Project no.507618

DELOS

A Network of Excellence on Digital Libraries

Instrument: Network of Excellence

Thematic Priority: IST-2002-2.3.1.12
Technology-enhanced Learning and Access to Cultural Heritage

D5.3.1: Semantic Interoperability in Digital Library Systems

Original date of deliverable: 28th February 2005
Re-submission date: 29th June 2005

Start Date of Project: 01 January 2004
Duration: 48 Months

Organisation Name of Lead Contractor for this Deliverable: UKOLN, University of Bath

Revision [Revised Final]

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)

Dissemination Level: [PU (Public)]

Semantic Interoperability in Digital Library Systems

Authors: Manjula Patel, Traugott Koch, Martin Doerr, Chrisa Tsinaraki
Contributors: Nektarios Gioldasis, Koraljka Golub, Doug Tudhope

Task 3: Semantic Interoperability
WP5: Knowledge Extraction and Semantic Interoperability
DELOS2 Network of Excellence in Digital Libraries

July 2004 – June 2005

Revisions following Review of March 2005

Date	Version	Modifications
28-Feb-2005	Final	
15-Apr-2005		Review comments received
Apr-Jun-2005	Revised Final	Overview changed to Introduction and added subsections on 1.1 Audience, Scope and Purpose 1.2 Relevance to DELOS NoE 1.3 Report Structure Changed section 1 from Introduction to Background Added section 9 on Guidelines and Recommendations Section 8 on a Research Agenda rewritten Added the 2 new JPAII activities in WP5 to the Conclusions References amalgamated, sorted into alphabetical order and moved to end of document. Additional web links added.

Table of Contents

1. Introduction

This report is a state-of-the-art overview of activities and research being undertaken in areas relating to semantic interoperability in digital library systems. It has been undertaken as part of the cluster activity of WP5: Knowledge Extraction and Semantic Interoperability (KESI). The authors and contributors draw on the research expertise and experience of a number of organisations (UKOLN, ICS-FORTH, NETLAB, TUC-MUSIC, University of Glamorgan) as well as several work-packages (WP5: Knowledge Extraction and Semantic Interoperability; WP3: Audio-Visual and Non-traditional Objects) within the DELOS2 NoE.

In addition, a workshop was held [KESI Workshop Sept. 2004] (co-located with ECDL 2004) in order to provide a forum for the discussion of issues relevant to the topic of this report. We are grateful to those who participated in the forum and for their valuable comments, which have helped to shape this report.

Definitions of interoperability, syntactic interoperability and semantic interoperability are presented noting that semantic interoperability is very much about matching concepts as a basis. The NSF Post Digital Libraries Futures Workshop: Wave of the Future [NSF Workshop] has identified semantic interoperability as being of primary importance in digital library research.

1.1 Audience, Scope and Purpose

Although undertaken as part of the activities of WP5, the intended audience of this report is the whole of the DELOS2 NoE and the Digital Library community at large. In fact, many of the issues relating to semantic interoperability in digital library systems are also relevant to other communities. It is therefore, a major aim of the report to integrate views from overlapping communities working in the area of semantic interoperability, these include: semantic web, artificial intelligence, knowledge representation, ontology, library and information science and computer science. The types of issue that the report has tried to address include:

- Why is semantic interoperability important in digital library systems and how can it be used effectively in these types of information systems?
- An analysis of different types or levels of semantic interoperability
- A clarification of the relationship between syntactic and semantic interoperability
- A description of relevant methodologies, prerequisites, standards and semantic services
- How semantic interoperability in digital library systems can be enhanced
- What are the relevant issues for the DELOS2 NoE and the Digital Library community at large?

1.2 Relevance to DELOS2 NoE Objectives

Section 2 (Network Objectives) of the Technical Annex of the FP6 DELOS2 Network of Excellence identifies a 10-year grand vision for digital libraries:

> **"digital libraries should enable any citizen to access all human knowledge any time and anywhere, in a friendly, multi-modal, efficient and effective way, by overcoming barriers of distance, language, and culture and by using multiple Internet-connected devices."**

We consider semantic interoperability to be an essential technology in realising the above goal. The overall objective of semantic interoperability is to support complex and advanced, context-sensitive query processing over heterogeneous information resources. The report examines several areas in which semantic interoperability is important in digital library information systems, these include: improving the precision of search, enabling advanced search, facilitating reasoning over document collections and knowledge bases, integration of heterogeneous resources, and its relevance in the information life-cycle management process. The report also investigates some theoretical issues such

as clarification and selection of relevant terminology, standardisation and interpretation and the differing levels of semantic interoperability in digital library environments. It notes that information structure; language and identifiable semantics are prerequisites to semantic interoperability, as is consensus building and standardisation.

Another of the major objectives of the DELOS network is to integrate and coordinate the ongoing research activities of the major European research teams in the field of digital libraries for the purpose of developing the next generation digital library technologies. Once again, we envisage that semantic interoperability will be crucial to the next generation of digital library technologies, which in turn will be strongly influenced by semantic web technologies.

1.3 Report Structure

Following a general introduction to semantic interoperability and what we hope to achieve from it in digital library systems, we consider its importance in terms of contexts and the information life-cycle in Section 3; also looking at some relevant usage scenarios that have been developed by various projects. Section 4 is concerned with more theoretical issues including: terminology which is currently in use; the constituents of semantic interoperability; advantages and disadvantages of standardization and interpretation and three levels of semantic interoperability in digital library systems (data structures, categorical data and factual data).

We go on to consider some of the prerequisites to enabling and enhancing semantic interoperability in Section 5, these include: standards and consensus building; the role of foundational and core ontologies; knowledge organisation systems (KOS); the role of semantic services; architecture and infrastructure and access and rights issues. Section 6 investigates the methods and processes that are currently being used to improve semantic interoperability. This section falls into two subsections, the first examining standardization of metadata schemas, mediation and data warehousing, while the second covers methods which are being applied to KOS, their concepts, terms and relationships.

The emerging use of semantic interoperability in digital library services is the topic of Section 7, in which we consider how library services such as: searching, browsing and navigation; information tracking; user interfaces; and automatic indexing and classification are being enhanced and implemented to provide advanced user services. In Section 8 we have attempted to identify gaps and areas that would benefit from further research and attention. Section 9 identifies some guidelines and recommendations that would benefit the DELOS2 NoE in advancing semantic interoperability in digital library systems and thereby working towards its 10-year grand vision (see section 1.2).

2. Background

The Internet and more particularly the Web, has been instrumental in making widely accessible a vast range of digital resources. However, the current state of affairs is such that the task of pulling together relevant information involves searching for individual bits and pieces of information gleaned from a range of sources and services and manually assembling them into a whole. This task becomes increasingly intractable with the rapid rate at which resources are becoming available online.

Interoperability is therefore a major issue that affects all types of digital information systems, but has gained prominence with the widespread adoption of the Web. It provides the potential for automating many of the tasks that are currently performed manually.

Ouksel and Sheth identify four types of heterogeneity which correspond to four types of potential interoperability [Ouksel and Sheth 2004]:

- System: incompatibilities between hardware and operating systems
- Syntactic: differences in encodings and representation
- Structural: variances in data-models, data structures and schemas
- Semantic: inconsistencies in terminology and meanings

As far as digital libraries are concerned, interoperability is becoming a paramount issue as the Internet unites digital library systems of differing types, run by separate organisations which are geographically distributed all over the world. Federated digital library systems, in the form of co-operating autonomous systems are emerging in a bid to make distributed collections of heterogeneous resources appear to be a single, virtually integrated collection. The benefits to users include query processing over larger, more comprehensive sets of resources as well as the promise of easier to use interfaces that hide systems, syntax and structural differences in the underlying systems.

We define interoperability very broadly as any form of inter-system communication, or the ability of a system to make use of data from a previously unforeseen source. Interoperability in general is concerned with the capability of differing information systems to communicate. This communication may take various forms such as the transfer, exchange, transformation, mediation, migration or integration of information.

The main focus of our attention in this report is semantic interoperability in digital library systems, the goal of which is to facilitate complex and more advanced, context-sensitive query processing over heterogeneous information resources. This is an area that has been identified as being of primary importance in the area of digital library research by the recent NSF Post Digital Libraries Futures Workshop [NSF 2003].

Semantic interoperability is characterised by the capability of different information systems to communicate information consistent with the intended meaning of the encoded information (as intended by the creators or maintainers of the information system). It involves:

- the processing of the shared information so that it is consistent with the intended meaning
- the encoding of queries and presentation of information so that it conforms with the intended meaning regardless of the source of information

Furthermore, an aspect of semantic interoperability between two or more sets of data is a situation where the meaning of the entities or elements, their relationships and values can be established and where some kind of semantically controlled mapping or merging of data is carried out or enabled. The provision of this semantic information and the mapping or merging process determines the degree of semantic coherence in a given service. Consequently, there are different levels of semantic coherence or interoperability. For example, an information transfer system may carry or refer to the necessary semantic information, whereas a system that caters for the integration of information would accumulate information together using a specific mapping or merging effort.

Bergamaschi et al. identify two major problems in sharing and exchanging information in a semantically consistent way [Bergamaschi et al. 1999]:

- how to determine if sources contain semantically related information, that is, information which is related to the same or similar concept(s)
- how to handle semantic heterogeneity to support integration of information and uniform query interfaces

Some of the critical issues in this area relate to providing adequate contextual information, metadata and the development of suitable ontologies. Achieving terminology transparency has been the focus of attention of many mediated systems such as MOMIS –Mediated Environment for Multiple Information Sources [Bergamaschi et al. 1999] that provides a reconciled view of underlying data sources through a mediated vocabulary, which also acts as the terminology for formulating user queries.

Metadata vocabularies and ontologies are seen as ways of providing semantic context in determining the relevance of resources. Ontologies are usually developed in order to define the meaning of concepts and terms used in a specific domain. The choosing and sharing of vocabulary elements coherently and consistently across applications is know as *ontological commitment* [Guarino et al. 1994] and is a good basis for semantic interoperability in independent and disparate systems.

Although ontologies have been hailed as the answer to semantic interoperability, concerns are being raised about the sufficiency of (static) ontologies to resolve semantic conflicts, cope with evolving semantics and the dynamic reconciliation of semantics. It is a fact of life that database schemas and other types of information get out of synchronisation with the original semantics as they are used (and misused). How effectively semantic conflicts are resolved (this may need to be done dynamically) will directly affect the inferences and deductions that are performed in answering a user query –and ultimately the results that are returned. Cui et al. describe a system that addresses some of these types of issues [Cui et al. 2002]. DOME –Domain Ontology Management Environment, provides software support for the definition and validation of formal ontologies and ontology mappings to resolve semantic mismatches between terminologies according to the current context. Further, Gal describes a system, CoopWARE [Gal 1999], which investigates issues relating to services that need to adapt ontologies to continuously changing semantics in sources. The semantic model is based on TELOS [Mylopoulos J. et al. 1990] making it flexible enough to support dynamic changes in ontologies.

In this report we have tried to cover the major issues that relate to semantic interoperability in digital library systems. Section 3 begins by looking at the importance of semantic interoperability in terms of providing context; knowledge life-cycle management and some use cases. Section 4 tackles some of the more theoretical and formal issues. Prerequisites, methods and processes to enhance semantics in library information systems are dealt with in sections 5 and 6. Section 7 is concerned with the role of semantic interoperability in digital library services such as searching, browsing and navigation; information tracking; user interfaces; and automatic indexing and classification. We conclude the report by identifying gaps and trying to recommend areas that would benefit from further attention as well as highlighting ways in which the DELOS2 NoE should maximise the use of semantic interoperability in the development of the next generation of digital library systems.

3. Importance of Semantic Interoperability in Digital Libraries

To understand the importance of semantic interoperability in digital libraries, we need to look at different contexts such as the traditional context of subject indexing and access, the integration of heterogeneous information sources in the digital world of the Internet and the context of improvements to the Information Life-Cycle Management. The importance of semantic interoperability in many elements of the information life-cycle is exemplified. Semantic interoperability is also important to many different communities and disciplines beyond Digital Libraries and many of them develop related activities. A suite of use cases from several major projects illustrates the breadth of applications where the need to integrate heterogeneous sources requires efforts to improve semantic interoperability which, in addition, often imply economic benefits.

3.1 Contexts

"Semantic is the key issue in order to solve all heterogeneity problems". (Visser 2004)

Interoperability is an important issue in all information systems and services. Without syntactic interoperability, data and information cannot be handled properly with regard to its formats, encodings, properties, values, and data types etc., not merged nor exchanged. Without semantic interoperability, the meaning of the used language, terminology and metadata values cannot be negotiated or correctly understood.

Interoperability is an important economic issue as well, as Dempsey [Dempsey, ARLIS 2004] points out: it is necessary to be able to extract a maximum value from investment in metadata, content and services by ensuring that they are sharable, reusable and recombinable. The improved services will allow users to focus on the productive use of resources rather than on the messy mechanics of interaction.

Semantic interoperability has shown its importance in several different contexts, communities and disciplines.

a) It has been important for a long time in the "traditional" context of subject indexing and access, to support search with some degree of semantic precision. Knowledge Organization Systems (KOS) have been used for a long time by for example abstract and index database providers to index the content of publications, databases and similar and to support subject searching. Online database hosts like Dialog and Silverplatter have quite a long time ago, in addition, started to offer cross-search solutions for subject access to databases hosted by them. Multilingual (with the help of the development of multilingual vocabularies) and multidisciplinary/cross domain search has been addressed as well.

b) A second and more recent context is the need for integration of heterogeneous, often distributed, information sources that became increasingly possible and requested with the development of the Internet.

The Internet, Intranets and in general information networks allow the reference to, usage of and integration of highly heterogeneous information sources plus the creation of new information out of many different sources (Data warehouses, Electronic Data Interchange, Business-to-Business, Peer-to-Peer architectures, Knowledge management systems, Digital Library services, eGovernment services etc.).

The resource environment e.g. in an university setting, became greatly expanded: besides library catalogues and abstract and index databases, digitized collections, licensed collections, remote preprint archives, institutional repositories, e-reserves, virtual reference, new scholarly resources, learning objects, web-based information and publications, subject gateways etc. became available and needed to be integrated in one seamless information space for the user. Information discovery here requires to be able to navigate across many sources by subject, by name, by place, by resource type or by educational level, with as little custom work, as little pre-coordinated agreement and as little terminological investigation as possible [Dempsey, ARLIS 2004]. A. Sheth describes even more advanced scenarios of navigation by following relationships that emerge dynamically from the integration of resources, rather than by elements common to the various resources.

A solution, a semantically well integrated digital library service, could be tried to be implemented as either a more or less centralised integrated and interoperable information service or as a "recombinant" library [Dempsey 2003] based on distributed and independent services and sources (e.g. based on a Web Services architecture): highly specialised presentation, application and content services, supported by common services would be made to cooperate.

Semantic interoperability would be enabled by terminology services, earlier integral with a particular search service, now potentially externally provided as third-party services. Even a federation of distributed, multilingual, formalised and enriched KOS could be offered as one such service.

c) A third context for Semantic Interoperability activities is improvement to the so-called information life-cycle management. Among other applications, this is often used to structure work in corporate Knowledge Management (see Section 2.2 below).

Semantic interoperability is important in many different communities and disciplines far beyond the sector of Digital Libraries, which is in the focus of this report. Among those are the Government, Museum, Educational and Corporate or Business sectors, already well known with regard to ambitious efforts.

Concerning Government information systems and eGovernment initiatives a recent "eGovernment Workshop on semantic interoperability" in Norway documented good practice and European cooperation (co-organised by the EU Commission) [eGovernment 2004]. Semantic interoperability based on data definitions and identification of data and metadata, are seen as an important prerequisite to eGovernment services such as: electronic interfaces between the business community and the public

sector facilitating exchange of data (e.g. for tax administration, statistics); common repositories of metadata and one-stop-shop web-based services for citizen access to data and metadata.

Semantic interoperability based on conceptual understanding of the shared information, data and knowledge interpretation, ontologies and agents, reconciliation methods and modelling of processes, is mentioned as a key element of EIF: the European Interoperability Framework [EIF 2004]. EIF is a set of standards and guidelines which describe the way in which organisations have agreed, or should agree, to interact with each other complementing national interoperability guidance by focusing on a pan-European dimension.

OntoGov is a recently started EU 6. F.P. STREP project dealing with semantics for life-cyle design of public services, tool development and the creation of a related domain ontology [OntoGov 2004].

Activities on a national level include: UK GovTalk and E-GIF which provide interoperability and metadata standards via e.g. a Government Category List, a Government Schemas Working Group and an Interoperability Working Group [GovTalk]; a portal of the Walloon Region applying a semantic web approach for interoperability. Leading terminological efforts for the support of Semantic Interoperability are carried out by Canadian [Canada] and Australian government agencies. Considerable semantic interoperability related efforts are undertaken in the GovStat project for the US Bureau of Labor Statistics [Efron et al 2004].

3.2 Information Life Cycle Management

Semantic interoperability is not only important in the "traditional" contexts of subject indexing and subject access to databases and documents, or when integrating heterogeneous information sources for the purpose of information discovery. It seems relevant in most of the stages of the so-called information life cycle.

3.2.1 Terminology

In the literature the terms information life cycle (management) and knowledge life cycle (management) are often used to represent the same concept just as the term knowledge is often used to refer to information.

One should distinguish between data, information and knowledge. According to D. Soergel [Soergel 1985, chapter 2], data is the form and information is the content, whereas knowledge has structure that ties together and integrates individual pieces of an image of the state of affairs and is the basis for action (he explains the nature of information through a cycle of: image - image of the state of affairs - new information based on observation, interpretation - updated image - action). On the other hand, M. Buckland [Buckland 1991] distinguishes between three categories of information, information-as-process, information-as-knowledge, and information-as-thing. He claims that information systems can deal with objects like documents or their representations, which are things, and not with processes or knowledge. Process in this context is the act of informing, and knowledge is what is perceived and communicated in that process.

We use here the term information life cycle management because it is data or information and not knowledge that have actually been incorporated in the so-called knowledge life cycles.

3.2.2 Models

There are many such models, each belonging to a different context and aiming at a different purpose; e.g. there are many life cycle models focusing on storage (the discipline of Computer Science), on information seeking (the discipline of Library and Information Science), information life cycles in companies and governments (the discipline of Knowledge Management) etc.

Information life cycle management models are just theoretical models, while the practical application is a more complicated process. As C. Borgman [2000, p. 108-109] says: the " ...cycle of creating, using, and seeking information can be viewed as a series of stages, but these stages often are iterative. People move back and forth between stages, and they may be actively creating, using and seeking information concurrently. People tend to manage multiple information-related tasks, each of which may be at a different stage in the cycle at any particular time."

We describe two information life cycle management models here, one from the Knowledge Management, and the other from the Library and Information Science community.

Although the one from Knowledge Management [Shadbolt et al. 2003] is called knowledge life cycle management, it is actually dealing with information as we define it (see above). This model is from a report by Advanced Knowledge Technologies Interdisciplinary Research Collaboration (AKT), focusing on tools and services for managing knowledge throughout its life cycle. Their model comprises six challenges, and serves as a means to classify AKT services and technologies.
They are:

- acquisition,
- modelling,
- reuse,
- retrieval,
- publishing and
- maintenance.

