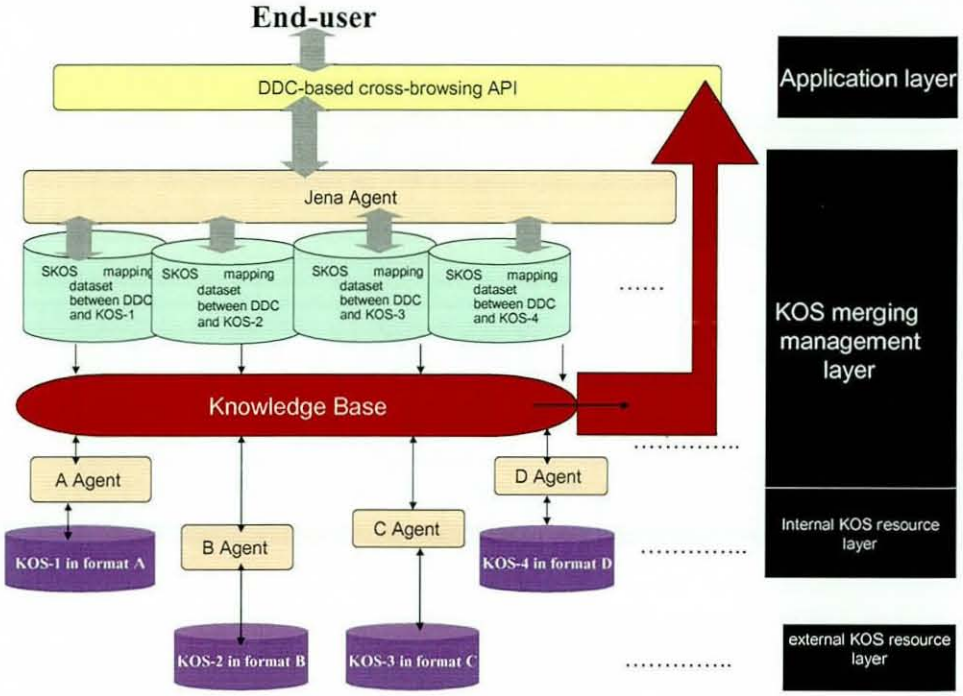


Figure 4: The terminology mapping service framework

For example, in Figure 4, it is assumed that the owners of KOS-2 and KOS-3 do not allow a third-party to hold their KOS, so their KOS have to be accessed remotely. But the owners of KOS-1 and KOS-4 may allow other systems to hold their data in another place. In the situation, KOS-1 and KOS-4 are set up in the local system.

4.Framework and meta-search engine

By now, this subject cross-browsing service can return the relevant conceptual terms, but end-users are more likely to be concerned with gaining the relevant metadata records through subject cross-browsing. It is important to make the subject cross-browsing service work with meta-search services provided by library service vendors. In this type of search process, a user can send a query to numerous information resources. The query is broadcast to each resource, and results are returned to the user. Thus, it is hoped that through interacting with the subject cross-browsing interface, the mapped conceptual terms from a particular KOS can become queries against the specific databases that are indexed by this KOS. In this context, a database registry that records the usage of KOS in different databases should be developed. The purpose of this registry is to make the mapped conceptual terms from each particular KOS become meta-search queries against the specific databases that are indexed by this KOS.

See Figure 5 as an example. The specific steps are described as follows:

1. Users interact with DDC-based subject cross-browsing interface, and get a number of mapped conceptual terms from various KOS;
2. Based on the database registry, the meta-search engine will split these mapped terms to become different queries against the databases indexed by the mapped KOS;
3. The results returned will be converted into a consistent format, and presented to the end-users;
4. A ranking algorithm should be developed based on the five different types of mapping relationships.

In this manner, an end-user can get the item-level metadata results by using the subject cross-browsing service and meta-search engines.
Note: This framework has not considered ranking algorithms.

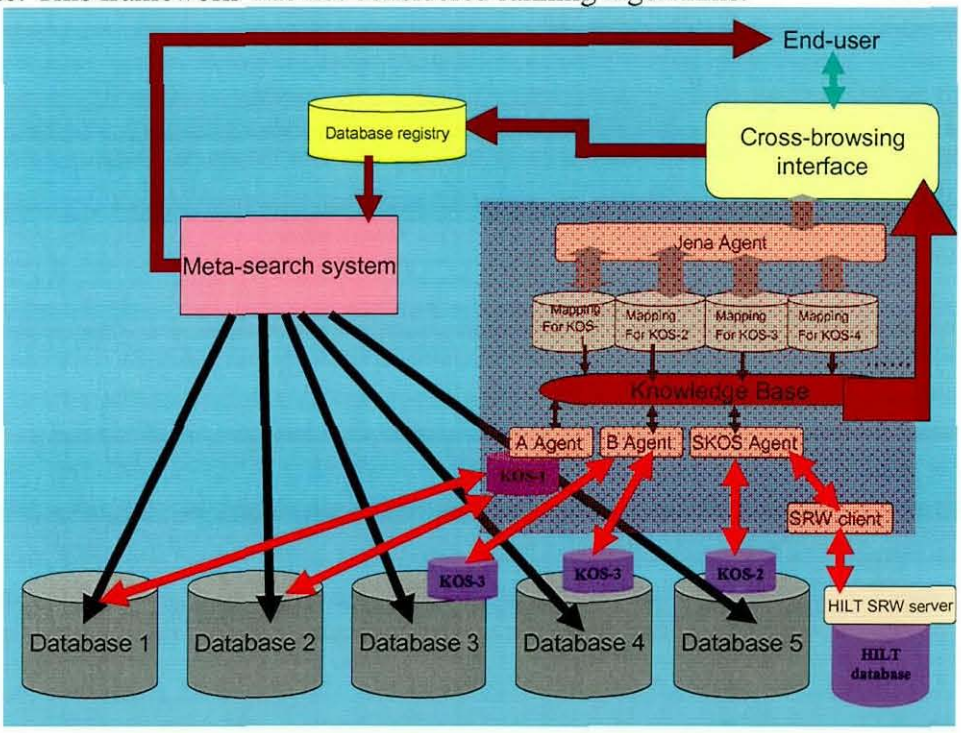


Figure 5: Cross-browsing and meta-searching

In the database registry, an RDF schema has been developed to describe the use of various KOS in different databases that can be accessed by a specific meta-search engine. The specification of this schema is listed in Table 2:

Element	Explanation
lsls2:database	Is a rdfs:class
lsls2:isindexedby	Is a rdfs:property lsls2:isindexedby rdfs:domain lsls2:database lsls2:isindexedby rdfs:range skos:conceptScheme this element is aimed to record the usage of various KOS in different databases that can be accessed by a specific meta-search engine.
dc:title	Database name
dc:subject	Database subject coverage, and recommend to use DDC number
dc:date	Database starting time

dc:format	Document format in the database
dc:language	Database language
dc:copyright	Copyright
dc:publisher	Database publisher
dc:identifier	Database URL
dc:type	The type of database

Table 2: The RDF schema used to describe the usage of various KOS in different databases that can be accessed by a specific meta-search engine.