For the different stages of the life cycle, specific "knowledge technologies" are expected to be fruitful. When discussing acquisition, their focus is on harvesting of ontologies from unstructured and semi-structured sources. In modelling they deal with modelling life cycles, and the coordination between Web services, as well as with mapping and merging of ontologies. Reuse refers to reuse of Web services via brokering systems and their experiments in mediating between problem solvers via partially shared ontologies. In the retrieval stage they focus on the transition from informal to formal media. In the publishing stage they demonstrate how formally expressed knowledge may be made more personal. The maintenance stage refers in their case to tools that respond to changes of language use in an organization over time.

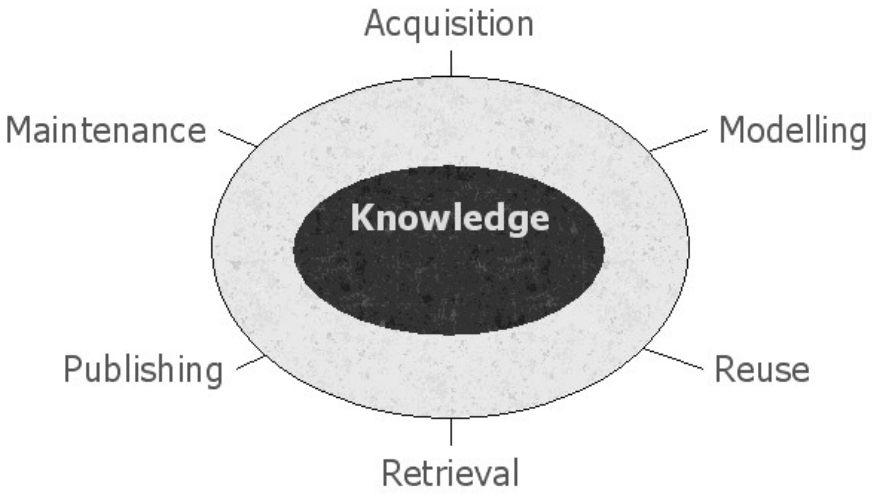

Figure 1: AKT's six knowledge challenges

The other is a library and information science approach, as described in G. Hodge [Hodge 2000]. The focus of her paper is on digital archiving based on an information life cycle approach. She identifies best practices for archiving at all stages of the information management life cycle:

- creation,
- acquisition,
- cataloguing and identification,
- storage,
- preservation and
- access.

Creation is the act of producing the information product. Acquisition is related to collection development, and the two represent the stage in which the created object becomes part of the archive or the collection. Identification provides a unique key for finding the object and linking that object to other related objects. Cataloguing is important for organization and access. Storage is a passive stage in the life cycle, although G. Hodge reminds us that storage media and formats have changed over time, which caused some information to be lost maybe forever. Preservation refers to preserving the content as well as the look and feel of the object. Access needs to be ensured and enabling it comes as a result of the previous stages.

Those two models overlap to a certain degree. Thus acquisition is an element of both life cycles. Preservation is a narrower term to maintenance, and access is a broader term to retrieval. In addition, taken together they might not be complete even for one given context.

3.2.3 Importance of Semantic Interoperability

Semantic interoperability issues seem relevant in each of the elements from the following extended list of information life cycle elements:

1 Creation, modification

2 Publication

3 Acquisition, selection, storage, system and collection building

4 Cataloguing (metadata, identification/naming, registration), indexing, knowledge organisation, knowledge representation, modelling

5 Integration, brokering, linking, syntactic and semantic interoperability engineering

6 Mediation (user interfaces, personalisation, reference, recommendation, transfer etc.)

7 Access, search and discovery

8 Use, shared application/collaboration, scholarly communication, annotation, evaluation, reuse, work environments

9 Maintenance

10 Archiving and preservation

In this information life cycle creators/authors, publishers, information systems managers, service providers and end-users are involved.

In the elements 5, 4 and 7, semantic interoperability activities seem most important, whereas these issues in the elements 2, 3, 9 and 10 are clearly less relevant.

1 When creating or modifying information, terminological resources can assist the creator in higher quality and better clarity of expression and, if done at this stage, assist in metadata generation for documents and objects. Semantic interoperability measures could support a more controlled terminological development in and between disciplines and communities. Creators/authors contribute to terminology development as well and thus help in developing and keeping up-to-date the vocabularies they might use as authoring or searching support.

2 Publication might involve tasks mentioned under 1 Creation and under 4 Cataloguing etc. Publishers might provide activities mentioned under 1, 3-6, 9, and 10, similarly to other actors like libraries and other memory institutions and intermediaries.
The phase of publication in the life cycle alone has no major benefit from semantic interoperability operations, other than including semantic and subject information, in the best of cases from controlled vocabularies, into the published documents.

3 Acquisition and collection building might use subject vocabularies and other semantic information when deciding about the inclusion of a document/object into a service or specific collection. Even automated procedures like harvesting need them to build subject specific collections (e.g. topical crawling). Interoperability activities allow including documents with equivalent but not identical terminology.

6 Information mediation activities can provide higher quality when based on semantically interoperable vocabularies, either operating in the background or offered to the user for navigation and support. Adaptive user interfaces can work with the users and translate to their own specific vocabulary and involve semantically corresponding information based on e.g. mapped vocabularies. They can provide different views of their resources and, thus, assist in personalisation as well as in reference and information transfer.

8 Terminological resources and semantic interoperability measures will improve use and evaluation of information. A greater benefit occurs in the scholarly communication and collaboration process with all related activities, again, clarifying the meaning of terms and concepts and their relationships.

9 Maintenance of information has to care for the links to related semantic information and vocabularies and to make sure that they are available as well. Version control and history/scope notes or similar are needed to preserve the terminology and meaning of documents to keep them understandable and to be able to follow changes over time.

10 Preservation of information includes taking care of the semantic information; the vocabularies, methods and tools, and not only the "raw" information of the documents (cf. 9). It may have extraordinary high requirements for interoperability with technologies expected to be available for a long time.

For the life cycle elements 5, 4, 7, semantic interoperability activities are of outmost importance. 5: Information integration, brokering etc. can basically not be carried out properly without such efforts. This whole report, including most of the use cases in Section 2.3, cover these three elements and provide examples from their realm.

Since vocabularies, semantic relationships and mappings are information (objects) as well, their life cycle: creation, acquisition, collection, modelling, identification, integration, mediation, search, use, maintenance and preservation etc. is of primary importance and a necessary prerequisite to improved semantic interoperability in all information life cycle contexts.

3.3 Use Cases

Every service trying to assemble data for cross search or to integrate data from (semantically) heterogeneous sources needs to address problems of semantic interoperability. This seems today to be the most frequent case in the digital information environment.

Here are a few short example use cases with some references to use cases developed in related projects and initiatives.

a) A web site gathers pages and documents from different research groups and sources. A controlled list of terms applying at least synonym control is needed to allow good search results (sufficient recall). The aim may be to enable cross-searching of databases, aggregation of the data contained within them, or the construction of software tools to present the information to the end-user, for example in a portal or Virtual Learning Environment.
Ex.: any kind of portal, OAI repository (e.g. arc), institutional repository system (DSpace, eprint etc.), even personal digital libraries.

Ex.: Use case Food Safety: Antimicrobials online A/OL [from SEMKOS]
A/OL is a web-based system that facilitates the dissemination of knowledge on preservation of food, extracted from scientific papers. Microbiology experts have assessed a large number of publications that contain data and results of experiments on microbial compounds. They extract the information that is usable in industrial applications, in particular on the effectiveness of the compounds in real food systems. The project is funded by the EU, the Dutch Ministry of Agriculture and DSM (multinational company producing food ingredients). In order to support the necessary quality of search in this new online database an ontology has been developed which needs extension and maintenance.

b) In order to provide for some precision in searching the same site, other semantic relationships between the key terms need to be specified, e.g. in a hierarchical thesaurus or a formal ontology. Dictionary, glossary or encyclopaedia information might be needed to clarify the exact meaning of the terms. A more advanced feature of search term expansion might be required.
[cf. SWAD-E use cases: Extend JISC-funded Subject Portal Project cross-search]

Both a) and b) may involve the creation and maintenance of general or domain specific vocabularies or the identification of suitable vocabularies, possibly to be merged and the markup/annotation/indexing of documents or metadata records with the new terms. [cf. SWAD-E use cases; OWL use case Web portals, Multimedia collections]
Authority Control Service [cf. SIMILE use case. OCLC service to provide subject classification and authority control to EPrints UK and DSpace: OCLC Metadata Switch Project].
Registries of (RDF) schemas/vocabularies (cf. SIMILE use cases).
Data mining and automated annotation on unstructured information, like journal abstracts [cf. SIMILE use cases].

SEMKOS Taxonomy use case, James Brooks [SEMKOS]:
"So, imagine a future scenario, perhaps for an abstracting & indexing service, or for annotation of PDFs, or for on-the-fly search engines using semantic expansion of a query:
My document is automatically indexed, the scientific names in it are, firstly, identified/recognized and, then, checked against recognized authority files. Spelling mistakes are recognized and simply corrected. After an invalid scientific name the correct name is added, say, in brackets OR added to an index OR mapped on the fly to the searched-for term, etc. If not as simple as this the document/resource is flagged to the attention of an editor. There could be automated addition of valid names to a master authority file OR one could automatically validate the status of names appearing in a thesaurus, so they are current. One could imagine software agents interrogating appropriate resources, mediating from place of description/validation/revision/detail addition to validated checklists/catalogues to index authority files, etc.
Now, there is often no single place that can be referred to, to validate a particular name for a particular (species of) organism. There are competing hierarchical schemes. There is often incomplete integration

between catalogues for the same taxa in different geographical regions. A problem for indexing resources, and the subsequent retrieval of them by other users, is that of what is the correct name to use. Thus, an organism may be known, in any one language, by more than one common name. Scientific nomenclature is not stable; historically, more than one name has often been applied to the same organism, in which case rules of precedence apply; alternatively, there are changing ideas as to what constitutes the limits of a particular genus, and new combinations may result when genera are split or when a species placed in one genus is shown to have greater affinities with another. Not to speak of common names. Integration of taxonomy and object databases (e.g. in Natural History Museums) is necessary."

c) Mapping and/or translation of terms is required if the improved search is to be applied to documents in different languages (bi- and multi-lingual collections). [cf. SWAD-E use cases: Multilingual image retrieval]

d) In order to allow browsing of structured collections and sub-collections of documents a (hierarchical) structure of categories (classification system) needs to be applied or built. This would allow exploration or search of all appearances of e.g. dogs on the site without having to specify each and every individual member of the family using all potential variants of names.
Ex.: Collections of (metadata on) journal articles [OWL use case Design documentation]

e) In the case of collections containing documents from or relating to different historical periods and/or with different political and cultural views, e.g. about Poland in the 18th Century, semantic interoperability measures identifying the different contexts and providing the appropriate vocabulary with relationship information (Poland's extension, boundaries, legal status at different times, seen by different players) are required (example from Kim Veltman). A similar task is the mapping of different views from different user groups and their specific vocabularies [OWL use case Corporate web site management].

f) If well structured documents with different origins are gathered, semantic interoperability operations need to be applied in two steps: approaches described above need to be preceded by deciding the equivalence of the semantic definition of data buckets/ metadata elements/ bibliog. fields/ properties/ attributes/ tags or similar with subsequent mapping, joining or splitting decisions.
Ex.: the standardisation of different metadata profiles into one common application profile as done in project Renardus to allow common cross- searching.
Ex.: Government metadata interoperability initiatives (cf. Section 2.1)

[from the SEMKOS Ontologies use case]:
Databases may have their meaning locked in textual documentation or in the structural format of the database, for example in the tables and keys of a relational database. Consistent access to Knowledge Organisation Systems requires advice, tools and standards that enable KOS owners to

- extract the semantics and the data model of data sets
- make this information available in a consistent, standards-based format
- provide mechanisms to access this information via the network

g) In case the documents or their metadata use values drawn from different KOS -or, as DCMI Grammar principles say- from different vocabulary encoding schemes (e.g. Library of Congress Subject Headings, North American Industry Classification System NAICS), these can be formally mapped (or linked via crosswalks) to improve Semantic Interoperability.
Ex.: Classification mapping in project Renardus to allow cross-browsing based on one common (switching) classification.
[cf. SWAD-E use cases: Extend JISC-funded Subject Portal Project cross-search]
[cf. SWAD-E use cases: Bized/SOSIG trials of data integration via classification mapping in the DESIRE project]
Ex.: Government interoperability work: Canada, e-GIF UK, EU (see Section 2.1)

Ex.: Geo-spatial integration via mapping of geographic names, places, regions, feature types etc.: Alexandria Digital Library.

The occurrence of different syntax encoding schemes (e.g. date string formatted in accordance with a certain formal notation, according to DCMI) requires rather a measure of assuring syntactic interoperability, a conversion into one common format.

These use cases could be written from the perspective of different actors or corresponding use cases could be added for:

- information creators (authors and metadata creators of documents, businesses, government agencies)
- information/service providers, intermediaries; machine-to-machine (vocabulary builders, web portal builders, collection organisers, indexers - assuming manual indexing, publishers of documents and databases, government agencies)
- end users trying to discover relevant information (information searchers, citizens, decision makers (decision support), educational needs, industrial contexts, electronic commerce)

All the mentioned measures imply to deal with and improve semantic interoperability. Approaches to be used are creating, extending, revising, maintaining, identifying, sharing, representing, using, syndicating, translating, mapping or merging vocabularies and applying them to the information system in order to allow human users (or machines) to improve the quality of their information discovery and search.

The economic value of improved Semantic Interoperability is described in the following use case:

Pest species in biological control [from: SEMKOS use case, Application B.3 page 18]
"Natural History Museums can be considered as archives of biodiversity. They harbour millions of specimens, providing first-hand information about geographical and historical presence of organisms, their characteristics, ecological environments, host-parasite relationships etc. The classification of these specimens/organisms is based on a taxonomic KOS, in which scientific names are internationally standardized, but which are fragmented and exist in many versions.

An interesting example taken from Systematics Agenda 2000 elucidates the economic importance of a reliable taxonomic identification of a pest species in biological control: In 1974 an introduced mealybug species was discovered in Zaire. This pest was costing cassava growers in West Africa nearly 1,4 billion dollars in damage each year. The species was described as Phenacoccus manihoti and a search was made for biological control agents in northern South America. When no effective parasites was discovered, a mealybug specialist was asked to re-examine the species and found that a closely related species, P. herreni, was found primarily in northern South America and that P. manihoti actually occurred further south. With this information, effective parasites were located and introduced into the infested areas of Africa.
The current reality of systems, that return on a question about spiders mostly Spiderman literature, is obviously inadequate."

4. Theoretical Considerations

As outlined in the previous chapters, the achievement of semantic interoperability is a complex task, which affects multiple levels and functions of information systems and the information process. In this section, we propose a systematic requirements analysis of the different constituent functions necessary to achieve overall semantic interoperability in digital library environment. We begin with a clarification of terminology that tends to be inconsistently used between the computer science and the libraries' community.

4.1 Terminology

We propose here one possible definition for each term for the purpose of this document. An exhaustive survey e.g. of definitions of ontology may be interesting but not very useful in order to make the intended meaning of this report more clear to the reader.

4.1.1 Universals and Particulars

From a knowledge representation perspective, concepts can be divided into **universals** and **particulars.** The fundamental ontological distinction between universals and particulars can be informally understood by considering their relationship with instantiation: particulars are entities that have no **instances** in any possible world; universals are entities that do have instances. **Classes** and **properties** (corresponding to predicates in a logical language) are usually considered to be universals. [after Gangemi et al. 2002, pp. 166-181]. E.g., Person or A being married to B are universals. John, Mary and John is married to Mary are particulars. General nouns and verbs of a natural language can be regarded to describe universals (polysemy not withstanding), whereas names describe particulars, [Steven Pinker 1994] describes the distinction of general nouns, verbs and proper names as innate functions of the human brain.

4.1.2 Ontology and Vocabulary

We follow here the definition of [Guarino 1998]:

*An ontology **is a logical theory** accounting for the intended meaning of a formal vocabulary, i.e. its ontological commitment to a particular conceptualization of the world. The intended models of a logical language using such a vocabulary are constrained by its ontological commitment. An ontology indirectly reflects this commitment (and the underlying conceptualization) by approximating these intended models.*

Guarino further defines a model as a description of a particular state of affairs, a world structure, whereas the conceptualisation describes the possible states of affairs or possible worlds of a domain consisting of individual items. Further, particular states of affairs are seen as instances (extension) of the conceptualisation. The ontology only approximates the conceptualisation, because its logical rules may not be enough to define all constraints we observe or regard as valid in the real world. In this sense, the formal vocabulary is a **part** of the ontology, but not an ontology in itself, which **is** a logical theory. The symbols of this vocabulary would normally refer to universals, as do the nouns and verbs of natural languages. Following this definition, a gazetteer is not an ontology, because it describes a particular world structure. A simple thesaurus which uses the broader term generic relationship [ISO2788] in the sense of IsA between concepts (universals) can however be regarded as a very simple form of an ontology. A controlled vocabulary clearly does not qualify as an ontology, but could be used to create an ontology [Qin, Jian & Paling, Stephen, 2001]. As the extent of formalization is not defined, there are varying opinions from which point on a terminological system qualifies as ontology.

4.1.3 Language and Vocabulary

A vocabulary is a set of symbols. A language consists of a vocabulary and a grammar that defines the allowed constructs of this language. A vocabulary alone does not qualify as a language.

According to http://www.wordiq.com: In mathematics, logic and computer science, a formal language is a set of finite-length words (i.e. character strings) drawn from some finite alphabet.

In computer science a **formal grammar** is an abstract structure that describes a formal language precisely: i.e., a set of rules that mathematically delineates a (usually infinite) set of finite-length strings over a (usually finite) alphabet. Formal grammars are so named by analogy to grammar in human languages.

Formal grammars fall into two main categories: *generative* and *analytic*.

- A generative grammar, the most well known kind, is a set of rules by which all possible strings in the language to be described can be *generated* by successively rewriting strings starting from a designated start symbol. A generative grammar in effect formalizes an algorithm that *generates* strings in the language.

- An analytic grammar, in contrast, is a set of rules that assume an arbitrary string to be given as *input*, and which successively *reduce* or *analyze* that input string to yield a final Boolean, "yes/no" result indicating whether or not the input string is a member of the language described by the grammar. An analytic grammar in effect formally describes a parser for a language.

In short, an analytic grammar describes how to *read* a language, whereas a generative grammar describes how to *write* it.

Note, that the term **alphabet** in the above is synonymous to the term **vocabulary**, and **not** to what normal people regard as an alphabet, and the term **word** in the above is synonymous to the term **phrase**, and **not** to what normal people regard as a word. This is enough reason for confusion. Therefore, the above must be read for normal people: A formal language is a set of possible, finite-length phrases, and not a vocabulary.

The linguist Noam Chomsky offers this definition of human language: First, he says that human language has structural principles such as grammar or a system of rules and principles that specify the properties of its expression. Second, human language has various physical mechanisms of which little is known but it does seem clear that "laterization plays a crucial role and that there are special language centres, perhaps linked to the auditory and vocal systems". The third quality of human language is its manner of use. Human language is used for expression of thought, for establishing social relationships, for communication of information and for clarifying ideas. Another characteristic of human language is that it has phylogenetic development in the sense that language evolved after humans had separated from the other primates. Therefore language must have had a selective advantage and must coincide with the proliferation of the human species. Finally, human language has been integrated into a system of a cognitive structure [Chomsky 1980, cited after Britta Osthaus].

Normally, a language also commits to the intended meaning of its symbols and constructs. In contrast to the ontology, it aims at enabling descriptions of states of affairs without intention to approximate the possible worlds. So, phrases like “my dog is a cat” or “the ship rains under the mountains”, are perfect English but violate our conceptualisation.

A suitable logical language, such as OWL, TELOS, KIF, RDF/S etc., allows for describing models of a particular state of affairs as instances of concepts defined in a formal ontology. Then, this language together with the vocabulary of the ontology can be seen as a specific language to describe valid models of this ontology.

4.1.4 KOS and Vocabulary

The term Knowledge Organization Systems (KOS) refers to controlled vocabularies as well as to systems/tools/services developed to organise knowledge (tools that present the organized interpretation of knowledge structures [Zeng 2004]). For the purpose of this report, we distinguish the contents, i.e. the vocabulary and the associated logical relationships, KOS in the narrower sense, from the software that may present the content. Some may require for a KOS to implement some logical structure, we will use it however in the context of this report for all kinds of knowledge organized as reference for use in information systems: from simple term lists up to taxonomies, [see definition in NKOS 2000; HILT 2001, App. F: Glossary]; Classification systems and elaborated ontologies on the side of the universals; Authority files and Gazetteers on the side of particulars.

In this sense, uncontrolled vocabularies, i.e. term lists without an organized editorial control, and controlled vocabularies, i.e. term lists with an organized editorial control are regarded special cases of KOS, and all kinds of KOS that deal with universals must contain a vocabulary in the narrower sense. We propose to distinguish proper names, such as place names or names of persons from terms due to their different role and nature. Even though the term “vocabulary” is frequently used as a synonym for KOS as we define it here, we propose not to use it in this sense because of the many ambiguities its use introduces with respect to other senses. In particular it may lead to a confusion of part and whole.

From an environmental perspective we define an “NKOS” as: Networked KOS, interactive information devices published in digital format. These are primarily aimed at supporting the description and retrieval of heterogeneous information resources on the Internet [Zeng 2004].

4.1.5 Schema, Data Model and Conceptual Model

The term schema typically stresses the structural aspect and even storage format. With more modern DBMS, the actual physical format is more and more hidden and irrelevant to the designer, e.g. the on-line dictionary SearchDatabase.com (http://searchdatabase.techtarget.com) writes:

“In computer programming, a schema (pronounced SKEE-mah) is the organization or structure for a database. The activity of data modelling leads to a schema. (The plural form is *schemata*. The term is from a Greek word for "form" or "figure.") The term is used in discussing both relational databases and object-oriented databases. The term sometimes seems to refer to a visualization of a structure and sometimes to a formal text-oriented description.”

Typically, the term schema is used to relate to the data structure as implemented, and not so much to refer to its intended meaning, in particular the meaning it has for real world described by instances of the schema. We prefer as a more general term data structure, defined in the same source as:

“A data structure is a specialized format for organizing and storing data. General data structure types include the array, the file, the record, the table, the tree, and so on. Any data structure is designed to organize data to suit a specific purpose so that it can be accessed and worked with in appropriate ways. In computer programming, a data structure may be selected or designed to store data for the purpose of working on it with various algorithms.”

From the point of view of standardization and semantic interoperability, this term makes a relevant abstraction from the internal organization of documents, metadata and databases.

Whereas computer science traditionally uses the term data model for the schema definition constructs, such as entity-relationship (E-R), XML DTD, others use it as product of the activity of data modelling, i.e. synonymous to schema. We propose to avoid the term. Use instead schema or conceptual model as appropriate.

On the other side, a conceptual schema is typically referred to as a map of concepts and their relationships. The difference between a conceptual schema and an implemented schema is typically in the omitting of data elements for the control of the information elements in the database such as keys, oid, locking flags, timestamps, etc., as well as in the explicit reference to real world concepts referred to by the schema constructs. The term conceptual model comes even closer to a logical formulation of the possible states of an application domain, so that normally a conceptual model can be regarded as a kind of ontology.

In many cases a data structure, abstract from its use to specify a storage layout, can also be seen as a special case of formal language to make statements about particular states of affairs. Its elements (fields, tables etc.) constitute a formal vocabulary, such as the Dublin Core Element Set. Similarly, an ontology can be used to define such a formal language, and hence data structures (such as the RDFS version of the CIDOC CRM.

However, this argument should not be used to regard the field names of a metadata structure as a kind of ontology. Most data structures do not qualify as ontologies, as data structure element definitions lack any formal approach to approximate a conceptualisation, e.g. the field *publisher* in DCES can be interpreted in at least three ways. Whereas concepts of an ontology are meaningful out of the context of a data structure, field names typically make sense only in the specific element hierarchy or connection. We regard the natural language interpretation of the gibberish of field names out of context as generally misleading, e.g. a field age in the CIDOC Relational Model has to be interpreted as stage of maturity of the referred art object; a field destination in the MIDAS schema at English Heritage is interpreted as destination of a wrecked ship on its last mission.

Metadata structures, often called metadata vocabularies or metadata frameworks should be regarded in the first place as schemas or conceptual schemas. Only in some cases may they be regarded as direct derivatives of an ontology.

4.1.5 Mapping and Crosswalks

The last term to be described here is the concept of schema mapping.

Semantic World defines mapping as: "the process of associating elements of one set with elements of another set, or the set of associations that come out of such a process. Often refers to the formally described relationship between two schemas, or between a schema and a central model." (http://www.semanticworld.org).

In the metadata community, the term crosswalk became fashionable:

A crosswalk is a semantic and/or technical mapping (sometimes both) of one metadata framework to another metadata framework.

Semantic mapping example :

- Dublin Core element title corresponds to the ADN element of title
- Dublin Core element type corresponds to the ADN element of learning resource type

Technical mapping example

- Technical mapping uses various programmatic solutions to transform metadata records or computer files. DLESE uses eXtensible Stylesheet Language Transform (XSLT) to programmatically change eXtenisble Markup Language (XML) metadata records to other formats. For example, the following shows Dublin Core XML elements and their corresponding ADN XML elements, http://www.dlese.org/Metadata/crosswalks/

Obviously both refer to the same process. What is called above as semantic mapping is lately in computer science also referred to as schema matching, whereas mapping implies the actual transformation algorithm. We prefer this definition:

"A schema mapping is the definition of a transformation of each instance of a data structure A into an instance of a data structure B that preserves the intended meaning of the original information and that can be implemented by an automated algorithm. The application domain expert ultimately judges the preservation of the intended meaning. A partial mapping may lose a clearly defined part of the original information. The actual schema map, i.e. the product of a mapping process, may also be called a *functor* for the data translation process."

We prefer the term schema matching to semantic mapping, because any mapping should be semantically correct.

4.2 Constituents of Semantic Interoperability in Digital Library Environments

In order to make the following distinction more obvious, let us regard an artificial, but realistic demonstration case about the integration of information objects related to the Yalta Conference in February 1945. This was the event officially designating the end of WWII. One can hardly find a better-documented event in history. We have created the demonstration metadata below from the information we found associated with the objects. The titles are as we have found them. The scenario is about how to make these information objects accessible by one simple request.

a) The State Department of the United States holds a copy of the Yalta Agreement. One paragraph begins, The following declaration has been approved: The Premier of the Union of Soviet Socialist Republics, the Prime Minister of the United Kingdom and the President of the United States of America have consulted with each other in the common interests of the people of their countries and those of liberated Europe. They jointly declare their mutual agreement to concert [http://www.fordham.edu/halsall/mod/1945YALTA.html].

A Dublin Core record may be:

Type: Text
Title: Protocol of Proceedings of Crimea Conference
Title: Declaration of Liberated Europe
Date: February 11, 1945.
Creator: The Premier of the Union of Soviet Socialist Republics
Creator: The Prime Minister of the United Kingdom
Creator: The President of the United States of America
Publisher: State Department
Subject: Postwar division of Europe and Japan
Identifier: ...

b) The Bettmann Archive in New York holds a world-famous photo of this event (Fig. 1). A Dublin Core record for this photo might be:

Type: Still Image
Title: Allied Leaders at Yalta
Date: 1945
Publisher: United Press International (UPI)
Source: The Bettmann Archive
RightsHolder: Corbis
Subject: Churchill; Roosevelt; Stalin

Figure 1: Allied Leaders at Yalta

The striking point is that both metadata records have nothing more in common than 1945, hardly a distinctive attribute.

c) An integrating piece of information comes from the Thesaurus of Geographic Names [TGN, http://www.getty.edu/research/tools/vocabulary/tgn/index.html], which may be captured by the following metadata:

TGN Id: 7012124*Names:* Yalta (C,V], Jalta (C,V)
Types: inhabited place(C), city (C)
Position: Lat: 44 30 N,Long: 034 10 E*Hierarchy:* Europe (continent) < Ukrayina (nation) < Krym (autonomous republic)*Note:* Located on S shore of Crimean Peninsula; site of conference between Allied powers in WW II in 1945; is a vacation resort noted for pleasant climate, & coastal & mountain scenery; produces wine, canned fruit & tobacco products.
Source: TGN, Thesaurus of Geographic Names

This could be, at least partly, formatted in DC as well:

Identifier: http://www.getty.edu/research/tools/vocabulary/tgn/7012124
Title: Yalta (C,V]

(Title) Alternative: Jalta (C,V)
Type: inhabited place(C)
Type: city (C)
(Coverage) Spatial: Lat: 44 30 N, Long: 034 10 E [DCMI encoding scheme=Box]
Source: TGN, Thesaurus of Geographic Names
Description: Located on S shore of Crimean Peninsula; site of conference between Allied powers in WW II in 1945; is a vacation resort noted for pleasant climate, & coastal & mountain scenery; produces wine, canned fruit & tobacco products.
(Relation) IsPartOf: Europe (continent); < Ukrayina (nation); < Krym (autonomous republic)

The keyword Crimea can finally be found under the foreign names for Krym, i.e. via another record (id=1003381). This example demonstrates a fundamental problem: in order to retrieve information related to **one** specific subject, information from multiple sources, including background knowledge, must be virtually or physically **integrated.** Integration affects:

1. Metadata structure and its intended meaning, such as Creator, Reference.
2. The meaning of terminology and related background knowledge, such as Allied Leaders and Allied Powers, The Prime Minister of the United Kingdom and Churchill.
3. The use of names and identifiers for concepts and real world items in data fields, such as Yalta, Jalta and TGN7012124.

As stated in section 1, semantic interoperability means the capability of different information systems to communicate information consistent with the intended meaning. Information integration is only one possible result of a successful communication. Other forms are querying, information extraction, information transformation, in particular from legacy systems to new ones. Since the emergence of different human languages, communication could be achieved in two ways: Either everyone is forced to learn and use the same language, or translators are found that know how to interpret sufficiently the information of one participant for another. The first approach is that of proactive standardization, the second that of reactive interpretation. This choice applies to all levels and functions of semantic interoperability and is a major distinctive criterion of various methods.

4.3 Standardization versus Interpretation

Standardization in order to achieve semantic interoperability in a digital library environment may comprise: the form and meaning of metadata and content schemata; shared concepts defined in KOS; use of names and construction of identifiers for concepts and real world items.

Standardization has the following advantages:

- Information can be immediately communicated (transferred, integrated, merged etc.) without transformation.
- Information can be communicated without alteration.
- Information can be kept in a single form.
- Information of candidate sources can be enforced to be functionally complete for an envisaged integrated service.

The disadvantages are:

- Source information needs adaptation to the standard.
- The effort of producing a standard, such as a terminology, can be very high.

- The standard has to foresee all future use. Introducing a new element is time-consuming and may cause upwards-compatibility problems. Necessarily in a changing world, it will always be behind the demands of the current applications.
- A standard is one for its domain. It cannot be optimal for all applications. The necessary selection becomes a political decision.
- Adaptation of information to a standard may require interpretation (manual or automatic).
- Adaptation of information to a standard may result in information loss.

Mechanical interpretation in a digital library environment may comprise: the **mapping** of metadata and content schemata (sometimes called crosswalks); **correlation** of concepts defined in KOS (sometimes called cross-concordances); **translation** of names and **reformatting** of identifiers for concepts and real world items.

Interpretation has the following advantages:

- Source information, in particular legacy data, needs no adaptation.
- Sources can serve additional local function.
- Only application relevant parts need interpretation.
- Interpretation can be optimised for multiple functions.
- Interpreters can easily be adapted to changes

The disadvantages are:

- Interpretation needs processing time during communication.
- The manual effort of producing the knowledge base for an interpreter, such as correlation tables for terminologies, can be very high (however, there are applications of automatic generation).
- The number of interpreters needed increases drastically with the number of formats.
- Interpretation of information may result in information loss, in particular affecting recall or precision of the overall system.
- Mechanical interpretation may not be possible at all.

The conclusions are that a comprehensive approach to semantic interoperability must consider an optimal combination of both alternatives for all functions:

A standard is elegant and efficient for specific applications. It is appropriate for problems with a low degree of necessary diversity and with high long-term stability. It hinders evolution and fruitful diversity. It reduces information. In order to be applied, it may need interpreters to generate input in standard form. Additional functions may need interpreters. A typical example is the Dublin Core Element Set.

Some of the inflexibility of standards can be avoided by designing extensible or modular standards with core functionalities and community specific extensions that do not invalidate the core functions, such as Dublin Core qualifiers. The CIDOC CRM [ISO21127] is also designed as a core standard. Its extension capability is based on the well-founded specialization (IsA) of object-oriented schemata. The combination of namespace schemas into application profiles [Dekkers 2001, Heery, R., Patel, M.] falls into this category. The idea has not been applied to KOS so far, however, some namespace assignment policies can be seen in this light.

A standard is inevitable when mission critical data have to be communicated, i.e. in cases, where certain data elements are necessary and inexact equivalence of meaning is not acceptable. In that case, the component sources have to **commit** to a common set of concepts or formats, sometimes called an Interlingua. Such a role is played by e.g. the EBTI and the EET thesaurus of the European Commission, which serve communication about customs regulation and education respectively. Obviously, a European law cannot be enforced on inexact matches between product terms in different European languages.

Interpreters are effective in environments with a high degree of necessary diversity and low long-term stability. They are elegant for cases, where only smaller portions of the source data have to be communicated to a target. Whereas a standard needs to support the sum of all functions in the intended integrated environment, interpreters can be flexible and selective to the needs of smaller subgroups within the integrated environment.

If the numbers of formats in use increases, interpretation may need to go through a common switching language, which reduces the number of interpreters needed, but increases the loss of precision. Effectively, such a switching language is nothing other than another standard. The LIMBER and SCHOLNET Projects took this approach with an English thesaurus in the middle. The CIDOC CRM is designed to be a switching language for schema mappings.

4.4 Levels of Semantic Interoperability in Digital Library Environments

In the current digital library technology, one can clearly distinguish 3 levels of information that are treated in a distinct manner and give rise to distinct methods to address semantic interoperability. These are:

1. **Data structures**, be it metadata, content data, collection management data, service description data.

2. **Categorical data,** i.e. data that refer to universals, such as classification, typologies and general subjects. Theoretically, one can regard all numbers to belong to this category.

3. **Factual data,** i.e. data that refer to particulars, such as people, items, places.

4.4.1 Semantic Interoperability and Data Structures

As outlined above, data structures describe possible states of affairs and support information control and management functions. The control and management functions are normally local to a system and not an object of semantic interoperability. A conceptual model can describe the others. From an ontological point of view, the respective elements of a data structure can be related to universals of the domain, but not to particulars. Characteristically, data structures encode the most relevant relationships in the domain, which should be kept explicit and intact. They provide only very abstract individual concepts, such as resource, agent, date, etc. The information content of data structures is extraordinarily small. It is very stable over time, because it relates directly to vital functions built into the system. Consequently, they are first-class candidates for standardization. Nevertheless, as the example in section 3.2 and the hundreds of metadata standards demonstrate, flexible and interpretative approaches are also necessary as described above.

The interpretative approach is based on schema mapping. It can be divided into the mediation and the data warehouse-style approach. In the first, queries are transformed to fit a source schema, and then only the answer set is transformed into the target format. In the second case, all source data are beforehand transformed into a target format. Which approach is better, depends on the update rate of the sources, their number and the complexity of the schema mapping. Several papers of Diego Calvanese and Lenzerini deal in much detail with these issues [see e.g. Calvanese et al. 1998].

A global schema serving as switching language for multi-step mapping or for information integration must be object-oriented, because it is not possible otherwise to relate the different abstraction levels of the universals of the involved data structures. E.g. one table may be about physical books, another about electronic documents, a third about tourist guides. It is economic and effective to develop the mapping services in a wider domain, an overarching ontology consisting mainly of relationships (such as the CIDOC CRM), as the necessary information is very compact and stable.

Further, abstractions that may be **fixed in one schema** by a respective table or relation type may be **categorical data** in another. Therefore, mapping algorithms depend in general on categorical data in the source instances. Frequently, these categorical data are locally standardized in KOS and still high-level concepts. Only if the logical structure of the respective KOS and the conceptualisation behind the involved data structures are compatible, can such mappings be implemented. This problem gives rise to a demand for standardizing or harmonizing the upper levels of KOS with categorical data.

4.4.2 Categorical Data

The number of categorical data is immense. Nations and communities build their own terminologies, from thousands to millions of terms. Terminologies are in constant evolution as new classes of phenomena come up or find the scientific or public interest. From an ontological point of view, terminologies are rich in individual concepts (classes), and very poor in relationships, except for IsA relations. Some researchers assume that all terminologies could be developed into ontologies. However, it is still theoretically not clear, if all human concepts as they appear in our data actually can be reduced to a rigid logical definition [see G. Lakoff, 1987], and if that effort will pay off. Obviously, as high-level concepts are fewer and more fundamental, formal treatment should start top-down.

Standardization of terminologies starts from controlled vocabularies in local databases to international KOS. Whereas the local degree of standardization is very high, the global one is poor compared to data structures. Terminologies are also tied to communities, not just to data. Therefore, KOS about categorical data frequently cannot be standardized across communities or nations even if they deal with the same subject. Curiously, modular approaches that would standardize high-level concepts and let lower level concepts be switched in are rarely discussed.

In order to achieve semantic interoperability, concepts must be matched. Since source data about the same item may dramatically differ in the level of detail of classification (e.g. Birma cat or feline), an exact match is frequently not possible. The interpretative approach is typically based on large translation tables that declare exact or inexact equivalences [Doerr 2001] of individual concepts. The same holds to a certain degree for geographical areas, even though they are not universals. Mapping of ontologies that contain rich relationships can be extraordinarily complex [see Doerr 2003].

4.4.2 Factual Data

Factual data are the largest group in number. Part of the factual data refer to facts that may appear only once in a digital library environment, such as the relation of a specific author, place and date to a publication. However, the particular author, place and date have already a high chance to reappear, so it must be possible to identify two references in order to achieve semantic interoperability.

In contrast to categorical data, factual data can only be identical or different (except for geographical areas). There are two strategies:

1. Encoding rules, such as for dates, try to ensure

    a) That no different items are taken to be the same, such as person names frequently suggest, and

    b) That the same item is not taken to be different, such as 3-9-2004 and Sept 3, 2004.

2. Descriptions of particulars are collected in KOS, and an artificial standard identifier is assigned to each object, such as a gazetteer id for a place. The naming and referencing of standardized universals poses a similar problem.

The sheer number of particulars makes the second approach only reasonable for very important items.

Data cleaning and duplicate detection algorithms can be regarded as interpretative approaches. They may in turn make use of KOS.