A simple metadata record using this schema is listed as follows:

```
<?xml version="1.0" encoding="UTF-8" ?>
   <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-
rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:lsls2="http://www-
staff.lboro.ac.uk/~lsls2/scheme#">
  <lsls2:database rdf:about="http://pubs.cs.uct.ac.za">
  <dc:title>UCT CS Research Document Archive</dc:title>
  <dc:description>UCT CS Research Document Archive is an
institutional repository in the department of computer
science, at University of Cape Town</dc:description>
  <dc:language>en</dc:language>
  <lsls2:indexedby
rdf:resource="http://www.acm.org/class/1998/acmccs98-
1.2.3.xml" />
  </lsls2:database>
</rdf:RDF>
```

In this manner, by using the subject cross-browsing service and the database registry, a metadata search engine can recognise how to use the concepts returned from the subject cross-browsing service to query different databases indexed by these concepts.

5. A software-based prototype

This software-based prototype system is an application that seamlessly combines the terminological resources from more than one KOS into a switch and browsing scheme. By using DDC, the users should be navigated to find relevant concept information across different KOS. This framework has been developed in the field of library science and computing science (DDC 000-099). Two other controlled vocabularies have been selected to map to DDC. One is the ACM computing classification, and the other is the UKAT thesaurus. The mappings have been established as shown in Figure 6.

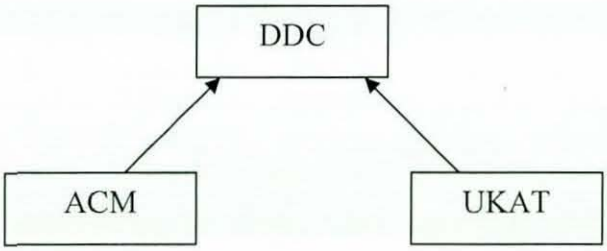


Figure 6: Mapping between DDC, UKAT, and ACM

The UKAT thesaurus is in SKOS-Core format, and the ACM Classification Scheme is in a simple XML format, see Appendix 1. In the ACM dataset, there is no URI to identify each concept. Thus, when mapping ACM to DDC, a URI is assigned to each concept of ACM in the mapping dataset, see Table 3.

DDC and ACM mapping data in SKOS-Mapping	ACM XML data
<skos:Concept rdf:about="http://www-staff.lboro.ac.uk/~lsls2/ddc.rdf/006.35"> <skos:notation rdf:datatype="http://iaaa.cps.unizar.es#notation">006.35</skos:notation> <skos:inScheme rdf:resource="http://www-staff.lboro.ac.uk/~lsls2/ddc.rdf"/> <skos:prefLabel xml:lang="en">**Natural language processing**</skos:prefLabel> <skos:broader rdf:resource="http://www-staff.lboro.ac.uk/~lsls2/ddc.rdf/006.3"/> <map:exactMatch **rdf:resource="http://www.acm.org/class/1998/acmccs98-1.2.3.xml/I.2.7"** /> //assigned URI for concept "natural language processing"// </skos:Concept>	<node id="I.2.7" label="**Natural Language Processing**"> <isComposedBy> <node label="**Discourse**" /> <node label="**Language generation**" /> <node label="**Language models**" /> <node label="**Language parsing and understanding**" /> <node label="**Machine translation**" /> <node label="**Speech recognition and synthesis**" /> <node label="**Text analysis**" /> </isComposedBy> </node>

Table 3: Mapping between KOS in different formats

In this context, two sets of mappings have been established in two SKOS-Mapping data sets. One set is the mapping between DDC and UKAT, and the other is between DDC and ACM. A Jena RDF API has been employed to process the two SKOS-Mapping data files at the same time. A DDC-based cross-browsing interface has been developed, and can be presented to the end-users. The end-users can select the concepts from DDC and get the mapped concepts from either ACM or UKAT. The specific steps are listed as follows:

4. When users select Concept A from browsing DDC, the Jena agent will process the two mapping datasets, and return the relevant URIs of the concepts mapped to Concept A of DDC from either UKAT or ACM.
5. Based on the URIs returned from the mapping datasets, two other agents are employed to deal with the UKAT data and ACM, and get more detailed information (e.g. preferred term, non-preferred terms, broader terms, related terms, etc) about the mapped concepts;
6. The returned terminological information will be converted into a consistent format, and presented to the end-users.

The two agents that are employed to process the ACM and UKAT are: a W3C DOM agent that is employed to deal with the ACM XML data, and an ARQ agent that is employed to deal with the UKAT SKOS data. The basic diagram of the prototype is shown in Figure 7.