In conclusion, we have proposed to classify approaches to semantic interoperability by five criteria: Standardization versus interpretation, and application to data structures, categorical data and factual data. We have argued that each of the six resulting classes deserves distinct theoretical and practical treatment, but also that these criteria are related to relevant commonalities of the different possible approaches.

5. Prerequisites to Enhancing Semantic Interoperability

In this section we describe prerequisites to enhancing semantic interoperability, i.e. elements of the necessary infrastructure and organisational issues to maintain distributed, disparate and heterogeneous information systems that can nevertheless communicate with each other or be accessed in a common way effectively.

Solutions to semantic interoperability cannot be successfully implemented based solely on an awareness of their importance (section 2) and a clarification of a theoretical framework (section 3), or just by running methods and processes on data (section 5) and establishing services (section 6).

There are a multitude of prerequisites, including technical infrastructure (architectures, protocols, syntactic solutions, encoding schemes and identifier systems), standards, organisational and legal matters, supporting services such as different registries and semantic knowledge bases and KOS (including foundational and core ontologies).

However, not all of these requirements are needed for isolated, closed-world projects and services. A lack of widely adopted and established solutions hampers open access and heterogeneous distributed services severely and prevents the generation of a maximum benefit from efforts and investments in interoperability.

We consider several areas that promote the achievement of semantic interoperability. In Section 4.1 we examine standards making and consensus building. Section 4.2 looks at the role of foundational and core ontologies, while Section 4.3 considers KOS in detail. The role of semantic services such as metadata and terminology registries is examined in Section 4.5 while Section 4.6 considers the role of architecture and infrastructure including: syntactic encoding; identifiers; protocols and the semantic description of web services. Finally, in Section 4.7 we consider matters relating to access and rights issues.

5.1 Standards and Consensus Building

Within a networked information environment, issues relating to accessibility, re-usability and interoperability are all of major importance in harnessing the potential of distributed heterogeneous resources. All of these aspects are under-pinned by the development of technical and metadata standards, which facilitate search, retrieval, evaluation, and sharing of information resources. We need appropriate technologies for the description, classification and indexing of resources using standard metadata and controlled vocabularies as well as for syntactical representations and protocols for communication.

For the purposes of this report, the term *standards* encompasses the following types of information:

- Official standards, national and international, which have gone through a formal standardisation process
- Specifications which are widely accepted
- Emerging standards which are in the process of being standardised
- De facto standards which have been widely adopted
- Guidelines for best practice

Many bodies exist for the creation of such standards, including the International Standards Organisation (ISO), the World Wide Web Consortium (W3C), the National Institute Standards Organisation (NISO), the American National Standards Institute (ANSI), the Library of Congress, the CEN/ISS, the IEEE, the Web3D Consortium, and the British Standards Institute (BSI). The process of standardisation necessarily involves consensus building with the aim of gaining widespread acceptance so that all standards-making organisations undertake an iterative process of consultation and review in their development.

The standards making machinery is a monolithic process with an increasing number of standards emerging at differing levels of information systems architecture corresponding to different levels and types of interoperability: systems and protocols (e.g. HTTP, SOAP, Z-39.50, OKBC, JDBC); syntax (e.g. XML); modelling (e.g. RDF, OWL, UML); and semantics (e.g. MARC, Dublin Core, IEEE LOM, CIDOC CRM, MPEG-7, <indecs>).

Standardisation and consensus building are processes, which precede potential interoperability and information exchange between information systems. Semantic interoperability in digital library environments is directly related to sharing and consistently using terminologies, so that rich or domain level interoperability can be achieved by negotiation and widespread acceptance with regard to shared concepts, terms and their meanings. As we will see in section 4.5 terminology services, such as registries and thesaurus servers, play an important role in facilitating consistent and coherent use of shared terms and their semantics. Section 3.3 has already characterised the advantages and disadvantages of standardisation in the area of differing types of KOS.

5.2 Role of Foundational and Core Ontologies

One of the well-accepted mechanisms for achieving semantic interoperability is the utilization of ontologies. According to a well-accepted formal definition, *"An ontology is a (possibly incomplete) axiomatization of a conceptualization"* [Guarino 1998]. According to the Wikpedia [Wikpedia], *"In computer science, an ontology is the attempt to formulate an exhaustive and rigorous conceptual schema within a given domain, a typically hierarchical data structure containing all the relevant entities and their relationships and rules (theorems, regulations) within that domain"*.

A class of ontologies of special interest are the so-called *Foundational Ontologies*, which are axiomatic ontologies which address very general domains [Masolo et.al. 2003]. Important foundational ontologies are the following:

- The *Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE)* [DOLCE], which aims to capture the ontological categories underlying natural language and human common sense.
- The *Object-Centered High-level Reference Ontology (OCHRE)* [OCHRE], which presents a revisionary view with respect to the standard notion of commonsense.
- The *Basic Formal Ontology (BFO)* [BFO], which has a meta-ontological flavour and is comprised of two components: The *Snap* Ontology of endurants, which is reproduced at each point of time and is used to characterize static views of the world, and the *Span* Ontology, an ontology of happenings and occurences and, in general, of entities which persist in time by perduring.

Putting the Foundational Ontologies aside, Ontologies can be classified into three main categories:

1. *Upper Ontologies*, where basic, domain-independent concepts as well as relationships among them are defined. According to the Wikpedia, *"an upper ontology is a hierarchy of entities and associated rules (both theorems and regulations) that attempts to describe those general entities that do not belong to a specific problem domain"*. Thus, an upper ontology is *"a core glossary in whose terms everything else must be described"*. Well known Upper Ontologies are:
   a) The *Cyc* Ontology, developed in the context of the Cyc project [Cyc]. The main objective of the Cyc project is to assemble a comprehensive ontology and database of everyday common-sense knowledge, so as to allow Artificial Intelligence (AI) applications to perform human-like reasoning.
   b) The *WordNet*, a database originally designed as a semantic network based on psycholinguistic principles [WorldNet]. It has been expanded by adding definitions and may also be viewed as a dictionary. WorldNet qualifies as an upper ontology, as it includes both general concepts and specialized concepts. The concepts may be related by the subsumption relation as well as by other semantic relations (e.g. part-of, cause etc.). The logical relations between the concepts are not precise, as the WorldNet has not yet been formally axiomatised. WorldNet has been extensively used in Natural Language Processing research.
   c) The *Suggested Upper Merged Ontology (SUMO)*, which is being developed by the IEEE P.1600 working group. SUMO aims to become the foundation ontology for several information processing systems [SUO WG]. It defines a class hierarchy as well as rules holding for the SUMO classes and relationships that may hold among them. The first release of SUMO was available in December 2000.

2. *Core (or Intermediate) Ontologies*, which are essentially the upper ontologies for broad application domains (e.g. the audiovisual domain). They may help in making real-world decisions for which Upper Ontologies may fall short, as the upper ontologies may be poor representations for certain problem domains. Core Ontologies are comprised of concepts and relationships that are thought to be basic in the broad application domain context (e.g. an event in the audiovisual domain). They often capture the semantics of well-accepted domain standards (e.g. ontologies capturing the MPEG-7 MDS in the audiovisual domain). When used in a more general context, Core Ontologies specialise concepts defined in the foundation ontologies.

3. *Domain Ontologies*, where concepts and relationships used in specific application domains are defined (e.g. a goal in the soccer video domain). The concepts defined in Domain Ontologies specialise the ones defined in both Upper and Core Ontologies, thus extending them with domain knowledge.

According to the above definitions, semantic interoperability depends mainly on the existence of well-accepted Upper and Core Ontologies, where basic concepts and relationships are defined. Then, the concepts defined in the Upper and Core Ontologies, are extended by appropriate Domain Ontologies. As the standards for metadata descriptions usually provide only general-purpose structures, the utilization of Core and/or Upper Ontologies capturing the semantics of the standards, together with Domain Ontologies that extend them with domain knowledge, are systematic mechanisms for the extension of standards.

The above approach has been successfully applied in the audiovisual domain. In Troncy [2003] an Upper Ontology capturing the concepts of genre, theme and technical process has been defined and a set of Domain Ontologies are extending its semantics. The ontologies guide the annotation of the audiovisual content and the metadata produced are compliant with the well-accepted *MPEG-7* [ISO/IEC JTC 1/SC 29/WG 11/N3966 2001] standard for audiovisual content metadata descriptions. In the context of the DS-MIRF framework [Tsinaraki et al. 2003, 2004a 2004b and 2004c], an OWL Upper Ontology fully capturing the MPEG-7 MDS semantics has been defined. The Upper Ontology defined in the context of the DS-MIRF framework is a successor of the Core Ontology defined in Hunter [2001], having two main advantages compared with it:

a) It is expressed in the OWL (Web Ontology Language) [McGuiness & Harmelan 2004], which is the dominant standard for ontology definition; and
b) It fully captures the MPEG-7 MDS (regarding classes, attributes, relationships and constraints) whereas the ontology defined in Hunter [2001] captures it only partially.

The DS-MIRF framework utilizes domain knowledge captured in OWL Domain Ontologies, defined systematically in order to extend the Upper Ontology. During audiovisual content segmentation, ontology-based semantic indexing takes place, resulting in the production of structured semantic metadata that describe the audiovisual content. The metadata produced are in OWL/RDF format and they are transformed, using appropriate transformation rules, into both MPEG-7 and *TV-Anytime* [TV-Anytime forum] compliant metadata, thus providing interoperability with software compliant with these standards.

The Cultural Heritage domain is seeking semantic interoperability, as there exist several metadata formats, followed by different institutions. Most of them have now been subsumed by the *CIDOC/CRM (Conceptual Reference Model)* [ISO/IEC, ISO/DIS 21127], which provides definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation. The CIDOC/CRM is regarded as a reference ontology for the interchange of cultural heritage information.

In order to allow the description, exchange and sharing of multimedia resources in the cultural heritage domain, a Core Ontology has been defined [Hunter 2002]. The ontology is based on the CIDOC/CRM and the MPEG-7; essentially MPEG-7 is combined with the CIDOC/CRM in [Hunter 2002].

An effort to utilise Upper and Core Ontologies in the Cultural Heritage domain is carried out in the context of the harmonisation of the *IFLA/FRBR (International Federation of Library Associations and Institutions, Functional Requirements for Bibliographic Records)* [FRBR 1998] standard for bibliographic records with the *CIDOC CRM*. According to the minutes of the 2nd Meeting on FRBR/CRM Harmonization, *"The main objective of the FRBR/CRM Harmonisation Group is to express the FRBR model as an object-oriented conceptual model, which can be regarded as a kind of formal ontology. The objective is not to "transform" the IFLA/FRBR model into something totally different or "better", nor of course to "reject" it or "replace" it – but to express the conceptualization of FRBR with the object-oriented methodology instead of the ER methodology as an alternative. It also is an opportunity to develop an actual ontology out of the IFLA entity-relationship model, with a formalism more suitable for Semantic Web related activities"*.

5.3 Knowledge Organization Systems

In order to understand the crucial role of Knowledge Organization Systems (KOS) as one major prerequisite for the improvement of semantic interoperability we need to go beyond the general formulations in the earlier chapters. We have to investigate the broad and rich variety of different types of KOS, the large number of KOS available, their size, access rights and examples of KOS really used for the purpose.

5.3.1 KOS are prerequisites to enhancing Semantic Interoperability

Establishing and improving semantic interoperability always requires vocabularies or better, Knowledge Organization Systems (KOS) to be used [Tudhope 2004]. Sometimes they need to be created (or extracted) first, in other cases existing vocabularies need to be transformed, mapped, merged or similar in order to make the vocabularies and/or documents or databases semantically interoperable. This is especially important if KOS are different with regard to their structure, domain, language or granularity (cf. the empirical analysis of projects in Zeng and Chan's article [Zeng 2004]).

All evidence shows that we cannot expect one KOS or general ontology to be applicable or even suitable to many different audiences and services. As the HILT Final Report states [HILT 2001]:

"3.1.1 Integrated Independence: A Basis for Interoperability?

There is evidence of growing agreement that interoperability in respect of subject schemes in a distributed environment is recognised as an issue [Miller, 2000] and that a standards based approach is the answer [Stam, 1990], but no evidence to suggest that one particular scheme or single approach [see, for example, Ledsham, 1999 and Garrod, 2000] will provide the whole answer - a conclusion borne out by the outcomes from the HILT Focus Group, Stakeholder Workshop, and Interim Report consultation exercise. These all favoured an integrative mechanism that would improve interoperability whilst allowing an independence of approach to be maintained as regards subject description in the various domains. That having been said, there are examples in the literature of Museums and Archives using library science approaches to subject description with apparent success. Inevitably, work in this area tends to focus on criticisms of heavily used schemes such as LCSH. The outcome is not always negative, but the general impression obtained from the literature is that no one scheme is - or, perhaps, can be - ideal for every purpose. This is borne out by other HILT results (see, for example, Appendix C, 'Perceived Strengths and Weaknesses of Controlled Vocabularies')."

HILT originated from discussions and ambitions in the UK Museum and Library community to develop a unifying High Level Thesaurus.

Experience indicates as well that we need to abstain from unnecessarily and prematurely narrowing down the "useful" types of KOS to e.g. thesauri and then to largely ignore to use and create solutions for the other types of KOS. Protocols and encoding schemes have been primarily developed for thesauri recently [cf. ZThes; SKOS]. As indicated below, historically a broad and rich variety of different types of KOS have been developed and used with good results.

Even though we do not further discuss this issue here, it should be mentioned that especially in an Internet context, a lot of people have been arguing that subject indexing and structuring using KOS or similar would not be necessary anymore because of the power of full text searching and the size, speed and timeliness of global search engines. After some years these arguments are much quieter. Instead, quite popular services that introduce (simple and low-level) indexing and structuring have appeared, such as the DMOZ Open Directory Project [DMOZ] with its collective information structuring, the social "tagging" of pictures in Flickr [Flickr] or the "folksonomy" (conflation of folk and taxonomy), bottom-up organisational categories emerging by pooling users tags, initiated in the blog aggregator/search site Technorati [Technorati]. It remains to be seen to what degree these initiatives are successful and if they manage to contribute to real semantic interoperability.

5.3.2 Taxonomy of KOS

Many different terms have been used and continue to be used to describe sets of terms in the different communities and disciplines making up the digital library field concerned with information systems and services. They are rarely defined, often used synonymously or heavily overlapping. Examples are: KOS, NKOS, Vocabularies, Ontologies, Schemas and Taxonomies. For definitions used here see Section 3.2.

The terminology depicting specific types of KOS is almost as fuzzy as the general terms. The NKOS initiative started developing a draft taxonomy of KOS [NKOS 2000] based on Gail Hodge's work [Hodge 2000]. It enumerates and defines the most frequently used types of KOS, their characteristics and purposes. In addition, it groups the types of systems into

- Term lists (words, phrases, sometimes including definitions);
- Classification and categorization systems (for the creation of subject sets)

- Relationship schemes (emphasizing relationships between terms and concepts). Semantic networks and ontologies, types that are not always placed in a KOS context, are here seen as relationship schemes alongside thesauri.
- Term Lists
    - Authority Files (e.g. LC Name Authority File)
    - Glossaries
    - Gazetteers
    - Dictionaries
- Classification and Categorization
    - Subject Headings (e.g. Medical Subject Headings MeSH)
    - Classification Schemes, Taxonomies and Categorization Schemes (e.g. Dewey Decimal Classification DDC, scientific taxonomies)
- Relationship Schemes (better: relational schemes)
    - Thesauri
    - Semantic Networks (e.g. WordNet, part of UMLS)
    - Ontologies

Even though the grouping seems not widely adopted (yet) and should be further discussed (e.g. the placement of Subject headings under Classification or the inherent view of classification systems as not being relationship schemes), the types and groups are used in this report since they illustrate the broad variety of KOS.

Rather than representing more or less good and useful vocabularies in general, as often claimed in discussions, the KOS types can be seen to span a continuum of systems between low levels of term control and lacking relationships between terms (and terms and concepts) at one end (e.g. simple term lists with some synonym control only) and systems with higher level conceptualisation, formal definition of terms and relationships and inference rules and definition of roles to support reasoning applications (e.g. advanced ontologies) at the other end.

This view indicates that different types of KOS can and very often are further developed/"upgraded" to a "higher level" of control or vice versa, simplified to be used in a context where the higher cost of the more controlled system cannot or need not to be carried. Such an activity is called KOS transformation.

One example of the prior case is the work started by FAO to develop the AGROVOC thesaurus into a full-fledged ontology [Soergel et al 2004]. The comparably "simple" semantic relationships in such a standard thesaurus (rather unspecified hierarchical, equivalence and associative relationships), the existing lead-in vocabulary and definitions and literary warrant in the scope notes, are the important and necessary basis of the "upgrading" work. Ontology development should not normally start from scratch as long as there is some kind of KOS available for the domain.

5.3.3 Number and size of KOS and NKOS

Many of the big online database hosts developed or used controlled vocabularies for their databases, about a few thousand probably. About 3000 controlled vocabularies are listed in records in a library OPAC in Toronto. The Library of Congress maintains a rather large list of so called Subject/Index Term Sources "MARC Code List for Relators, Sources, Description Conventions (of MARC 21)" for

cataloguing purposes [MARC Code List]. A few thousand publicly known and maintained vocabularies is what used to be listed in older printed registries.

A few web sites provide overlapping and limited lists of vocabularies available for free on the WWW [Koch; SWAD-E Thes; HILT lists], taken together between 100 and 200 vocabularies. Taxonomy Warehouse, a private company, has started building a vocabulary registry [Taxonomy] but it is still far from what would be needed to support distributed usage of vocabularies on the net (cf. the NKOS Registry work [Vizine-Goetz 2001]). The issue of KOS registries is covered in Section 4.5

Lists of ontologies are provided at SchemaWeb [SchemaWeb] and the DAML Ontology Library [DAML], which lists and categorises 282 DAML ontologies.

KOS come in all sizes and especially general-purpose systems developed over a long time for large audiences can grow very large. They stretch from a few terms, like in many ad-hoc ontologies via large thesauri like the CABI thesaurus containing 50 000 terms to huge systems like AAT with 125 000 terms. A term record can, in addition, be quite rich, contain terms in several different languages, many related terms with different relationships, scope notes etc.

5.3.4 Methods and processes applied

Since KOS are always involved when dealing with semantic interoperability, methods and processes applied to them are presented in other parts of this report. The syntactic issues, representation and encoding systems are covered in Section 4.6.1; identification and naming in Section 4.6.2; protocols for accessing and serving terminology in Section 4.6.3; registries of KOS (metadata about the vocabularies), and maybe of their terms and concepts, are described in Section 4.5; KOS transformation, correlation, mapping and other methods to enhance semantic interoperability are treated in Section 5 and the usage of KOS in digital library services in Section 6.

Efforts in all these areas are argued for in Linda Hill's and colleagues research agenda [Hill 2002] on the integration of KOS into digital libraries.

5.3.5 Availability, Rights

The vision of a free and interoperable distributed usage of vocabularies is easy to formulate but in real life there are many obstacles. Vocabularies are hard to discover and evaluate (suitable registries and metadata about the systems is missing), still most vocabularies are not digitally available at all or digital public versions are far from complete, there is rarely syntactic interoperability (owners have not adapted to standardised representation, identification, encoding etc.), owners do not yet prepare for terminology services nor provide term/concept level metadata, service protocols are not ready, nor commonly agreed upon etc.

Probably most important, by far the majority of large and well-maintained vocabularies are not freely available. IPR issues are not clarified; license models are not developed or tested for such vocabularies. To allow local institutions or even individuals to tailor and augment vocabularies is an option far from being realised.

5.3.6 Examples of the usage of KOS in Semantic Interoperability Applications

To our knowledge there is no overview and statistics available regarding the usage of KOS in (public) digital information systems. A couple of hundred abstract & index databases hosted by commercial vendors have been using KOS for a rather long time, predominantly scientific, technical and medical vocabularies.

On a national level the HILT project investigated the use of KOS in the UK. Most libraries, museums and archives seemed to use in house vocabularies, only one third of the responding JISC collections and services mentioned published systems with DDC, LCSH, the UNESCO thesaurus and the HASSETT, Humanities and Social Sciences Electronic Thesaurus, as the most frequently occurring [Nicholson et al, HILT II Final Report, App. B]. Examples of other frequently mentioned KOS in digital service contexts are the AAT, AGROVOC, CABI, GEMET thesaurus, the group of mda thesauri and the MeSH Subject Headings.

In a very useful project overview, Zeng and Chan mention more than 40 KOS (not a complete list) as being involved in projects addressing semantic interoperability issues [Zeng 2004]. Among those are:

Subject Headings (Term lists):

CHS Canadian Subject Headings
FAST Faceted Application of Subject Terminology
GSAFD Genre terms for fiction
GFSH General Finnish Subject Headings
LCSH Library of Congress Subject Headings
LCSHac LC Children's Headings
MeSH Medical Subject Headings
Rameau Répertoire d'autorité-matière encyclopédique et alphabétique unifié
RVM Répertoire de vedettes-matière (Canada)
Sears List of Subject Headings
SHL Subject-Heading Language (National Library in Warsaw)
SWD Schlagwortnormdatei
ULAN Union List of Artist Names

Classification systems:

BIOSIS Concept Codes
BSO Broad System of Ordering
DDC Dewey Decimal Classification
EI Classification
ICONCLASS
LCC Library of Congress Classification
MSC Mathematics Subject Classification
NLMC National Library of Medicine Classification
PACS Physics and Astronomy Classification Scheme
PTC Polish Thematic Classification
RVK Regensburger Verbundklassifikation
SAB Klassifikationssystem för svenska bibliotek
SIC Standard Industrial Classification
UDC Universal Decimal Classification
U.S. Patent and Trademark Office Patent Classification

Relationship schemes:

AAT Art and Architecture Thesaurus
Allärs Allmän tesaurus på svenska (Finland)
GLIN Thesaurus for the Global Legal Information Network
English Heritage Thesaurus
ERIC Thesaurus
HEREIN European Heritage Information Network, Thesaurus
INSPEC Thesaurus
IZT Informationszentrum Sozialwissenschaften Thesaurus (Germany)
LIV Legislative Indexing Vocabulary

National Monuments Thesauri (English Heritage)
TCT Thesaurus of Common Topics (Poland)
TGM Thesaurus of Graphic Materials (Library of Congress)
TGN Getty Thesaurus of Geographic Names
Thésaurus de l'architecture (Merimee)
UMLS Unified Medical Language System
UNESCO Thesaurus
WordNet
YSA Yleinen suomalaninen asisanasto (Finland)

Several of the projects mentioned, themselves comprise several different vocabularies (UMLS: 100; Renardus: 10; H. W. Wilson megathesaurus: 12), many KOS are multilingual.
Quite a few of the projects work with vocabularies from two or all three groups of KOS.

5.5 Role of Semantic Services

With the advent of machine-processable data (see section 4.6.1) comes the prospect of interoperability, which is increasingly regarded as being important in realising the goal of accessing and reusing data. However, for semantic interoperability to take place requires sharing and consistent use of terminologies, which can only result from a community basing its practices on well informed, published, authoritative information.

As data migrates from its original source to an integrated system, it is necessary to ensure that it continues to be interpreted correctly otherwise disastrous consequences will ensue in terms of query processing. Several types of terminology services have emerged in the digital library world, with the aim of supporting semantic reconciliation and thereby enhancing semantic interoperability, they include: registries or repositories of metadata and semantics; metadata schema registries; registries of crosswalks or mappings between vocabularies; and ontology servers as well as other types of terminology services.

Although the functionalities provided and the target audiences vary, they are all characterised by the following features:

- they provide authoritative, trustworthy content
- they have a collections and a persistence policy
- they are concerned with making information readily accessible

We distinguish between registries and repositories in that a repository is merely concerned with the collection of some corpus of data, whereas a registry has an additional layer, which caters for policy and management issues, as well as providing user level services.

Terminology services play an important role by supporting the following types of functionality:

- disclosing concepts, terms and semantic relationships
- promoting consistent use of vocabularies
- publication of semantics
- providing examples of use and best practice
- making accessible information relating to provenance, currency, authoritativeness, deduction, and reasoning processes [McGuinness and Pinheiro da Silva, 2003]

The types of entities that are used to determine semantic proximity and that support semantic reconciliation include: vocabularies, classifications or taxonomies and thesauri.

Semantic interoperability requires domain-level consensus on the structure, concepts and terminology to be used in knowledge representation. Semantic registries serve an informational purpose by collecting together appropriate information and tracking developments in a relevant area. As

mentioned earlier, several such services are emerging; they are aimed at collaborative development of metadata vocabularies and their harmonisation at a domain level.

5.5.1 Metadata Registries

Metadata registries come in various guises, depending on their intended target audience and the functionalities that they are required to support. For example, the Environmental Data Registry developed by the US Environmental Protection Agency [EPA] and the National Health Information Knowledgebase hosted by the Australian Institute of Health and Welfare [NHIK] are both concerned with collecting and maintaining information relevant to their respective domains in a repository to support their target communities. Both of these initiatives use the ISO/IEC 11179 standard for data registries [ISO/IEC 11179] as their foundation.

5.5.2 Metadata Schema Registries

A metadata schema registry is often developed in order to manage the evolution of a single vocabulary. An example is the one maintained by the Dublin Core Metadata Initiative [DCMI]. The primary purpose of this registry is to support the evolution of the Dublin Core Metadata Element Set [DCMES] including its manifestation in multiple languages.

Other metadata schema registries hold as their content, multiple vocabularies and the relationships between them. Examples include the DESIRE [DESIRE registry] and SCHEMAS [SCHEMAS Registry] Registries, their primary functions being to provide a publication environment for the disclosure of customised metadata vocabularies or application profiles [Heery 2000, Baker 2001]; the Metadata for Education (MEG) Registry [MEG Registry] for supporting the UK education community and the CORES Registry [CORES Registry].

Such registries play an important role in making apparent trends in the usage of various vocabularies and even individual terms and hence promote the consensus building process, which is a foundation for semantic interoperability.

5.5.3 Registries of Mappings

Section 3.1.4 introduced the concepts of crosswalks, schema mapping and schema matching. Creating mappings from one metadata framework to another is one of the most important and widely used ways in which differences are reconciled to enable automatic processing and integrated access to heterogeneous information systems –in particular for legacy data.

Resource discovery and information retrieval across a wide variety of sources is a major driver. Consequently, consistency of use becomes very important. Metadata registries, which develop and maintain authoritative information with regard to formal mappings and relationships between multiple metadata schemas, are essential in this process. Registries maintained by standards bodies have not yet emerged, however several examples of informal registries and collections of mappings which are accessible over the Web do already exist:

Examples of Mappings

Crosswalks from the Alexandria Metadata Schema to Other Schemas
http://www.alexandria.ucsb.edu/public-documents/metadata/crosswalks.html

DLESE http://www.dlese.org/Metadata/crosswalks/index.htm

Getty Information Institute. Metadata Standards Crosswalk
http://www.getty.edu/research/institute/standards/intrometadata/3_crosswalks/index.html

IEEE http://ltsc.ieee.org/doc/wg12/LOM_1484_12_1_v1_Final_Draft.pdf

LC Network Development and MARC Standards Office, http://www.loc.gov/marc/

Metaform: Crosswalks, Crosscuts, & Mappings, State and University Library at Göttingen, Germany (SUB)
http://www2.sub.uni-goettingen.de/metaform/crosswalks.html#Crosswalks

OCLC crosswalks http://www.oclc.org/research/projects/mswitch/1_crosswalks.htm

Examples of Registries and Directories of Mappings

MAAT Metadata Standards Crosswalks http://www.sinica.edu.tw/~metadata/tool/mapping-foreign.html

Mapping between metadata formats, http://www.ukoln.ac.uk/metadata/interoperability/

MIT DSpace Metadata Mapping http://libraries.mit.edu/guides/subjects/metadata/mappings.html

5.5.4 Other Terminology Services

So far we have considered registries, which have been primarily developed for use by humans. However, for automated processing it is necessary that similar types of information be made available to software and software agents through machine interfaces. Section 4.6.3 considers various protocols that have been developed for terminology services, largely to provide programmatic interfaces to thesauri (CERES, Zthes, ADL, and SKOS).

In addition, there has been a recent proliferation of work in the area of ontology servers. An ontology server not only enables the publication and disclosure of the semantics that are being used in applications and web services, but it can also be queried by software agents roaming the Web in order to retrieve semantics [Hendler 2001, Gibbons 2003]. This in turn facilitates the type of reasoning and inference required to make the vision of the Semantic Web [Berners-Lee et al. 2001] a reality.

A major function of an ontology server is to enable inferencing as well as automated query processing, so that tasks may be performed with little or no human intervention. There is a considerable amount of work being done in this area with several groups of researchers investigating the issues involved [Patel 2004, FIPA 2000, Suguri 2001, Volz 2003, Pan 2003, and Farquhar 1996].

5.6 Role of Architecture and Infrastructure

Automated processing of information requires several architectural components to form an infrastructure to support it. Below we examine the role of syntactic encoding in achieving semantic interoperability; digital resource identification; development of various protocols for accessing terminology services and the semantic description of web services.

The following architectural and infrastructural requirements are more important for open access solutions and broad heterogeneous distributed services than for isolated projects. In the latter, architecture and identifiers can in particular be “home-made” and ad-hoc and even encoding systems and protocols can be proprietary and only loosely based on standards like XML or the ADL thesaurus Protocol. Good and comprehensive solutions require these prerequisites; they enable vastly improved interoperability and form a significant step towards the vision of a (real semantic) Semantic Web.

5.6.1 Syntactic Interoperability and Encoding Systems

Digital processing of information requires it to be encoded in a machine-processable form using encoding syntax or mark-up such as that provided by MARC, XML DTDs, XML, RDF, OWL, VocML

and MathML. Syntactic heterogeneity is concerned with differences in the representation and encoding of data. While syntactic interoperability is not necessarily a prerequisite for all types or levels of semantic interoperability (see Section 3.3), it is required for the automated integration of information from multiple sources.

Matthews et al. discuss the modelling of thesauri for the Semantic Web [Matthews et al.]. They describe and analyse several approaches that have been proposed using Semantic Web ontology languages such as DAML+OIL, RDFS and OWL:

- a term-based approach that models terms as a class of resources
- a subclass approach in which terms are modelled as classes themselves
- a term with categories approach
- a concept-based approach

Vizine-Goetz et al. on the other hand have opted to use MARC as the format of choice in their project, which is concerned with mappings between controlled vocabularies such as LCSH [Vizine-Goetz et al. 2004].

Syntactic interoperability is achieved when compatible forms of encoding and access protocols are used to allow the information systems concerned to communicate with each other. However, this does not mean that each system can process the data in a manner consistent with the intended meaning. For example, one system may use an entity called "Actor" and another one called "Agent". With syntactic interoperability, data from both systems may only be retrieved as distinct, even though they may have exactly the same meaning. The role of protocols in enabling programmatic access to KOS is considered in Section 4.6.3.

5.6.2 Digital Resource Identification

In the traditional Library world, identifiers such as: the International Standard Book Number (ISBN); the International Standard Serial Number (ISSN); the Serial Item and Contribution Identifier (SICI) and the Book Item Contribution Identifier (BICI) have been used in order to identify and access resources or their specific parts.

Given a networked information environment several initiatives have emerged proposing various schemes for identifiers such as the Digital Object Identifier (DOI) [DOI], the Uniform Resource Name (URN) [URN], Persistent Uniform Resource Locators (PURLS) [PURL] and the Archival Resource Key (ARK) [ARK].

One consequence of operating in a global digital information environment is that the unique identification of resources becomes a major issue. Concepts, metadata terms, controlled vocabularies and the relationships between these various types of entities need to be identified so that they can be automatically referenced and processed. Unambiguous identification is of particular interest for the aggregation of information relevant to a specific entity across multiple resources.

On the Web, the Uniform Resource Identifier (URI) [URI] provides for the unique identification of resources, it can be used to identify resources such as images, places, music, documents and people. More importantly in the context of digital library systems, it is used to uniquely identify individual concepts, terms and relationships that constitute a KOS, so that it is possible to distinguish between entities with the same label and thereby the semantics associated with them. However, it should be noted that technically unique identifiers are insufficient for semantic interoperability, which involves the matching of concepts and the negotiation of unique references for particular concepts.

Additional issues, which complicate the achievement of semantic interoperability, include:

- Versioning or sequencing of concepts and terms as their semantics evolve over time

- Identification of derivatives or alternate formats for resources which have the same intellectual content

5.6.3 Protocols

A prerequisite for semantic interoperability is the use of standardised protocols for query and access to terminology. The adoption of common standards has the benefit of enabling a logical division of effort. KOS resources, search interfaces, cataloguing/indexing/mapping tools and indexed collections using common thesaurus protocols may all be developed by separate institutions, and may be physically hosted in separate locations.

Linda Hill and colleagues have argued for a general KOS service protocol from which protocols for specific types of KOS can be derived [Hill et al 2002]. The idea is to provide programmatic access to KOS content by various types of (Web) services, as opposed to thinking only of interactive human interfaces. Thus, in future a combination of thesaurus and query protocols might permit a thesaurus to be used with a choice of search tools on various kinds of database. This includes not only controlled vocabulary search applications but also collections without controlled metadata. For example, semantic query expansion services could be used with both free text and controlled vocabulary indexed collections.

A variety of interchange format specifications for the representation and dissemination of thesaurus data have been developed and are in use today. These are tagged text formats such as the MARC21 Authority format as used by the J. Paul Getty Trust, XML based formats such as the ZThes DTD [Zthes], and RDF representations such as SWAD-Europe's SKOS-Core schema [SKOS]. In order to facilitate distributed thesaurus access, a platform neutral access protocol should be used to manipulate thesaurus data. Protocols for retrieving thesaurus data are closely linked to thesaurus representation formats. The CERES [CERES], Zthes and ADL [ADL] protocols are reviewed by Binding and Tudhope [Binding and Tudhope 2004].

The Californian CERES/NBII Thesaurus Partnership Project (CERES) developed a general protocol standard for distributed thesaurus communication. This project was a collaboration between the California Environmental Resources Evaluation System (CERES), and the US Geological Survey Biological Resources Division (USGS/BRD) to facilitate access to environmental information. The aim was to construct an integrated controlled environmental vocabulary together with the tools that would enable it to be used for metadata creation and query construction, in both stand-alone and Web systems. This involved the development of a 'general-purpose thesaurus applications programming interface' to broker communication between the thesaurus and client applications. A working demonstration was provided on the project Web site. CERES developed an HTTP protocol using an RDF (XML) thesaurus representation format that followed the NISO Z39.19 standard.

The Zthes Z39.50 profile for thesaurus navigation (ZTHES), 'an abstract model for representing and searching thesauri', was based on the Z39.50 protocol following ISO 2788 [ISO 2788]. Thus part of the specification concerns the representation of thesaurus database records for Z39.50 implementation. It was intended, however, that the model could be general enough for use in other base communication protocols and an XML thesaurus DTD is given for the model. The Zthes profile has been used to make some thesauri available (via Z39.50) on the Internet by means of a Zthes-compliant Z39.50 server. Subsequently Zthes has been used as part of the ZING, 'Z39.50-International: Next Generation' effort, in the SRW Search/Retrieve Web Service protocol [SRW]. While looking to build on and facilitate access to Z39.50 systems, SRW includes both SOAP and URL-based access mechanisms. Implementations of the protocol exist, however access is restricted to individual data elements in the thesaurus model.

The ADL (Alexandria Digital Library) thesaurus protocol [ADL] is intended as a lightweight, stateless programmatic interface to thesaurus servers, based on XML and HTTP. The protocol's model of a thesaurus closely follows Z39.19 and the definition is specified in an XML DTD and corresponding

XML schema. Unlike the wider Z39.50 context of Zthes, the ADL protocol is focused on 'downloading, querying, and navigating thesauri'. A sister gazetteer protocol has also been developed. A generic, open source Java thesaurus server is supplied and demonstration forms illustrate the five independent services.

Although it is possible to produce a browsing hierarchy via combinations of primitive calls, we argue that the overheads resulting from the round-trip network latency of repeated calls to the service provider would hamper the performance of interactive interfaces over common Web bandwidth restrictions. Therefore an appropriate composite provision in the protocol is desirable. The Simple Knowledge Organisation System (SKOS) API is a more recent development, which defines a core set of methods for programmatically accessing and querying a thesaurus based on the SWAD-Europe project's SKOS-Core schema [SKOS]. It has been implemented as a web service API and a demonstration is available, including server, client and sample data. This approach builds on the ADL protocol's provision of composite groupings of primitive requests for individual data elements, such as get-broader and get-narrower (with parameters for the number of levels).

The current trend towards service-oriented architectures (SOA) brings an opportunity of moving towards a clearer separation of interface components from the underlying data sources, via the use of appropriate Web services. There are many advantages to this approach: platform neutral dissemination of thesaurus content, leveraging existing intellectual effort in the compilation of thesauri - exploiting common representations, etc. However, in an SOA, basing distributed protocol services on the atomic elements of thesaurus data structures and standard relationships is not necessarily the best approach; client operations that require multiple client-server calls would carry an overhead, as each function call introduces an element of round-trip network latency. This would limit the interfaces that could be offered by applications adhering to the protocol. We argue that Web interfaces offering advanced thesaurus services require protocols which group primitive thesaurus data elements (via their relationships) into composites, to achieve reasonable response rates.

5.6.3 Semantic Description of Web Services

A *Web Service* is a software program that can be accessed via the Internet through its exposed interface (e.g. a query service built on top of the information system of a cultural heritage institution). Web services are identified by their URIs. Web service interface descriptions declare

a) The operations that can be performed by a web service;
b) The message types exchanged during the interaction with the web service; and
c) The physical location of ports, through which information should be exchanged.

Bindings define the computers and ports where messages should be sent. Web services are usually deployed in web servers and can be invoked by any software component (including web services), independently of its implementation [Cabral et. al. 2004].

Web services initially aimed to revolutionize eCommerce and enterprise-wide integration. These expectations were not met; current standard technologies for web services (e.g. WSDL [Christensen et. al. 2001]) provide only syntactic-level functionality descriptions. Web services usually offer little more than a formally defined invocation interface, in the form of human-oriented metadata that describe the service function and the organization that developed it (e.g. through UDDI descriptions [UDDI Consortium 2000]). Although applications may invoke web services using a common, extensible communication framework (e.g. SOAP [W3C 2003]), the lack of machine-understandable semantics makes human intervention necessary for automated service discovery and composition within open systems [Cabral et. al. 2004].