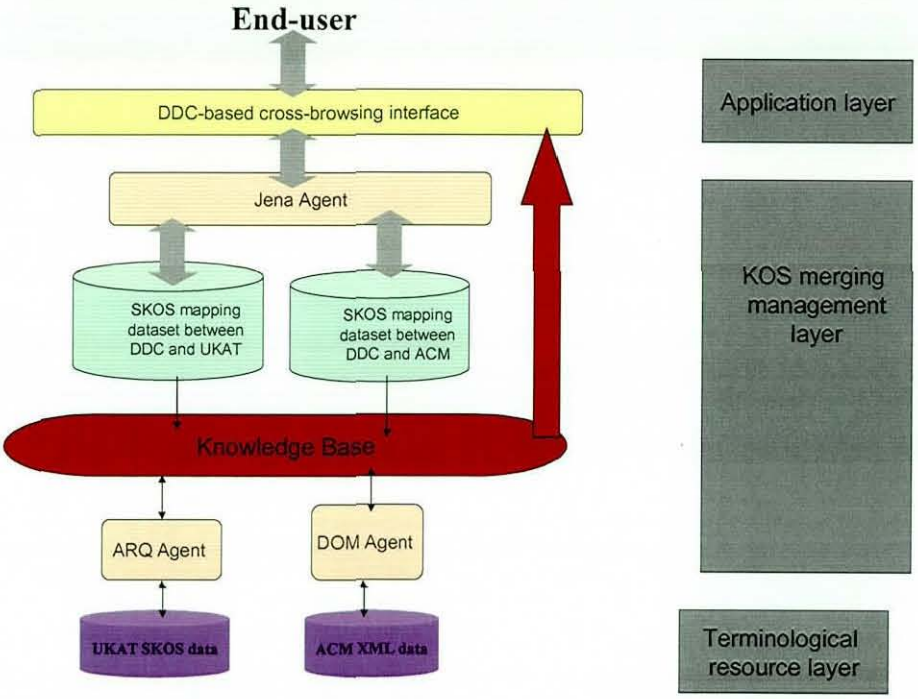


Figure 7 The infrastructure of the prototype system

From the perspective of librarians, it could be decided that DDC is not good enough to be an appropriate scheme for subject browsing in a library portal system. This may be because the inexperienced user may not be familiar with the terminology provided by DDC. In this case, here, it is assumed that the information science taxonomy developed by Hawkins (2003) is effectively used in a particular organisation as a subject browsing structure. Thus, it is required to map DDC to this taxonomy. As mentioned in 2.5, only exact match is used during this process. However, it has been found that some specific concepts in this information science taxonomy cannot map to the relevant DDC. There are three main reasons for this. First, DDC used in this research is not complete. Second, concepts in the information science taxonomy are newer than DDC concepts. Third, in some parts of this taxonomy, the granularity is more detailed than DDC's relevant part. Thus, it was decided to use two levels of this taxonomy to map to DDC. It was found that 90% of concepts in two levels of this taxonomy can map to the relevant DDC concepts.

Appendix 4.1: Mappings between Information Science Taxonomy (Source: Hawkins 2003) and DDC

1.information science research
exact map to 020.Library & Information science

1.1 basic concepts, definition, theories, methodologies, and applications
exact map to: 020.7 Education, research, related topics

1.2 properties, needs, quality, and value of information
broad map to 020.7 Education, research, related topics

1.3 statistics, measurement
exact map to 001.4 Research; statistical methods

1.4 information retrieval research
exact map to 025.524 Information search and retrieval

1.5 user behaviour and uses of information systems
broad map to 025.04 Information storage and retrieval systems

1.6 human-computer interface
broad map to 004.6 Interfacing and communications

1.7 communication
exact map to 004.6 Interfacing and communications

1.8 opertations research/mathematics
exact map to 004.0151 Mathematical principles

1.9 history of information science
exact map to 020.9 Historical, geographic, persons treatment

2. knowledge organisation
broad map to 025.4 Subject analysis and control

2.1 thesauri, authority lists
exact map to 025.49 Controlled subject vocabularies

2.2 cataloguing and classification
exact map to 025.3 Bibliographic analysis and control

2.2.1 general classification
exact map to 025.43 General classification systems

2.2.2 subject-specific classification
exact map to 025.46 Classification of specific disciplines and subjects

2.3 abstracting
exact map to 025.4028 Abstracting techniques

2.3 indexing
exact map to 025.48 Subject indexing

2.3 reviewing
exact map to 028.1 reviews

2.4 standards and protocols-1
exact map to 025.30218 Standards

2.4 standards and protocols-2
exact map to 025.40218 Standards

3. the information profession
exact map to 023 Personnel management

3.1 information professionals
exact map to 023.2 Professional positions

3.2 organisations and society
narrow map to 020.601 International organizations

3.2 organisations and society-1
narrow map to 020.603 National, state, provincial, local organizations

4. Social issues
narrow map to 021 library relationships

5. the information industry
exact map to 000 computer science, information & general work

5.1 knowledge management
broad map to 001.01 theory of knowledge

5.1 information management
exact map to 020.68 Management

5.2 marketing
exact map to 004.0688 Computers--marketing management

5.2 players
exact map to 021 library relationships

5.3 economics and pricing
exact map to 025.11 Finance

5.4 marketing and ecommerce
broad map to 004.0688 Computers--marketing management

6. publishing
exact map to 070 News media, journalism & publishing

6. distributation
exact map to 025.6 Circulation services

6.1 print
exact map to 070.17 Print media

6.2 electronic
exact map to 070.5797 Electronic publications (Digital publications)

6.3 secondary publishing
exact map to 002.0216 Lists, inventories, catalogs

6.4 scholarly communicattion
exact map to 025.060012 Scholarly Web sites

7. publishing and distribution
exact map to 004.6 Interfacing and communications

7.1 internet
exact map to 004.678 Internet (World Wide Web)

7.2 intranet
exact map to 004.682 intranet

7.3 software
exact map to 011.77 Computer programs and software

7.4 hardware
exact map to 004.64 Kinds of hardware

7.5 multimedia
exact map to 006.7 Multimedia systems

7.6 document management
exact map to 025.17 Administration of collections of special materials

7.7 AI
exact map to 006.3 Artificial intelligence

7.7 expert system, intelligent agents
exact map to 006.33 Knowledge-based systems

7.8 telecommunications
exact map to 004.69 Specific kinds of computer communications

7.9 security, access control, authentication, encryption
exact map to 005.8 Data security

8. electronic information systems and services
exact map to 025 Operations of libraries, archives, information centers

8.1 information searching and retrieval systems and services
exact map to 025.04 Information storage and retrieval systems

8.2 customised information systems, alering, current awareness
exact map to 025.5 Services for users

8.3 document delivery systems and services
exact map to 025.6 Circulation services

8.4 geographic information systems
exact map to 025.0691 Geography--information systems

9. subject-specific sources and applications
exact map to 025.06 Information storage and retrieval systems devoted to specific disciplines and subjects

9.1 physical sciences
broad map to 025.065 Science--information systems

9.2 life science
exact map to 025.0661 Medical sciences--information systems

9.3 social sciences
exact map to 025.063 Social sciences--information systems

9.3 humanities
exact map to 025.060013 Humanities--information systems

9.4 business
exact map to 025.0633 Economics--information systems

9.5 law
exact map to 025.0634 Law--information systems

9.5 political science, and government
exact map to 025.0635 Public administrations--information systems