Semantic Web Services (SWS) have been introduced in order to

a) Augment web services with rich formal descriptions of their capabilities; and
b) Facilitate the automated discovery, composition, dynamic binding, and invocation of services within an open environment.

A Semantic Web Service is a semantically described service. Sophisticated description models are utilized in SWSs, which can be enhanced with ontologies enabling both machine interpretability of the SWS capabilities and integration with domain knowledge. However, *Semantic Service Description frameworks* are needed, which should provide the infrastructure for supporting semantic interoperability among web services.

Current efforts in SWS infrastructure development can be characterized along three orthogonal dimensions [Cabral et. al. 2004]:

- *Usage Activities*, which define the functional requirements that should be supported by a framework for SWSs.
- *Semantic Web Service Architecture*, which describes the components needed for accomplishing the activities defined for SWSs.
- *Service Ontology*, which aggregates all the concept models related to the description of SWSs, and constitutes the knowledge-level model of the information describing and supporting their usage.

The semantic ontology dimension is fundamental in defining SWSs, as it represents both the service capabilities and the restrictions applied to the use of a specific service. The service ontology essentially integrates at the knowledge-level the information that has been defined by web service standards (e.g. UDDI, WSDL, etc.) with related domain knowledge. Three main semantic web service frameworks have been developed [Cabral et. al. 2004]:

1. The *Internet Reasoning Service - IRS-II* [Motta et. al. 2003], a SWS framework that allows applications to semantically describe and execute web services. IRS-II is based on the *UPML (Unified Problem Solving Method Development Language)* framework [Omalyenko et. al. 2003]. UPML distinguishes between four categories of components:
   - *Domain models*, which describe the domain of an application (e.g. cultural heritage, sports etc.).
   - *Task models*, which provide a generic description of the task to be solved (e.g. search for resources related to specific concepts), specifying the input (e.g. concepts) and output types (e.g. resources), the goal to be achieved (e.g. location of resources relevant to the concepts specified) and applicable preconditions (e.g. existence of available information sources).
   - *Problem Solving Methods (PSMs)*, which provide abstract, implementation-independent descriptions of reasoning processes that can be applied to solve tasks in a specific domain.
   - *Bridges*, which specify mappings between the different model components within an application.

   Web service invocation is capability-driven in IRS-II. This is supported by a task-centric invocation mechanism provided by the framework.

2. The *OWL-S* (previously DAML-S) [OWL-S Coalition 2003] framework, which consists of a set of OWL ontologies designed for describing and reasoning over service descriptions. OWL-S allows describing services that can be expressed semantically, and yet grounded within a well-defined data typing formalism. This is achieved utilizing the expressivity of description logics and the practical feasibility found in the emerging web service standards. OWL-S consists of three main upper ontologies:
   - The *Profile*, which is used to describe services so as to support service discovery (e.g. thesauri, search services etc.). The profile class can be sub-classed and specialized, thus supporting the creation of profile taxonomies capable of describing different classes of services.
   - The *Process Model*, which describes the composition (or orchestration) of the flow of control and the execution sequence of one or more services. It is used both for reasoning about possible compositions and controlling the enactment/invocation of a service. Three process classes have been defined:
     - The *atomic process*, which is a single, black box process description with exposed inputs, outputs, preconditions and effects (IOPEs). Inputs and outputs relate to data

channels, where data flows between processes. Preconditions specify facts of the world that must be asserted in order for an agent to execute a service. Effects characterize facts that become asserted given a successful execution of the service, such as the physical side effects the service execution has on the physical world.

- ♦ *Simple processes*, which provide a means of describing service or process abstractions. They have no specific binding to a physical service, and thus have to be realized by an atomic process or expanded into a composite process.
- ♦ *Composite processes*, which are hierarchically, defined workflows consisting of atomic, simple and other composite processes. These process workflows are constructed using a number of different composition constructs (e.g. Sequence, Unordered, Choice, If-then-else, Iterate, Repeat-until, Repeat-while, Split, and Split+join etc.).

The profile and process models provide semantic frameworks whereby services can be discovered and invoked, based upon conceptual descriptions defined within OWL ontologies

- The *grounding*, which provides a pragmatic binding between a concept space and the physical data/machine/port space, thus facilitating service execution. The process model is mapped to a WSDL description of the service, through a thin grounding. Each atomic process is mapped to a WSDL operation, and the OWL-S properties used to represent inputs and outputs are grounded in terms of XML data types. Additional properties pertaining to the binding of the service are also provided.

3. The *Web Service Modeling Framework (WSMF)* [Fensel & Buntler 2002], which provides a model for describing the various aspects related to web services. Its main goal is to fully enable e-commerce by applying Semantic Web technology to web services. WSMF is centred on two complementary principles: (a) a strong de-coupling of the various components that realize an e-commerce application; and (b) a strong mediation service enabling web services to communicate in a scalable manner. Mediation is applied at several levels (i.e. mediation of data structures, mediation of business logics, mediation of message exchange protocols and mediation of dynamic service invocation). WSMF consists of four main elements:
   - *Ontologies* that provide the terminology used by other elements.
   - *Goal repositories*, where the problems that should be solved by web services are defined.
   - *Web Service descriptions* that define various aspects of web services
   - *Mediators*, which bypass interoperability problems.

   The implementation of WSMF has been assigned to two main projects:
   - The *Semantic Web enabled Web Services (SWWS)* (SWWS Consortium 2003), which will provide a description framework, a discovery framework and a mediation platform for web services, according to a conceptual architecture.
   - The *WSMO (Web Service Modeling Ontology)* [WSMO WG], which will refine WSMF and develop a formal service ontology and language for SWS. WSMO service ontology includes definitions for goals, mediators and web services. A web service consists of a capability and an interface. The underlying representation language for WSMO is F-logic. The main feature of the WSMO architecture is that the goal, web service and ontology components are linked by four types of mediators:
     - ➢ *OO mediators*, which link ontologies to ontologies,
     - ➢ *WW mediators*, which link web services to web services,
     - ➢ *WG mediators*, which link web services to goals, and finally,
     - ➢ *GG mediators*, which link goals to goals.

5.7 Access and Rights Issues

Access to knowledge and organisation schemes is very often inhibited due to these being deeply embedded in proprietary information systems or due to IPR reasons. For semantic interoperability to

take place it is necessary that there is ready access to and availability for sharing and reuse of these types of information. Digital Rights Management (DRM) will play an important role in bringing about (or not) a situation of wide accessibility.

The proliferation of digital information has resulted in new issues being raised with regard to rights issues, both for copyright holders and users of digital resources. Libraries have traditionally been involved with IPR issues largely as intermediaries that manage rights between holders of copyright over resources and users of those resources.

The large-scale infringement of copyright made possible by peer-to-peer systems such as Napster [Napster], has led to the emergence of DRM systems. DRM refers to the general concept of expression of terms of access and use, as well as the enforcement of those terms through technology. These technologies are therefore aimed at increasing the scope of control that rights-holders can assert over their intellectual property. Typically the rights are expressed using a Rights Expression Language (REL) such as the Open Digital Rights Language (ORDL) [ODRL], the XML based XrML [XrML] or the MPEG-21 REL [Smith 2004]. The concerns over DRM systems relate to the extent to which they erode the principle of "Fair Use" [Coyle 2004] that Libraries rely on heavily in order to make accessible to the general public materials that are under copyright. The major problem being that the ambiguity of the term "fair use" is difficult, if not impossible, to express in RELs. The Digital Millennium Copyright Act [DMCA 1998] further backs copyright owners' right to encode their own intellectual property regime into a digital, machine-processible format.

The erosion of the "fair use" doctrine has to some extent resulted in the more flexible Creative Commons license scheme for digital materials:

> "Creative Commons defines the spectrum of possibilities between full copyright — *all rights reserved* — and the public domain — *no rights reserved*. Our licenses help you keep your copyright while inviting certain uses of your work — **a "some rights reserved" copyright**."

The emergence of a globally connected information environment has also led to the investigation of federated DRM [Martin et al. 2002].

6. Methods and Processes to Enhance Semantic Interoperability

It is important to get an overview of the different methods, processes and techniques in use to enhance semantic interoperability. Standardization and translation approaches need to be covered for all three information levels: data structures, categorical and factual data (cf. Section 3.3 and 3.4).

Advantages and disadvantages of the standardization and the two translation approaches, mediation and data warehouse approach, are described in Section 5.1. The translation approaches are implemented via mapping of source schemata to a global schema. As in database integration techniques, the target schema can be a fixed (employing schema integration and modular approaches) or a constantly adapted schema (requiring continuous mapping, matching and translation).

The use of foundational and core ontologies for the mapping between schemata and the role of core ontologies when improving interoperability between different KOS is outlined and leads to the second half of this sections overview of methods.

Whereas 5.1 focuses on the level of data structures, section 5.2 looks at the level of categorical data, describing methods applied to KOS, their concepts, terms and relationships.

Based on a published comprehensive empirical study, a more detailed account of methods to enhance semantic interoperability in information systems based on the usage of KOS is provided. We look first at the multitude of methods used recently and predominantly in the field of Library and Information

Science (LIS). Approaches in the Ontology and Semantic Web communities are then outlined. Our comparison finds a broader overlap when it comes to the translation approaches than in the area of standardization approaches between the two communities. Finally, we try to sketch an integrated view.

6.1 Standardization of metadata schemas, mediation and data warehousing

As described in section 3.3, there is a general trade-off between standardization and translation. Basically, they are both, competing and complementary principles. The fundamental question is, which degree of semantic equivalence the user wants or can afford to support. In an information discovery scenario, the requirements for exactness are relatively low; each improvement of recall combined with a reasonable precision is regarded as progress. An inexact equivalence of more systems might be more effective than an exact equivalence of fewer systems. In scientific research scenarios one may require that the answer to a query has statistical value and is exhaustive. In this case, just the opposite holds.

Current focus in digital library systems is on the first, with minimal standards such as the Dublin Core. The other side should be given equal attention.

The choice between standardization of metadata schemas, mediation and data warehousing depends on social factors, economic factors, technical feasibility, and convergence of technical solutions, innovation rate and the above quality considerations.

6.1.1 Standardization

The process of standardization requires an already emerging good practice or convergence of solutions in a wide application field, which is also an expression of underlying semantic homogeneity. Further there must be an economic benefit in using the standard.

For instance, metadata in the proper sense (data about data) are a smaller component of the overall data holdings. In particular bibliographic metadata schemata are widely independent from the subject matter the described literature is about. Being mostly determined by the library processes and information discovery processes; they constitute good candidates for standardization. Nevertheless, fiction and non-fiction may already be treated differently with respect to the subject, whereas other scientific and collection data may employ very different schemata.

The standardization process needs to identify those parts of the semantics behind the various metadata and data structures that are

- common to multiple applications,
- stable and widely used,
- and form a functionally complete unit.

Standardization of elements not mature enough is counterproductive, and reduces the value and hinders adoption. A very “cautious” standard is Dublin Core, which is frequently accused of being an over-simplification. The CIMI application profile, for instance, might be regarded in some of its parts as an example of doing too much. A good standard should leave space for compatible extensions. See also 5.1.5.

Semantic interoperability is typically an economic driving force, particularly for those players that do not dominate a market. Strong players in the market may try to fight off competition by not following standards. Innovative players may be hindered by a standard in exploiting the added value of their solutions. This demonstrates that standardization has a strong social component. There are two solutions:

1. Bring all key players together and find consensus.

2. Form a core group of sufficiently representative players that elaborates an optimal solution and convince a growing community by its quality.

The second process has its risks and may be slower, but it is often cheaper and may yield results of higher quality and longer validity.

In order to achieve semantic interoperability among data described by data structures that cannot be integrated into one common standard due to

- lack of maturity
- lack of consensus or
- cost of adaptation,

two translation-based approaches are possible: mediation or data warehousing. In the case of the digital library, *one can fairly assume that there is no requirement of updating local systems via a global access system*. This reduces greatly the complexity of translation-based approaches.

6.1.2 Mediation

Mediators [Wiederhold] use a virtual global schema that generalizes over the relevant parts of the source schemata. Global access is achieved by translating each query into the local schemata, collecting the answers and attempt integration of the received answers in real time ("on the fly"). The sources are not changed, but a source specific so-called wrapper software takes over communication of the local source with the global access system. This approach is technically the most costly one.

Its advantages are:

- no change of source system
- immediate response of source updates (e.g. ideal for distributed booking systems)
- high scalability if the wrappers are intelligent enough

Its disadvantages are:

- conversion of both, queries and answer sets
- inefficient joins across source systems
- inefficient data integration.

The last point is particularly important if the sources are expected to contain many duplicate entries that are referred to in different ways. Failure to detect such duplicates in the sequence makes joins over such items impossible. E.g. one may like to query the transitive closure of co-authorship in order to find out clusters of scientific subjects. Mediators are the choice for a low degree of heterogeneity, and sources with high standards of unique reference to identical items.

In a recent project [Amato 2004], a new variation of the mediation approach was presented. Heterogeneous metadata following several different standards, i.e. DC, SCORM, MPEG7 are stored as given by the data provider in one digital library. The digital library access system internally mediates requests from outside expressed in any of these formats to access all documents in the digital library. Obviously this works well for a limited number of similar formats.

6.1.3 Data warehouse approach

Alternatively, in a data warehouse-like approach one may extract in regular intervals all relevant source data and integrate them into a separate database. The best-known example of this technique is the harvesting of metadata. Since the data are not collected and queried at the same time, data can be transformed by a multi-step process which can be customized to special problems at low cost. Even semi-automatic algorithms can be applied. This is the major advantage of this method. In particular, complex algorithms of duplicate detection and elimination can be applied effectively.

The advantages of the data warehouse approach are:

- no change of source system
- efficient joins across source systems
- efficient data integration

Its disadvantages are:

- translation of all (also not-requested) data
- complex update of the data warehouse if source data are deleted
- limited scalability due to central storage

Since data storage media are cheap, the duplication of metadata is no issue. The data warehouse approach is the choice for data that mostly increases but rarely is deleted, that contain many duplicates in different encoding, and for high degrees of heterogeneity. One should regard this situation as the most characteristic for digital libraries in a wider sense. In order to increase scalability, mediators can connect multiple data warehouses with compatible schemata, in order to combine the virtue of both approaches.

Mediators and data warehouses are implemented via mapping of source schemata and the global schema, which intends to produce a specification (or *functor*) to translate data encoded in a source schema into data encoded in the target schema under preservation of meaning (this subsumes the term "crosswalk").

In database integration two forms are distinguished: a fixed target schema, or a target schema which may be adapted for each new source schema if necessary in order to optimally integrate it. This distinction is also known as the Global-as-View (GAV) approach, in which a global schema is defined in terms of the source schemata, and the Local-as-View approach, in which a global schema is defined independently from the source schemata.

In a GAV approach, query reformulation reduces to simple rule unfolding (standard execution of views in ordinary databases). However, changes in information sources or adding a new information source requires an administrator to revise the global schema and the mappings between the global schema and source schemata. Thus, GAV is not scalable for large applications. LAV scales better, and is easier to maintain than GAV because the global schema is independent of the source schemata. Recently, all sorts of combinations of both approaches are discussed [Koffina et al. 2004].

Making a practical distinction, we deal in the following with GAV under "schema integration", whereas the LAV approach is discussed in 5.1.5 as a mapping approach.

6.1.4 Schema integration and modular approaches

Schema integration in the sense of GAV requires a highly controlled environment and frequently substantial changes in the source systems in order to come up with a satisfactory global schema. This is typically the limiting factor of this approach. In the sequence, access and distribution techniques are relatively simple and well understood. The GAV approach preserves optimally source semantics, such that very high requirements for the preservation of source semantics can be fulfilled. However the global schema may tend to become so complex, that common semantics might be hidden in different integrated structures, or at least querying for more abstract properties may become impractical for the user.

A modular approach to organising a schema is through the notion of an *application profile*. In broad terms, application profiles are a type of element set that draws on metadata terms from extant vocabularies and customises them for a local or specific application [Heery 2000]. The concept of application profiles first emerged from the DESIRE project. Within the DESIRE and SCHEMAS projects (see section 4.5.3) application profiles were used as a means of disclosing terms that had been used in particular applications to facilitate the reuse of terms and thereby enhance the potential for

interoperability between disparate systems. Additionally, the SCHEMAS project made significant early advances with regard to the formulation of application profiles in a machine processible format using RDF Schemas [Baker 2001]. This work continued to mature in the MEG Registry project. Differentiating between element sets and application profiles is a useful means of distinguishing where and how terms are defined as opposed to how they are used and adapted in practice. Application profiles encourage the modular organisation and structuring of knowledge. Element Sets declare a unique set of terms and definitions and in effect make apparent semantic knowledge which is available for re-use, the ideal being that terms are identified by means of Uniform Resource Identifiers (URIs), preferably persistent.

Machine processible encoding of application profiles makes possible automated data mining and querying across vocabularies, making apparent emerging trends and patterns in metadata vocabulary and term usage, aiding the process of consensus building and hence semantic interoperability.

6.1.5 Mapping, matching and translation

In the LAV approach, the mapping is a non-trivial interpretation of a source schema in terms of the global schema. Even though there are many attempts to automate this process, it is still mainly in the hand of the expert to define and to decide which the best interpretation of his/her source is in terms of the global schema.

One can distinguish two forms: The *core approach*, in which the global schema represents a minimal global denominator, and a maximal approach, in which the global schema is a common semantic generalization over the sources.

In a core approach, such as Dublin Core, data under the global schema can mostly be represented as joins of data paths or deductions of the local schemata. This facilitates a mediator approach, but there is a considerable loss of semantics. Therefore DC is discussed as a “finding aid”.

In a *maximal approach*, the only practical solution is that the global schema provides an appropriate set of primitives, by which the source schemata can be described; otherwise the global schema would become impractically large. This is the approach of the CIDOC CRM. As a result, the mappings can mostly be represented as joins of data paths or deductions of the global schemata. Such a mapping can preserve the granularity of aggregation of links and nodes of the most analytical source, e.g. the analytical creator – creation relationship of a sculptor and a bronze statue (model-mould-cast).

Once the granularity is preserved, finer semantic distinctions can be introduced by adding suitable typologies to better map classes and relationships. Those typologies should be described in KOS (see section 5.2).

In practice, no one hinders in principle to extend the global schema in an LAV approach, so that the integration is improved. However, those extensions should not affect existing mappings in order to preserve the scalability of the approach. For that reason, the initial global schema should be carefully crafted from the beginning with such extensions in mind. Necessarily it must employ subsumption for classes, properties and relationships.

Depending on the quality to be achieved, the mapping may be such that

1. the global schema *contains* the semantics of the local schema, i.e. each source query can be replaced by an equivalent global query.
2. the classes and properties of the global schema *subsume* those of the source systems, i.e. each source query answer set is a subset of a respective global query answer set.
3. the global schema is an approximation of the local schema, i.e. each source query answer set has a relevant overlap with a respective global query answer set.
4. the global schema preserves the constraints (in particular cardinality constraints) of the source schemata.

Quality (1) can hardly be achieved in a LAV scenario, and quality (2) is typically sufficient. Obviously, quality (2) is still sufficient for statistically relevant and exhaustive query processing of the integrated sources. Quality 3 supports only discovery scenarios. Quality (4) is typically useless, because there is *no update requirement* from the global access system. It is rather *counterproductive*, because it will create incompatibility of more and less constrained sources and does not contribute to either recall or precision. Constraint enforcement should be seen as a data acquisition problem. One may recommend applying quality (2) for a maximal approach, and quality (3) for a core approach.

Finally, mappings can be distinguished by the source from which semantics are drawn. Those sources may provide the complete mapping, or provide an incomplete set of equivalences between source and target constructs a schema matching. The most important sources are:

- Elicitation of expert knowledge
- Indication by linguistic similarity of class and property names
- Comparison of identical data items stored in equivalent constructs
- Structural similarity of source and target constructs
- Background knowledge and other mappings

Since the manual mapping process is one of the hardest bottlenecks in information integration (as e.g. experienced by RLG in the Cultural material Initiative, or the Canadian Heritage Information Network), recent research is targeted at automating it using the above sources. However it comes to a problem of understanding real world semantics behind the used data structures. Since, at least in the shorter term, manual intervention and control is necessary, there should be a systematic investment in generically semi-automatic methods, which at the moment are unfortunately quite rare.

6.1.6 Usage of Foundational and Core Ontologies

If we look at the mapping process as an a priori intellectual problem, we encounter the difficulty of understanding the intended meaning of both, the source and the target schema constructs. Even a good description of both admits some degree of freedom of interpretation, but frequently there is no more documentation than a field name. In practice, a schema that is for a longer time in use may be used for situations initially not foreseen, so that interpretation changes with use.

In few words, we have at least four mental models: The source creator's, the source user's, the target creator's and the target user's. Hopefully, all four share basic conceptualizations of their world, else information integration would make not much sense. Experience from mapping exercises show that a good specification of a common conceptualization, i.e. a core ontology, can be very helpful as a common language to express intended meaning. In particular, it allows not only to state overlaps between source and target constructs, but also to objectify the differences and the reasons for these differences.

Foundational ontologies play a particular role by offering ready made logical formulations of basic semantic distinctions such as different kinds of parthood, identity, unity etc. [Masolo 2003], which are typically intuitively assumed but not available to people as conscious constructs. Often they can be used to explicitly explain intuitively felt differences between different people's concepts.

Experience from the CIDOC CRM and ABC Harmony [Doerr, Hunter, Lagoze 2003] harmonization shows that it is not only a question of having just one ontology. Different ontologies may be more or less suited to express the common concepts in the different applications that allow for an integration of the respective views. E.g. for a news integration system, a distinction of a hero from a criminal might be less suited than the distinction of people having attracted public attention from those not having. An ethical information system may make just the opposite choice. More practical importance for digital library schemata have distinctions such as material and immaterial items, in contrast to tangible and intangible items etc.

In other words, the core ontology should be suitable for the intended aspect of integration, and both the mapping and the ontology should make clear the assumed functionality they support.

A core ontology may be a basis to create a global schema itself. In that case, it is not only helpful to harmonize the mappings from different sources to a common interpretation of the target, but it can also be used as an intermediate to support mapping of a set of sources to multiple target schemata, in particular standardized schemata.

Explicit, exhaustive annotation of the intended meaning of a number of source and target schemata by a common ontology finally opens the way to apply algorithms to automatically determine compatibility between schemata and to develop automatic generators for mapping algorithms.

6.1.7 KOS Compatibility with Core Ontologies

Consider two schemata describing archaeological objects: Schema A contains a field *object type*; schema B contains a field *material* and another *function*. For schema A, a KOS defines the term *flint scraper*. For schema B we need two terms: material *flint* and function *scraping* in two different hierarchies. This example demonstrates how a general KOS for object types may become useless for a more analytical schema. The solution is to analyze the object terms into the function and material facet as appropriate, i.e. *scraper* = object which is *made for*: *scraping*, and flint *scraper* = *scraper* which is *made from*: *flint.*

Another complication is that the context of use of a KOS may lead to generally not applicable conclusions. E.g. a hierarchy of functions with objects in mind may come to the conclusion that *liquid storage* is a generalization of *liquid transport*. This is true for the object made for this function, but not for the function itself.

A very frequent example is the transition from object to subject: Whereas talking about *bridges* implies talking about *bridge construction*, *bridge construction* is not a kind of *bridge*! (See LCSH).

Another case is the following: Schema A describes archaeological objects, i.e. virtually any physical object. Schema B describes coins. Schema C describes objects with monetary value, i.e. coins, paper money, Kauri shells etc. This poses two problems:

a) In order to retrieve equivalent results from schema A and C, a mediator or translator must be able to associate the schema C as a whole with a term characterizing “object with monetary value” (*money* in the AAT) and this term must be part of the hierarchy of terms in a KOS employed for schema A to describe the object types.
b) Coins may include medals or not. Depending on such choice, schema B is subsumed by schema C or not.

In order to make KOS independent from idiosyncratic forms of use, a core ontology must be used to identify common facets and use contexts (such as “subject”), and to analyze and harmonize the KOS with respect to the dominant facets of the domain. In turn, it must be possible to explain the meaning of metadata and other data fields in terms of the core ontology or other high-level concepts shared by the respective KOS. Under this condition KOS servers may be used as independent Web services that allow for mediating terminology between systems using different schemata.

6.2 Methods applied to KOS, their concepts, terms and relationships

Building upon the theoretical considerations in Section 3.3 and 3.4 primarily, we intend in this section to provide a more detailed account of methods to enhance Semantic Interoperability in information systems based on the usage of KOS. We will focus on approaches, which have been used recently,

predominantly in the field of Library and Information Science (LIS), trying to occasionally show obvious parallels from the ontology and semantic web communities.

This overview draws heavily on papers by Zeng and Chan, Visser, the HILT project and Doerr [Zeng 2004, Visser 2004, HILT 2001, Doerr 2001].

6.2.1 Approaches as found in LIS contexts

Traditionally, the principle options for improved semantic interoperability have been described as

a) integration of existing KOS,

b) mapping between KOS, or

c) creation of a new KOS.

The investigators of project HILT, devoted to improvements of terminology services for the JISC Information Environment in the UK, indicate the multiplicity of possible solutions in a large but rather unsystematic matrix of nine main options and numerous second level options and combinations with capability enhancements such as:

- adding thesaural structure,
- building new scheme-specific micro-thesauri,
- mapping to existing domain-specific micro-thesauri,
- adding mappings to local terms,
- ensuring multilingual capability,
- allowing community control,
- machine-assisted processing,
- AI-assisted processing,
- providing user training,
- providing flexible facilities to aid users,
- facilitating user mind maps,
- ensuring consistent application of indexing terms via training and/or monitoring,
- providing user assistance for optimal retrieval, terminologies interoperability agency.

[Nicholson, D., Wake, S., & Currier, S. (2001) and HILT (2001) Tab. 2]

Here are both KOS and indexing enhancements, involvement of several KOS, processing enhancements, and search and user support measures listed. Different actors could apply one or several options in combinations and sequences of actions (e.g. adopting a mapping service in the short term and compiling a single scheme in the long term) are possible.

Regarding the decision process for a given service, Zeng and Chan summarize:

"The choice of a basic approach plus any combination of the possibilities mentioned above may bring various end-products and require different amounts of time and resources. Any method and

combinations with other processes may have pros and cons. When a particular method is employed, it is necessary to conduct a comprehensive investigation in order to identify potential problems."

We want here to take a further step towards a more systematic description of the available options, especially as far as KOS are directly involved.

After analysing many recent projects and services, Zeng and Chan [Zeng 2004] carry out a methodological analysis of vocabulary association and integration resulting in the following alternatives (cf. the descriptions, illustrations and examples in the paper):

1) Derivation/Modeling

A specialized or simpler vocabulary is developed with an existing, more comprehensive vocabulary as an initial model.

Recent example: Facet analysis can play a key role in facilitating semantic interoperability by deconstructing and systematising complex, pre-coordinated Subject Headings that might otherwise prove intractable for mapping purposes. Facets (almost always) constitute mutually exclusive groupings of concepts. Single concepts from different facets are combined together when indexing an object - or forming a query. If Subject Headings can be de-coordinated into individual concepts then the mapping exercise can potentially be simplified. The OCLC FAST project [FAST] has taken some initial steps in this direction by investigating automatic means to convert LCSH headings via a simplified syntax into a faceted representation. Within DELOS, this line of research is being developed by the JPA-2 project, *Ontology-driven interoperability*. This will include a study of the relationships that govern formation of valid (and useful) compound terms and how to combine such terms dynamically. The eventual aim is a principled and formalized treatment of facet structure and the syntactical relationships underlying the composition of concepts.

2) Translation/Adaptation

Terms translated from a controlled vocabulary in a different language.

3) Satellite and Leaf Node Linking

Specialized thesauri are treated as satellites of a super-structure often made accessible via an integrated search interface. Leaf nodes in a tree structure can be used to link to a specialized vocabulary for sub-topics of the node.

4) Direct Mapping

Establishing equivalence between terms in different controlled vocabularies or between verbal terms and classification numbers. Usually has an intellectual review component.

Ex.: Renardus [Koch, Neuroth, Day 2001]; OCLC's mappings involving LCSH and DDC [Vizine-Goetz et al 2004].

Ex. Ontology merging tools: Anchor-PROMPT [Noy and Musen 2001] as a plug-in to Protege-2000; Chinaera, an interactive merging tool based on the Ontolingua editor.

5) Co-occurrence Mapping

Works at the application level, i.e. in metadata records, where the group of subject terms from different vocabularies (or from free text) actually results in loosely-mapped terms to be used for mapping between vocabularies or directly for retrieval.

6) Switching

In translating equivalent terms in different vocabularies, a switching language may be used as an intermediary. It can be a new system (e.g. UMLS Metathesaurus) or an existing system (e.g. DDC in project Renardus [Renardus]). The anchor terms can reside in an Interlingua. A classification structure as the backbone enables hierarchical browsing.

This is one of the most frequently used approaches.

7) Linking Through a Temporary Union List

Terms that are not conceptual equivalents but are closely related linguistically may be linked to enhance retrieval. It is only a temporary linking, based on word matching, corresponding to the terms from a query (ex.: MACS). No new vocabulary product is created, not even a concordance of words.

8) Linking Through a Thesaurus Server Protocol

Establishing a linked environment through a thesaurus server protocol (ex.: Alexandria Digital Library Thesaurus protocol). Different local thesauri answer queries for certain terms. No new vocabulary product is created.

We might add to Zeng and Chan's presentation, that there seem to be variants and combinations of these methods, e.g. the way the Californian CERES project [CERES] and the National Biological Information Infrastructure (NBII) merged and integrated different KOS to develop an Integrated Environmental Thesaurus supported by a Thesaurus "Networking" Tool Set. This is a merging and creation of a vocabulary that combines and transcends methods 1 and 3 above.

Other examples might be the "creation" of the European GEMET thesaurus. General Environmental Multilingual Thesaurus [GEMET] or the planned construction of a Multilingual Mapped Forestry Thesaurus by the Global Forest Information Service (GFIS)/FAO via mapping between concepts and terms from AGROVOC, CABI and several other thesauri.

A parallel when it comes to classification could be the FAT-HUM Classification to be created by the University College London FATKS project integrating the best features of BBC (Bliss), UDC and BSO [FAT-HUM]. The purpose is to create a faceted classification in the areas of religion and visual arts, building on the three distinct but closely interconnected classifications of concepts:

- broad classification representing the universe of knowledge: sciences, established disciplines and subjects
- more detailed faceted classification tested in two areas of humanities: religion and visual arts
- classification of generally applicable concepts (common auxiliaries).

Zeng and Chan differentiate as well between approaches for establishing interoperability among different existing KOS according to the structures and characteristics of these KOS: KOS of different structural types (e.g. classifications vs. thesauri) and KOS of similar structural types.

Implementers have to be aware of the both theoretically and practically more complicated nature of the former approach and have to avoid errors emanating from treating the different types of KOS as having identical mechanisms, elements and purposes.

Factors that influence how successfully one vocabulary can be associated with another [from Vizine-Goetz et al 2004, referring to Lancaster and Smith (1983), Doerr 2001 and others]:

- Extent of overlap in the subject matter
- Level of specificity of terms
- Degree of pre/post-coordination

- How the vocabulary codes equivalence, hierarchical, and other relationships
- Differences in word use, e.g. common versus scientific names
- Differences in meaning resulting from different classifications of terms

In order to store and manage the links which may be established by the mapping approaches, several options have been used:

a) Special fields in authority records (ex.: MARC 21 Authority Format)

b) Concordance tables carrying semantic relationships and equivalences, managed through database software

c) Semantic Network as a backbone for clusters of equivalent terms where often semantic types are assigned to (UMLS Semantic network).

d) Lexical Database like WordNet with semantic relations between synonym sets.

6.5.2 Approaches as found in Ontology and Semantic Web contexts

In an information integration system each information source has an ontology describing the meaning of the contents (this section follows Visser's tutorial [Visser 2004] closely). The integration takes place either via a common ontology or via fixed mappings between different ontologies. Required are

a) specialized tools and editors to support the process of building an ontology and

b) an ontology language based on Description Logics and subsumption reasoning for the computation of relations between information sources and for the validation of the integration results.

Several different techniques can be used in order to compare meaning:

1 Semantic matching with synonyms:
Word matching on the basis of synonym sets (e.g. WordNet); disambiguation
Encoding: OWL [OWL] notions of synonymy: equivalentClass; equivalentProperty; sameAs

2 Matching with taxonomies (synsets in a hierarchy)
Visser points to approaches based on the length of the connecting path between synsets or the amount of shared linguistic information, to the use of concept lattices to infer relations between concepts and to semantic matching e.g. equivalence, overlapping or mismatch between the extension of concepts at a given node.

3 Disambiguation of homonyms using top-level ontologies (e.g. Cyc [Cyc], DOLCE [DOLCE]) for establishing the global context.

4 Comparing feature sets (attributes, parts, functions) of concepts.

5 Matching by classification of large sets or complex structures into the goal taxonomy (e.g. with Description Logics as in OWL).

Three general approaches (architectures) are used for ontology-based information integration:

a) single-ontology approach where all information sources are linked to a single ontology which can be a global ontology with shared vocabulary or a combination of modules from different ontologies establishing a similar view of the domain.

Ex: OntoBroker [OntoBroker] (frame-based representation system)

Methods: structural enrichment (structural resemblance, definition of terms); meta-annotation

b) multiple-ontology approach where each information source has its own ontology, shared vocabulary is missing and mapping is performed between the different ontologies. Heterogeneous views are supported. Inter-ontology mappings are however hard to define in reality and costly.

Ex: Observer [OBSERVER] (pure Description Logic language CLASSIC)

Methods: defined mapping: translation between ontologies, 1:1 mappings between classes and values; lexical relations: quantified inter-ontology relationships from linguistics; top-level grounding: relate all ontologies to a top-level ontology; find semantic correspondences.

c) hybrid approach where ontologies of single information sources are built using elements from one global shared vocabulary which makes the ontologies comparable.

Ex: MECOTA/BUSTER [BUSTER] (pure Description Logic language OIL, OWL)

Visser concludes that:

- the mapping between ontologies still is ad-hoc or arbitrary rather than well-founded
- there is a need for investigation on a theoretical and empirical basis
- there is a lack of methodologies supporting the development and use of ontologies
- the methodology should be language independent

6.2.3 Integrated view

The methodological approaches of different communities active in the field of semantic interoperability appear not to be dramatically different.

This report's theoretical considerations (Section 3.3) differentiate between two main routes to achieve semantic interoperability in Digital Library environments: Standardization and Interpretation.

1) Proactive standardization, information transformation: using the same language (plus: extensible or modular standards with interlingua/core).

This seems to correspond to Visser's a) Single ontology approach. The mentioned techniques are, however, rather different from LIS techniques. Zeng's methodologies 1-3 belong to this group, even subsuming KOS merging.

LIS:

- Derivation/Modeling
- Translation/Adaptation
- Satellite and Leaf Node Linking

Ontology:

- structural enrichment (structural resemblance, definition of terms)
- meta-annotation

2) Reactive interpretation, information integration: translation between languages (common switching language reducing the number of interpreters needed).

This seems to correspond to Visser's b) Multiple-ontology approach listing similar techniques as applied in LIS.

Zeng's methodologies 4-8 belong to this group, even subsuming KOS correlation.

LIS:

- Direct Mapping
- Co-occurrence Mapping
- Switching
- Linking Through a Temporary Union List
- Linking Through a Thesaurus Server Protocol

Ontology:

- Related to Direct Mapping: Defined mapping: Translation between ontologies, 1:1 mappings between classes and values;
- Related to Direct Mapping and to Linking: Lexical relations: quantified inter-ontology relationships from linguistics;
- Related to Switching: Top-level grounding: relate all ontologies to a top-level ontology;
- Related to Linking: Find semantic correspondences.

7. Semantic Interoperability in Digital Library Services

A considerable amount of work is being undertaken and investigated in the area of automated and semantically enhanced library services. Although some user level services are now beginning to appear, it is notable that much of this work is still at the stage of developing demonstrators and prototypes. Typical digital library services include the following:

- Searching, browsing, navigation
- Cross searching and cross browsing
- Brokerage services
- Information tracking
- Transformation of data
- User interface design
- (Semi-)Automatic indexing and classification
- Mapping services
- Translation support for multiple languages

Below we consider some of the work that has been published. It should be noted that the researchers and investigators concentrate on differing areas of traditional library services and their focus on automation and advanced services varies considerably. Tudhope [Tudhope 2004] provides a useful overview of technological solutions to terminology services as well as an analysis of demonstrators and services being developed in the UK.

For example, Renardus [Renardus] is a distributed Web-based service, which provides integrated searching and browsing access to quality controlled Web resources from major individual subject gateway services across Europe (funded by the EU's Information Society Technologies 5th Framework Programme until 2002). Renardus uses a generic broker architecture, which is achieved using the Z39.50 search and retrieval protocol. All levels of semantic interoperability mentioned in this report (Section 3.4) have been applied:

1. Data structures are integrated via mapping and adapting the metadata elements of all contributing gateways to a common Renardus application profile that specifies the required data fields, their semantics, syntax, encoding systems and cataloguing rules.

2. Factual data has been integrated to a minor degree, via common encoding rules for dates, a common document type vocabulary and similar.

3. Categorial data has been integrated by intellectual bilateral mappings from the common DDC classification system to all different classification systems in use in the participating gateways [Koch, Neuroth and Day].

Standardised mapping relationships are expressed between a pair of classes (and not between a DDC class and individual resources). To support the practical effort, Renardus has adapted a mapping tool developed by the German CARMEN project. The outcome is a unified cross-browsing structure at the Renardus site, identical to the DDC classification system, linking from each DDC class to equivalent parts of the distributed gateways (their local classification structures). Several browsing support features have been introduced, i.e. a graphical fish-eye view of classes topologically surrounding the one actually inspected. In addition, the advanced search system makes use of the DDC mapping to greatly expand both precision and recall of the results of a topical query.

A very early major project aimed at automated classification of Web pages based on a established classification system was the Nordic WAIS/World Wide Web Project [Nordic WAIS/World Wide Web Project 1995; Ardö et al. 1994; Koch 1994]. In this project automated classification of the World Wide Web and WAIS (Wide Area Information Server) databases using Universal Decimal Classification (UDC) was experimented with. A WAIS subject tree was built based on two top levels of UDC, i.e. 51 classes.

GERHARD (German Harvest Automated Retrieval and Directory) is a robot-generated Web index of Web documents in Germany [GERHARD 1998; Möller et al. 1999; GERHARD 1999]. It is based on a multilingual version of UDC in English, German and French, adapted by the Swiss Federal Institute of Technology Zurich. GERHARD's approach included advanced linguistic analysis.