9.6 news

exact map to 070 News media, journalism & publishing

9.7 education, library and information science, ready reference
exact map to 025.52 Reference and information services

9.8 encyclopaedias
exact map to 030 general encyclopaedic works

9.8 databases of theses and dissertations
broad map to 011.75 Theses and dissertations

9.8 full-text database management systems
exact map to 005.759 Full-text database management systems

10. libraries and library services
exact map to 026 Libraries for specific subjects

10.1 library description and types
exact map to 026 Libraries for specific subjects

10.2 library services
narrow map to 025.5 Services for users

10.3 library automation
exact map to 025.00285 Libraries automation

10.3 library operations and strategic planning
exact map to 025 Operations of libraries, archives, information centers

10.4 library consortia and networks
exact map to 021.65 networks

10.4 library cooperatives
exact map to 021.64 Cooperation

10.5 digital and virtual libraries, hybrid libraries
exact map to 020.2854678 Internet libraries

10.6 education and training
exact map to 020.7 Education, research, related topics

11. government and legal information and issues
exact map to 021.8 Relationships with government

Appendix 4.2: Information Science Taxonomy (Source: Hawkins 2003)

1. Information science research
   Basic concepts, definitions, theories, methodologies, and applications
   Properties, needs, quality, and value of information
   Statistics, measurement
   1.31 Bibliometrics, 1.32 citation analysis, 1.33 scientometrics, 1.34 informetrics
   Information retrieval research
   User behaviour and uses of information systems
   Human-computer interface
   Communication
   Operations research/mathematics

History of information science, biographies

2. Knowledge organisation
   Thesauri, authority lists
   Cataloguing and classification
   Abstracting, indexing, reviewing
   Standards and protocols

3. The information profession
   Information professionals
   Organisations and society

4. Societal issues
   Information ethics, plagiarism, credibility
   Information literacy, lifelong learning
   The information society

5. The information industry
   Information and knowledge management
   Markets and players
   Economics and pricing
   Marketing, e-commerce
6. Publishing and distribution
   Print
   Electronic
6.3 Secondary publishing
6.4 Scholarly communication
7. Publishing and distribution
   Internet
   Intranets, Web conferencing
   Software
   Hardware
   Multimedia
   Document management
7.7 AI, expert systems, intelligent agents
7.8 Telecommunications
7.9 Security, access control, authentication, encryption

8. Electronic information systems and services
   Information searching and retrieval systems and services
   Customised information systems, alerting, current awareness
   Document delivery systems and services
   Geographic information systems

9. Subject-specific sources and applications
   Physical sciences
   Life science
   Social sciences, humanities, history, linguistics
   Business
   Law, political science, government
   News
   Education, library and information science, ready reference

Other/multidisciplinary

10. Libraries and Library services
10.1 Library description and types
10.2 Library services
10.3 Library automation, operations and strategic planning
10.4 Library consortia and networks, coalitions, cooperatives
10.5 Digital and virtual libraries, hybrid libraries
10.6 Education and training

11. Government and legal information and issues

Appendix 5: Evaluation scenarios

Note: In this appendix, different questions were asked, and these questions were designed to map to the five heuristics listed in Section 3.4.4.

Appendix 5.1: Scenario 1

Objectives:

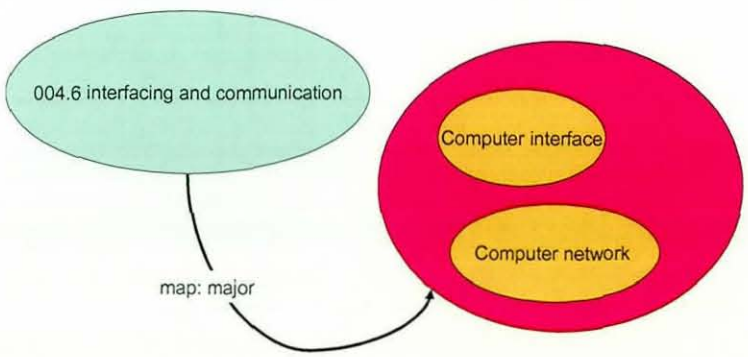
1. To help the evaluators to get familiar with browsing functions of the prototype;
2. To examine the effectiveness of the methods to combine a number of relevant concepts into a bag, and map the bag to a compound concept;

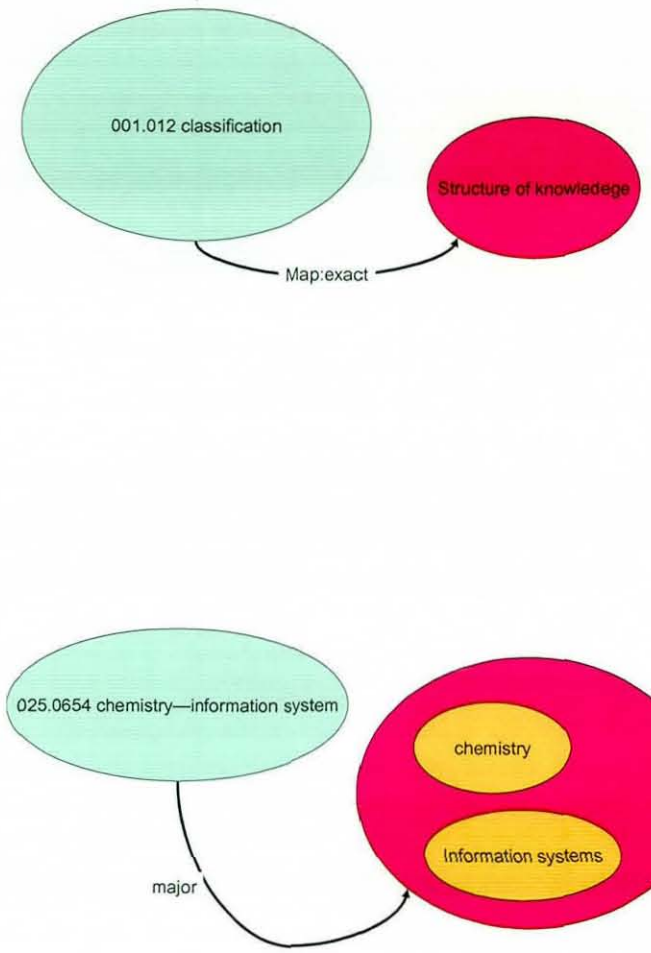
Interface used: Interface0

Tasks:

Please follow the following steps:

1. Please click the checkbox called "Mapping to UKAT"
2. Look at DDC concepts in the list, select DDC concept called "001.012 classification", and click the "search" button
3. Read the results carefully
4. Click the "clear" button
5. Select DDC concept called "025.0654 chemistry—information systems" in the list, and then click the "search" button
6. Read the results carefully
7. Click the "clear" button
8. Select DDC concept called "004.6 interfacing and communication" in the list, and then click the "search" button
9. Read the results carefully
10. Click the "clear" button
11. Look at the following pictures





Picture 1: the mappings from DDC to UKAT

Questions:

1. Are the five identified mapping relationships (exact, major, minor, broad, narrow) appropriate to represent the actual conceptual mapping between DDC and UKAT? (map to Heuristic 1—semantic extensibility)
2. What additional relationships could be identified to help the users select the relevant mapped concept? (map to Heuristic 1—semantic extensibility)
3. Is it useful if users could add Boolean operators to combine the concepts when a concept in the browsing structure returns more than one concepts from UKAT? (map to Heuristic 1—semantic extensibility)

Appendix 5.2: Scenario 2

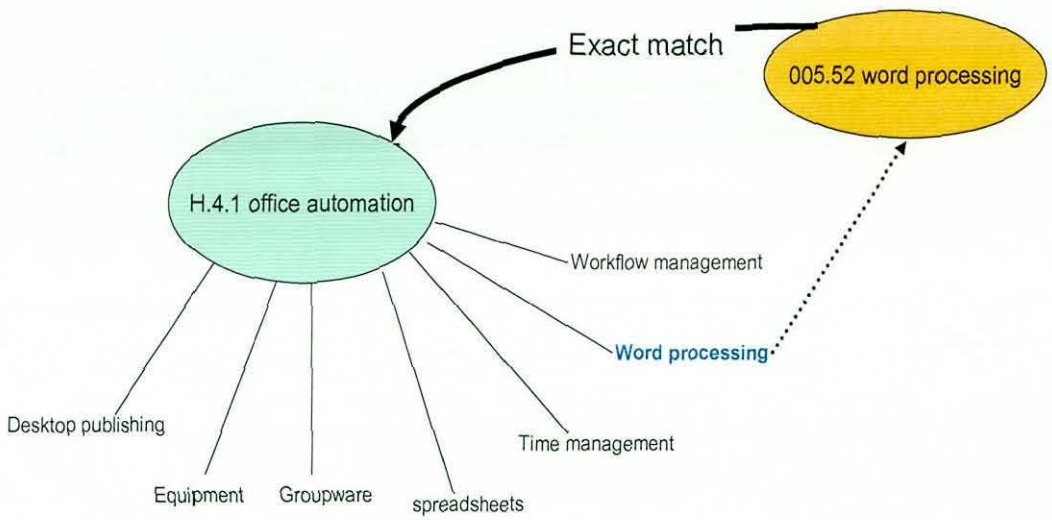
Objective:

1. To clarify the element to be mapped.

Interface used: Interface1

Tasks: Please follow the following steps:

1. Please click the checkbox called "Mapping to ACM";
2. Select DDC concept called "004.0151 Mathematical principles", and click the "search" button;
3. Read the results carefully;
4. Click the "clear" button.
5. Select DDC concept called "004.696 Videotex" in the list, and then click the "search" button;
6. Read the results carefully;
7. Click the "clear" button.
8. Select DDC concept called "005.52 word processing" in the list, and then click the "search" button;
9. Read the results carefully;
10. Click the "clear" button.
11. See Picture 2.



The terms under the concept "H.4.1 office automation" are not concept in ACM These terms have no notations.
Logics: If a concept (C-1) in ACM include a number of terms, and one of these terms is equivalent to a DDC concept (C-2), then we exactly map C-1 to the C-2.

Picture 2: The mapping from DDC to ACM

Questions:

1. Do you think the mapping presented in Picture 2 is correct? If not, please express the correct way to present the mappings. (map to Heuristic 1—semantic extensibility)
2. Are there too many terms displayed on the screen? (map to Heuristic 3—user interactivity)
3. In Picture 2, is there an alternative way to display a mapped classification concept that includes a large number of terms? (map to Heuristic 1—semantic extensibility and Heuristic 3—user interactivity)

Appendix 5.3: Scenario 3

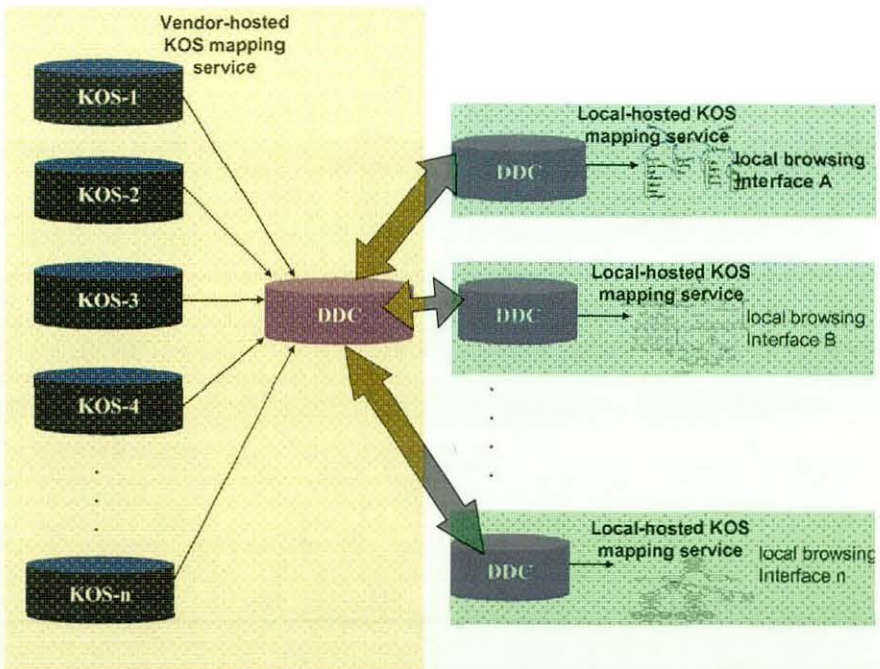
Objectives:

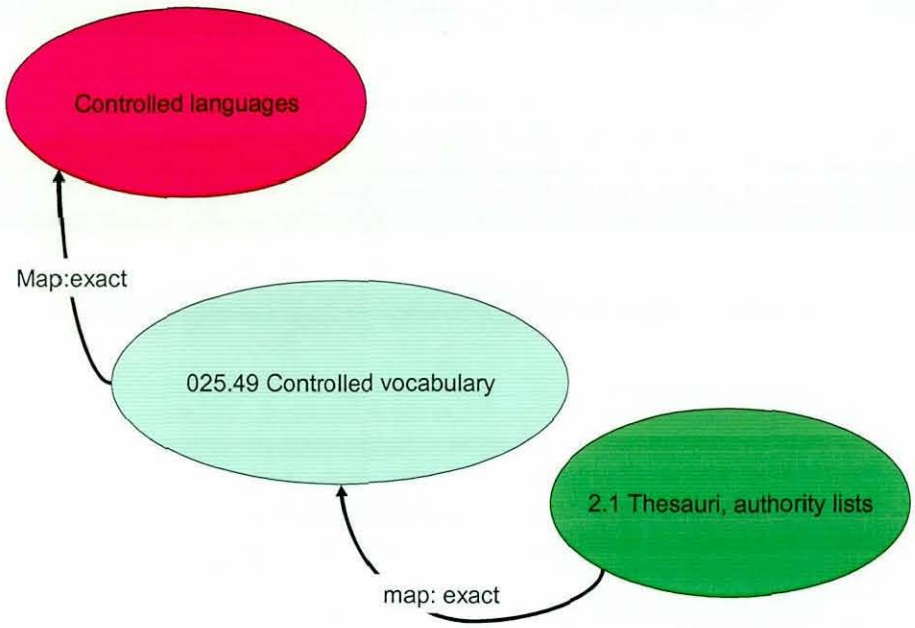
1. To test the usability of the mappings between a local information science taxonomy and DDC;
2. To test whether DDC can be a middleware between a local taxonomy and different KOS used by different information resources.

Interface used: Interface21

Tasks: Please follow the following steps:

1. Click the checkbox called "mapping to UKAT";
2. In the list, select the concept called "7.7 AI", and click the "search" button;
3. Read the results carefully;
4. Click the "clear" button.
5. In the list, select the concept called "2.1 thesauri, authority lists", and click the "search" button;
6. Read the results carefully;
7. Click "clear" button.
8. In the list, select the concept called "10.6 education and training", and click the "search" button;
9. Read the results carefully;
10. Click the "clear" button.
11. See picture 3.





Picture 3: The mappings from a local taxonomy to DDC

Questions:

1. Is there any advantage in using a local taxonomy as a subject browsing interface rather than using DDC? (map to Heuristic 3—user interactivity)
2. Do you think there is the possibility of data loss by switching terminological information from a local taxonomy to different KOS via a DDC? Can you describe some potential solutions to solve the problems you have identified? (map to Heuristic 5—technical adaptability, and Heuristic 1—semantic extensibility)
3. Who should create the mappings between DDC and a local taxonomy? Who should create the mappings between DDC and other vocabularies? (map to Heuristic 4—cultural feasibility)
4. How deep should the mappings between DDC and a local taxonomy be created? How deep should the mappings between DDC and other vocabularies be created? (map to Heuristic 1—semantic extensibility and Heuristic 3—user interactivity)

Appendix 5.4: Scenario 4

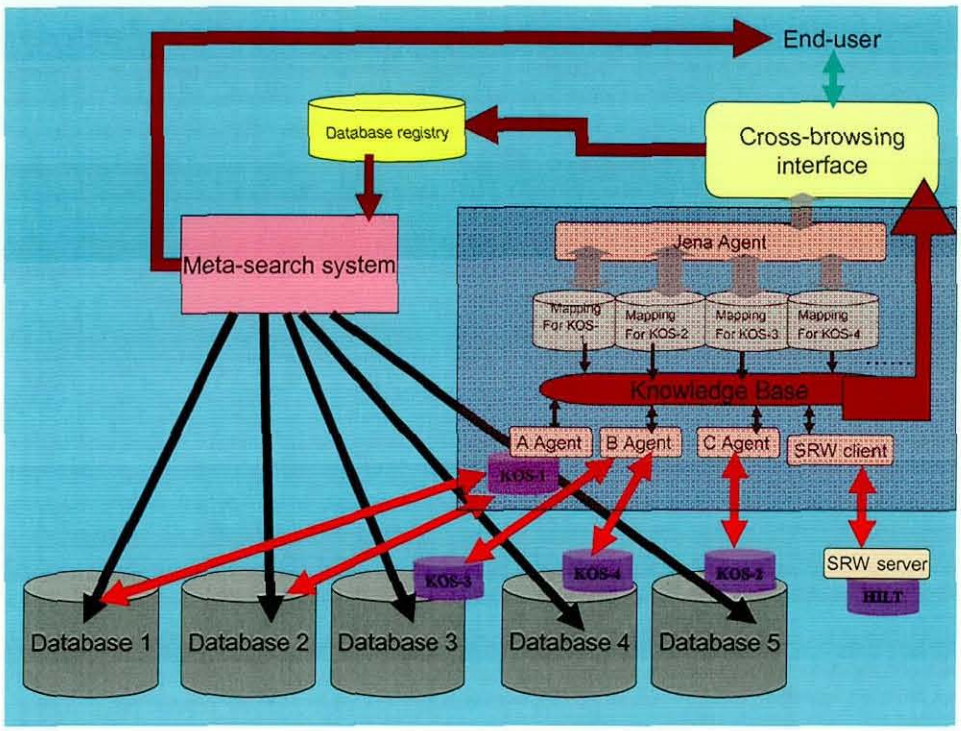
Objectives:

1. To test the viability of the cooperative work between a terminology mapping service and meta-search engines?

Interfaces used: Interface21 and Interface3

Tasks:

1. **To find the mapped concepts:** In Interface21, click the checkboxes called "mapping to ACM" and "mapping to UKAT", select the concept called "7.5 multimedia", and click the "search button";
2. Read the results carefully;
3. Click the "clear" button;
4. **To find the online databases indexed by the UKAT and ACM:** In interface3, click the checkboxes called "The databases indexed by ACM Computing Classification" and "The databases indexed by UKAT Thesaurus";
5. Read the results carefully;
6. **To find item-level metadata records cross the online databases indexed by the UKAT and ACM:** Please imagine that, through the database registry, a meta-search engine will automatically use the mapped concepts in UKAT to cross-search the databases indexed by UKAT, and use the mapped concepts in ACM to cross-search the databases indexed by ACM;
7. Please imagine the item-level metadata records are returned from the databases indexed by UKAT and ACM;
8. See Picture 4.



Picture 4: The technical architecture of this framework

Questions:

1. Can you identify the technical and cultural difficulties in implementing this system in a real situation? (map to Heuristic 5—technical feasibility)
2. In this system, what factors do you like the most or least? Why? (map to Heuristic 3—user interactivity)
3. What other additional functions could be further developed in this system to improve subject cross-browsing? (map to Heuristic 2—technical adaptability)

Appendix 5.5: Final general interview questions

1. Based on your own experience and understanding, can you give me your general comments about this framework?

2. Do you have some other suggestions as to how to generate some similar subject cross-browsing services to navigate users to find relevant item-level metadata records? (Heuristic 5—technical feasibility)

3. Is DDC an appropriate starting point (subject cross-browsing) to conduct a search? Why? (map to Heuristic 3—user interactivity)

4. Is DDC a good switch language to switch terminological information between local taxonomies and different KOS used by different information resources? Could you suggest any other controlled vocabulary that can be used as a switch language? (map to Heuristic 1-semantic extensibility)

5. Because there are a wide variety of terminology resources that use different formats, access protocols, schemes and that are located in different servers, do you think this framework could adapt to access and search most of these various KOS in distributed information environments? Can you identify any problems when this system accesses different terminology resources in distributed information environments? (map to Heuristic 4—cultural feasibility and Heuristic 2—technical adaptability)

6. In order to retrieve item-level metadata records across different collections via a subject cross-browsing interface, it is necessary for this framework to work with other information services. Can you identify any problems when this framework system works with the meta-search engine and database registry shown in Picture 4? Is there any other subject-related services with which this terminology mapping system can interact with in M2M ways? (map to Heuristic 4—cultural feasibility and Heuristic 2—technical adaptability)

7. As shown in Picture 4, do you think the core part of this subject cross-browsing service is technically feasible? From a technical point of view, can you identify any problems when implementing a real system such as this? (Heuristic 5—technical feasibility)

8. Who should create the mappings between the switch language and different KOS? Who should create the mappings between the local taxonomy and DDC?

Glossary

AAT: Art and Architecture Thesaurus, developed by the Getty Institution.

API: An application programming interface (API) is a range of developed functions, procedures, methods, classes or protocols that an operating system, library or service offers to facilitate requests asked by different computer programs.

ARQ API: A Java API to manipulate RDF data, in which SPARQL query could be constructed.

BS8723: It is a UK-based national standard to provide guideline for the construction of structured vocabularies for information retrieval.

Caption: Statement of the subjects represented by a *notation* in a classification scheme.

Collection registry: A collection registry is used to "assist other applications, such as portals, virtual learning applications or research services, to discover and devolve materials that match their users' interests in their research, learning and teaching. It contains quality description information (metadata) about resources and services"[9].

Compound concept: Concept constructed from single conceptual terms.

Co-occurrence mapping: This is based on the construction of links between concepts from different KOS based on the co-occurrence of instances.

DDC: Dewey Decimal Classification.

DCMI: Dublin Core Metadata Initiative.

Direct mapping: This refers to establishing concept mappings between only two controlled vocabularies.

Explicator Project: A UK-based Project aims to convert different astronomy vocabularies into SKOS format, and then improve interoperability between these vocabularies.

HASSET: Humanities and Social Sciences Electronic Thesaurus

HILT: HIgh Level Thesaurus Project

Institutional repository: An Institutional Repository is "an online locus for collecting, preserving, and disseminating -- in digital form -- the intellectual output of an institution, particularly a research institution"[10].

---

[9] Mimas Annual Report. http://www.mimas.ac.uk:8080/reports/annual/year0607/Online/

[10] Cochrane, T., 2007, The challenge of developing successful institutional repositories in the research sector, http://www.nzvcc.ac.nz/files/u2/Challenge_Tom_Cochrane.ppt

Interoperability: Ability of two or more systems or components to exchange information and to use information that has been exchanged.

IPSV: The Integrated Public Sector Vocabulary.

Java Applet: This is "a program written in the Java programming language that can be included in an HTML page, much in the same way an image is included in a page"[11].

JENA: Jena is a Java framework for building Semantic Web applications. It provides a programmatic environment for RDF, RDFS and OWL, SPARQL and includes a rule-based inference engine.

JISC: Joint Information Systems Committee.

KoMoHe Project: A German project supervised a terminology mapping effort, in which 'cross-concordances' between major controlled vocabularies were organized, created and managed.

KOS: "The term knowledge organization system is intended to encompass all types of schemes for organizing information and promoting knowledge management. Knowledge organization systems include classification and categorization schemes that organize materials at a general level, subject headings that provide more detailed access, and authority files that control variant versions of key information such as geographic names and personal names. Knowledge organization systems also include highly structured vocabularies, such as thesauri, and less traditional schemes, such as semantic networks and ontologies. Because knowledge organization systems are mechanisms for organizing information, they are at the heart of every library, museum, and archive"[12].

LCC: Library of Congress Classification.

LCSH: Library of Congress Subject Headings.

M2M: Machine to machine interaction.

Mapping: The process of establishing relationships between the terms, notations, or concepts of one vocabulary and those of another.

Mapping collaboration: The mapping work between different vocabularies could be distributed to different organizations.

Meta-search (also called federated search): This refers to use a search engine to cross-search a number of information resources.

---

[11] http://java.sun.com/applets/

[12] www.clir.org/pubs/abstract/pub91abst.html

NetBean: This refers to both a platform for the development of applications for the network (using Java), and an integrated development environment (IDE) developed using the NetBean Platform[13].

Notation: Sets of symbols to represent a concept.

OAI-PMH: The Open Archives Initiative Protocol for Metadata Harvesting.

OCLC: Online Computer Library Centre.

OPAC: Online public access catalogue.

OWL: Web Ontology Language, which was developed by W3C.

Post-coordination: Controlled terms are combined to form concepts at the time of indexing.

Precision: Ratio of relevant items retrieved to total items retrieved.

Pre-coordination: Controlled terms are constructed from several concepts so that fewer terms are needed for indexing.

Query expansion: In a vocabulary, the thesaurus-based relationships, such as BT, NT, RT, USE, etc., can be used to "expand" terms semantically.

RDF: Resource description framework is a data model based upon the idea of making statements about Web resources in the form of subject-predicate-object expressions, called *triples* in RDF terminology[14].

Recall: Ratio of retrieved relevant items to relevant items in the database.

Renardus Project: A European project aims to establish an academic subject gateway service in Europe, co-ordinated by national initiatives.

Scope note: In a controlled vocabulary, an explanation of how the term is to be used.

Semantic web: It is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.

SKOS: Simple Knowledge Organisation System.

Social tagging: A social tagging system is a collaboratively generated, open-ended labeling system that enables Internet users to categorize content such as Web pages, online photographs, and Web links.

[13] http://netbeans.org/
[14] http://www.w3.org/TR/rdf-concepts/

Source vocabulary: Language or vocabulary which serves as a starting point when seeking a corresponding term in another language or vocabulary[15].

SPARQL: Simple protocol and RDF query language is a RDF query language. It can be used to express queries across diverse data sources, whether the data is stored natively as RDF or viewed as RDF via middleware. SPARQL contains capabilities for querying required and optional graph patterns along with their conjunctions and disjunctions[16].

SRW/U: Search/Retrieve Web protocol (SRW) and Search/Retrieve via URI are web services protocols for querying indexes and databases on the web and returning search results. The idea behind them is similar to the older Z39.50 protocol for search and retrieve. However, SRW/U is based on modern Web 2.0 technology and uses HTTP, web browsers, and XML. It is therefore simpler to implement. Queries in SRW/U are expressed using the Common Query Language (CQL)[17].

STAR Project: The Semantic Technologies for Archaeological Resources Project.

STITCH Project: A Dutch project aims to improve semantic Interoperability between different KOS to access multiple cultural heritage collections.

STERNA Project: Semantic Web-based Thematic European Reference Network Application.

Structural model: This refers to the ways to create the mappings between more than two vocabularies[18].

Target vocabulary: Language or vocabulary in which a term is sought corresponding to an existing term in a source language or vocabulary[19].

Taxonomy: Structured vocabulary using classificatory principles as well as thesaural features, designed as a navigation tool for use with electronic media[20].

Terminology registry: A service is constructed to "list, describe, and point to a variety of vocabularies. It can hold vocabulary information, such as member terms, concepts and relationships, and provide terminology services, for both human inspection and m2m access"[21].

---

[15] BS 8723-4, 2008, Structured vocabularies for information retrieval, Part 4:Interoperability between vocablaries, London: British Standards Institution

[16] http://www.w3.org/TR/rdf-sparql-query/

[17] SRW: Search/Retrieve Webservice. http://srw.cheshire3.org/SRW-1.1.pdf.

[18] BS 8723-4, 2008, Structured vocabularies for information retrieval, Part 4:Interoperability between vocablaries, London: British Standards Institution

[19] BS 8723-4, 2008, Structured vocabularies for information retrieval, Part 4:Interoperability between vocablaries, London: British Standards Institution

[20] BS 8723-3, 2008, *Structured vocabularies for information retrieval, Part 3:Vocabularies other than thesauri*, London: British Standards Institution.

[21] Golub, K. and Tudhope, D., 2008, *Terminology registry scoping study (TRSS): Excerpt on metadata,* http://www.ukoln.ac.uk/projects/trss/dissemination/metadata.pdf

Terminology resource: In this research, a terminology resource can refer to a variety of KOS-related sources that are published in relevant digital formats. These might include:

- Terminology services, such as OCLC Terminology Service, HILT Terminology Service, etc.—which were developed as shared services that allow other services to access their terminological data;
- Controlled vocabularies, which were used to index important collections, and were represented in well-defined encoding formats, and published on the Web;
- Mapping sets between different controlled vocabularies which were represented in well-defined encoding formats, and published on the Web;
- Local vocabularies which were used by a library portal system for local subject indexing and cataloguing.

Terminology Services (TS): Terminology services 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. They can be m2m or interactive, user-facing services and can be applied at all stages of the retrieval process"[22].

UDC: Universal Decimal Classification.

URI: A Uniform Resource Identifier (URI) is "a compact string of characters used to identify or name a resource on the Internet. The main purpose of this identification is to enable interaction with representations of the resource over a network, typically the World Wide Web, using specific protocols"[23].

W3C: World Wide Web Consortium

XML: The Extensible Markup Language (XML) is a general-purpose *specification* for creating custom markup languages. It is classified as an extensible language because it allows its users to define their own elements. Its primary purpose is to help information systems share structured data, particularly via the Internet, and it is used both to encode documents and to serialize data[24].

XTM: XML Topic Map is a standard for the representation and interchange of knowledge, with an emphasis on the findability of information. It represents information using topics (representing any concept, from people, countries, and organizations to software modules, individual files, and events), associations (representing the relationships between topics), and occurrences (representing information resources relevant to a particular topic)[25].

---

[22] Tudhope, D., Koch, T., and Heery, R. 2006, *JISC Terminology service review*, http://www.ukoln.ac.uk/terminology/TSreview-jisc-final-Sept.html

[23] http://www.statemaster.com/encyclopedia/Uniform-resource-identifier

[24] http://itbimba.com/tutorials/xml.htm

[25] http://en.wikipedia.org/wiki/Topic_Maps

Z39.15: Z39.19 is a US-based national standard used to present guidelines and conventions for the contents, display, construction, testing, maintenance, and management of monolingual controlled vocabularies.

Z39.50: Z39.50 is a client-server protocol for searching and retrieving information from remote computer databases[26].

Zthes: Zthes is an abstract model for representing thesaurus terms, and made concrete by an XML format for representing thesauri according to this model.

[26] http://en.wikipedia.org/wiki/Z39.50