Online Computer Library Center's (OCLC) project Scorpion [Scorpion] built tools for automated subject recognition, using the Dewey Decimal Classification (DDC). The basic idea was to treat a document to be indexed as a query against a DDC knowledge base. The results of the "search" were treated as subjects of the document. Scorpion also used clustering, to refine the result set and to further group documents falling in the same DDC class [Subramanian, Shafer 1998]. The SMART (System for Manipulating And Retrieving Text) weighting scheme was used. The tool was used e.g. to support cataloguers classifying web pages and to support end user searching in CORC. OCLC currently works on releasing FAST (Faceted Application of Subject Terminology) [FAST], based on the Library of Congress Subject Headings (LCSH), which are modified into a post-coordinated faceted vocabulary. FAST could serve as a tool for automatic indexing, similar to the role the DDC database had in Scorpion.

WWWLib [Wolverhampton Web Library] is a manually maintained library catalogue of British Web resources, within which experiments on automating its processes were conducted [Wallis & Burden 1995; Jenkins et al. 1998]. The original classifier from 1995 was based on comparing text from each document to DDC captions. In 1998 each class mark in the DDC captions file was enriched with additional keywords and synonyms.

"All" Engineering ["All" Engineering resources on the Internet] is a robot-generated Web index of about 300000 Web documents, developed within the DESIRE project [DESIRE project], as an experimental module of the manually created subject gateway Engineering Electronic Library (EELS) [Koch & Ardö 2000]. The Engineering Index (Ei) thesaurus was used; in this thesaurus, terms are

mapped to the independent Ei classification system. The DESIRE project proved the importance of applying a good controlled vocabulary in achieving good classification accuracy: 60% of documents were correctly classified, using only a very simple algorithm based on a limited set of heuristics and simple weighting. Another and later variant, Engine-e [Engine-e], used a slightly modified automated classification module to the one developed in "All" Engineering [Lindholm, Schönthal & Jansson 2003].

The project BINDEX (Bilingual Automatic Parallel Indexing and Classification) [HLT Project Fact sheet: BINDEX] focused on classification of journal articles. The aim of the project was indexing and classifying abstracts from engineering papers in English and German, using English INSPEC thesaurus and INSPEC classification, FIZ Technik's bilingual Thesaurus "Engineering and Management" and the Classification Scheme "Fachordnung Technik 1997". Several natural language processing techniques were used.

FACET. Binding and Tudhope discuss programmatic access to KOS services and the requirements that advanced interfaces pose for networked KOS access protocols [Binding and Tudhope 2004]. Illustrations are given from the FACET web demonstrator, which explores how a thesaurus can be integrated into the search interface and the potential of semantic expansion in querying collections indexed with faceted metadata.

APAIS. Australian Public Affairs Information Service Thesaurus (employs slightly extended version of Zthes).

FAO. Zisman et al. discuss experiences from applying Web service wrappers in an 'information bus' approach to the development of a prototype system that integrated various FAO data sources with disparate organisation and structure [Zisman et al. 2000].

FATKS Project. FATHUM - A faceted classification for the humanities [FATHUM].
Faceted web demonstrator from UCL

HILT. The HILT project [HILT] has explored the possibilities of a high-level thesaurus to provide terminology services at the collection level for UK higher educational communities.

OCLC services. Vizine-Goetz et al. discuss results from an OCLC project to create inter-vocabulary associations automatically [Vizine-Goetz et al. 2004]. The case study mapped the ERIC thesaurus to the Library of Congress Subject Headings by encoding the vocabularies according to MARC (MAchine Readable Cataloging) standards, automatically matching vocabulary terms, and storing mapping data as machine links. The OAI protocol is used to provide access to a vocabulary with mappings, via a browser to human users and through the OAI-PMH Web service mechanisms to machines.

Bibster. A Semantics-Based Bibliographic Peer-to-Peer System [Broekstra et al. 2004] describes a peer-to-peer system for exchanging bibliographic data. It exploits ontologies in data-storage, query formulation, query routing and answer presentation.

REMINDIN': Semantic Query Routing in Peer-to-Peer Networks based on Social Metaphors [Tempich et al. 2004]. The authors define a method for query routing that lets peers observe which queries are successfully answered by other peers, memorise this observation and subsequently use this information in order to select peers to forward queries to in the future.

Semantic Search. The authors describe two implemented semantic search systems built on top of TAP [TAP] which augment traditional Web search results with relevant data aggregated from distributed sources [Guha et al. 2003]

A Thesaurus Server. Matthews et al. describe a demonstrator (SophiaM) for storing and browsing concept-based thesauri in the context of the Semantic Web [Matthews et al.].

ONION (Ontology compositION). Mitra and Wiederhold describe a system that enables semantic interoperation among various information sources by articulating the ontologies associated with them [Mitra and Wiederhold 2001]. An articulation focuses in the semantically relevant intersection of information sources (semi-automatic). They concentrate on the large and diverse number of bio-information sources that are available.

GeoXwalk Gazetteer Project. "The principal purpose of geoXwalk is to provide a shared service within the JISC Information Environment (IE) that can underpin geographic searching. The rationale behind the project is that there is currently no unified entry point to assist in geographic searching within the existing academic network as each information provider/service adopts different geographic coding conventions (some use postcodes, others place names, some grid references etc.).The geoXwalk gazetteer should provide researchers and teaching staff with access to an on-line gazetteer for reference and cataloguing purposes."

8. Implications for a Research Agenda

Below is a list of some major research, development and organisational activities, which emanate from this report on issues of semantic interoperability. It is by no means exhaustive or prioritised. It could serve as input for discussions and planning in relevant communities and disciplines.

Schema mapping tools
Schema mapping tool sets, which are intuitive enough for the domain expert to specify the mapping on the intended meaning level and to verify the achieved degree of preservation of the intended meaning in a mapping - and which are formal enough so that the mapping algorithm can be created automatically or with the help of an IT expert with rudimentary or no domain knowledge at all. These tools must use powerful graphical visualization mechanisms and modes of interaction.

Mapping of vocabularies, multilingual terminologies as particular case

Terminology services: practical experience and research

Investigate the operation of such terminology services in different service architectures
e.g. XML Web Services

Thesaurus and KOS protocols need to be improved by:

- possible provision of more complex services, such as semantic expansion (beyond basic broader and narrower expansion) or concept space interchange
- more advanced natural language functionality for identifying controlled terminology in free text (documents or query)
- cross-mapping provision (important for semantic interoperability)
- possible data-dependent filters such as the number of postings associated with a concept.

Develop and agree upon service protocols

Develop a platform neutral access protocol, which is not closely linked to specific KOS representation formats

Develop integrated KOS development tools for distributed usage on the net

Investigate mediation between concept representation in numeric form or in images and in text

Research on the specifics of terminology mining

Develop graphical tools; contextualization tools; visualization tools

Investigate interoperability issues when combining terminology efforts with applications such as search engines, Content Management Systems or web publishing software

Investigate the contribution of KOS to knowledge based interactive tools for the Semantic Web

Create systematic discussion and common research between relevant communities,
e.g. traditional NKOS and Ontology; Semantic Web and Digital Libraries; Library and Information Science, Linguistics, Computer Science, Artificial Intelligence; technical and application/content standard activities etc.

Cooperate with the linguistic and language engineering community

Common developments with the family of ISO TC 37 standards for terminology management, lexicography and computerized terminology

Tools to create, maintain, and deploy data standards

Best practice recommendations in usage of standards

Develop best practice guidelines how to convert vocabularies into digital services and into a suitable and standardised syntax and exchange format; how to provide term/concept level metadata

Conversion from KOS to ontologies (and vice versa?)

Develop common KOS representation formats

Develop and try to agree on a taxonomy of KOS

Research the power of query languages developed for XML, RDF, OWL when applied to KOS

Propose unique identifier standards for concepts, terms and relationships

Develop semantic registries, vocabulary registries for both human and machine usage
Discovery tools for vocabularies

Evaluation and assessment criteria for digital library systems based on achieving semantic interoperability.

Evaluation of vocabularies

Investigate IPR issues related to vocabularies; develop and test license models for such vocabularies

9. Recommendations and Guidelines

There is no doubt that semantic interoperability is crucial to the next generation of digital libraries, and as mentioned earlier in this report, it has been widely identified as such [NSF Workshop, eGovernment 2004, EIF 2004, OntoGov 2004]. It is therefore essential that issues relating to semantic interoperability be considered as an inherent part of future Digital Library research and development

(see section 8). As a result of undertaking this study, the authors would like to make some recommendations and outline some guidelines relevant to the future work of the DELOS2 NoE:

1. DELOS2 should raise awareness of the importance of semantic interoperability within the Digital Library community.
2. The Digital Library community needs to converge on a set of definitions and semantics for various terms to facilitate communication with other communities (see section 4.1).
3. DELOS2 should raise awareness of Digital Library standards and issues amongst other communities.
4. The achievement of semantic interoperability is a multi-level issue affecting many functions of information systems. Sections 4, 5 and 6 of this report cover many of the aspects that should be taken into account when developing a digital library system.
5. Semantic interoperability is relevant to all aspects of information life-cycle management from creation and integration to archiving and preservation (see section 3.2).
6. Digital library architectures should make use of standard access and query protocols as far as possible.
7. Digital library architectures should cater for interoperation and interaction with distributed third-party services such as terminology servers rather than building these functions into the system itself.
8. Digital library architectures should consider making use of service-oriented architecture (SoA) and semantic web services (see section 5.6.3).
9. Any digital library proposing federation, mediation or integration of heterogeneous resources needs to consider issues relating to interoperability, automation and semantic interoperability.

10. In addition, these types of systems need to pay careful attention to their user interfaces in order to hide syntax and structural differences in the underlying systems.
11. Advanced, context-sensitive query processing over heterogeneous information resources requires the matching of concepts. Attention therefore needs to be paid to the development of conceptual models, ontologies and schemas to ensure that this will be possible. The initial global schema should be carefully crafted from the beginning with extensions in mind. Necessarily it must employ subsumption for classes, properties and relationships.
12. We have found that KOS are central to achieving semantic interoperability (see section 6.2). It is therefore crucial that they are adequately accessible and properly managed. Vocabularies, semantic relationships and mappings are information objects themselves, their life cycle: creation, acquisition, collection, modelling, identification, integration, mediation, search, use, maintenance and preservation etc. is of primary importance and a necessary prerequisite to improved semantic interoperability.
13. Experience indicates that we need to abstain from unnecessarily and prematurely narrowing down the "useful" types of KOS to e.g. thesauri and then to largely ignore to use and create solutions for the other types of KOS. Historically a broad and rich variety of different types of KOS have been developed and used with good results.
14. It should be recognised that there are several barriers to the vision of a free and interoperable distributed usage of vocabularies. Vocabularies are still hard to discover and evaluate; most vocabularies are not digitally available at all or digital public versions are far from complete; there is rarely syntactic interoperability; owners do not provide term/concept level metadata; and service protocols are not ready, nor commonly agreed upon.
15. However, probably most important, by far the majority of large and well-maintained vocabularies are not freely available; IPR issues are not clarified; and license models are not developed or tested for such vocabularies.
16. In order to facilitate distributed thesaurus access, a platform neutral access protocol should be used to manipulate thesaurus data. At present, protocols for retrieving thesaurus data are closely linked to thesaurus representation formats.
17. Finally, given the importance of semantic interoperability to future digital library systems, DELOS2 should develop evaluation and assessment criteria for digital library systems based on achieving semantic interoperability.

10. Concluding Comments

This report is a state-of-the-art report produced by DELOS WP5, the cluster concerned with knowledge extraction and semantic interoperability. Our goal was an ambitious one entailing the examination and integration of diverse and disparate work being undertaken in the area of semantic interoperability in digital library systems. The report is necessarily disjoint in some parts, reflecting the diverse approaches and range of work currently in progress. This should not be seen as a failing, but the sign of a healthy research area in which debate is thriving and leading to a better understanding of relevant issues.

A report of this nature cannot be complete or comprehensive in all areas, we are aware that while some areas have been covered in depth, others provide only an overview (depending on the interests and expertise of the authors). Hopefully we have provided an overview of the area and enough information and relevant references for those interested to follow-up various issues in detail.

Within DELOS2 JPAII, work on semantic interoperability in digital libraries is being taken forward in two further tasks:

Task 5.4: Interoperability of eLearning applications with digital libraries (TUC, UKOLN, IU)
Task 5.5: Ontology-Driven Interoperability (TUC, IU, NTNU, ULund, Sztaki, UGlam, AUEB, Imperial, DSTC)

Task 5.4 is focusing on the education domain whilst Task 5.5 is concerned largely with the cultural heritage domain, both tasks will result in a demonstrator. Between them, the two tasks bring together partners from each of the clusters in the DELOS2 NoE as well as additional external partners, promoting further collaboration and integration.

References

"All" Engineering resources on the Internet : a companion service to EELS", (31 January 2003), (EELS, Engineering E-Library, Sweden).
http://eels.lub.lu.se/ae/

"DESIRE project". (30 March 1999) (Lunds Universitets Bibliotek).
http://www.lub.lu.se/desire

"eGovernment Workshop on semantic interoperability. Exchange of Good Practices." Organised by the Norwegian Ministry of Trade and Industry and the Brønnøysund Register Centre in cooperation with the European Commission Brønnøysund Norway, 22-23 June 2004.http://www.brreg.no/workshop/

"Engine-e", (13 February 2004), (Lund University Libraries).
http://engine-e.lub.lu.se/

"FAST : Faceted Application of Subject Terminology", (OCLC projects),
http://www.oclc.org/research/projects/fast/

"GERHARD - Navigating the Web with the Universal Decimal Classification System" (September 1999), (GERHARD).
http://www.gerhard.de/info/dokumente/vortraege/ecdl99/html/index.htm

"GERHARD: German Harvest Automated Retrieval and Directory", (20 July 1998), (GERHARD),
http://www.gerhard.de/

"HLT Project Fact sheet: BINDEX", (14 November 2001), (HLTCentral),
http://www.hltcentral.org/projects/print.php?acronym=BINDEX

"Scorpion", (OCLC software).
http://www.oclc.org/research/software/scorpion/default.htm

<indecs> Framework for Rights Management.
http://www.indecs.org/

2nd Workshop on Semantics in Peer-to-Peer and Grid Computing (SemPGRID), WWW Conference 2004, May 2004, New York, USA.
http://www.cs.vu.nl/~frankh/abstracts/SEMPGRID04.html

ADL Thesaurus Protocol.
http://www.alexandria.ucsb.edu/thesaurus/protocol

Amato, G. Gennaro, C. Rabitti, F. Savino, P., Milos: A Multimedia Content Management System for Digital Library Applications, in Research and Advanced Technology for Digital Libraries, ECDL 2004 Bath, UK, Proceedings, Springer 2004, pp. 14-25.

Archival Resource Key (ARK).
http://www.cdlib.org/inside/diglib/ark/

Ardö, A. et al. (1994), "Improving resource discovery and retrieval on the Internet: The Nordic WAIS/World Wide Web project summary report", NORDINFO Nytt, vol. 17, no. 4, pp. 13-28.

Australian Public Affairs Information Service Thesaurus.
http://www.nla.gov.au/apais/thesaurus/

Baker T., Dekkers M., Heery R., Patel M. and Salokhe G. (2001),
What Terms Does Your Metadata Use? Application Profiles as Machine-Understandable Narratives, Journal of Digital Information Vol 2(2), November 2001.
http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Baker/

Bergamaschi S., Castano S. and Vincini M. (1999), Semantic Integration of Semistructured and Structured Data Sources, ACM Sigmod Record, Vol 28(1) March 1999, pp54-59

Berners-Lee T., Hendler J. and Lassila O. (2001), The Semantic Web, Scientific American, Vol 284(5) May 2001 pp28-37

Binding C., Tudhope D. (2004). KOS at your Service: Programmatic Access to knowledge Organisation Systems. Journal of Digital Information, 4(4),
http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Binding/,
http://www.comp.glam.ac.uk/~FACET/webdemo/

Borgman, C. (2000). From Gutenberg to the Global Information Infrastructure: Access to Information in the Networked World. MIT Press.

Broekstra J., Ehrig M., Haase P., van Harmelan F., Menken M., Mika P., Schnizler B., Siebes R. (2004), Bibster –A Semantics-Based Bibliographic Peer-to-Peer System,

Buckland, M. K.(1991). Information as thing. Journal of the American Society for Information Science 42 (5): 351-360.

BUSTER. Bremen University Semantic Translator for Enhanced Retrieval. http://www.semantic-translation.com

Cabral, L. Domingue, J. Motta, E. Payne, T. and Hakimpour, F. (2004), *Approaches to Semantic Web Services: An Overview and Comparisons*, in Proc. of the 1st European Semantic Web Symposium (ESWS2004)

Canada, Government Information: Thesauri and Controlled Vocabularies,
http://www.collectionscanada.ca/8/4/r4-280-e.html

CERES Thesaurus Protocol and browser.

http://ceres.ca.gov/thesaurus/

Chomsky, N. (1980) Human language and other semiotic systems. In Thomas A. Sebok and Jean-Umiker-Sebok (eds.): Speaking of Apes: A Critical Anthology of Two-Way Communication with Man. New York: Plenum Press, pp. 429-440.

Christensen, E. Curbera, F., Meredith, G., Weerawarana, S. (2001), *Web Services Description Language (WSDL)*, W3C Note 15,
http://www.w3.org/TR/wsdl

CIDOC Conceptual Reference Model (CRM).
http://cidoc.ics.forth.gr/

CORES Registry.
http://www.cores-eu.net/registry/

Coyle K. (2004), Rights Management and Digital Library Requirements, Ariadne Issue 40, July 2004, http://www.ariadne.ac.uk/issue40/coyle/intro.html

Creative Commons.
http://creativecommons.org/

Cui Z., Jones D., O'Brien P. (2002), Semantic B2B Integration: Issues in Ontology-based Approaches, ACM SIGMOD Record, Vol 31(1) March 2002, pp43-48

Cyc. http://www.cyc.com/cyc/technology/whatiscyc

D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati, "Description Logic Framework for Information Integration";, In Proc. of the 6th Int. Conf. on the Principles of Knowledge Representation and Reasoning (KR'98), 1998, pages 2-13

DAML Ontology Library. http://www.daml.org/ontologies/

Dekkers M., "Application Profiles, or how to Mix and Match Metadata Schemas", Cultivate Interactive, issue 3, 29 January 2001. URL: http://www.cultivate-int.org/issue3/schemas/

Dempsey, Lorcan (2003). The recombinant library: portals and people.
http://www.oclc.org/research/staff/dempsey/dempsey_recombinant_library.pdf

Dempsey, Lorcan (2004). Interoperability: the value of recombinant potential (PowerPoint: 2.6MB/28 slides) Presentation given at ARLIS 2004, the 32nd Annual Conference of the Art Libraries Society of North America, April 17, 2004, New York City (USA).
http://www.oclc.org/research/presentations/dempsey/arlis04.ppt

Dempsey, Lorcan (2004). Terms and conditions ... libraries, subject terminologies and the web (PowerPoint: 4.3MB/52 slides) Opening address to Dewey Editorial Policy Committee Retreat, February 16, 2004, Dublin, Ohio (USA).
http://www.oclc.org/research/presentations/dempsey/dewey_20040316.ppt

DESIRE Registry, http://desire.ukoln.ac.uk/registry/index.php3

Digital Object Identifier. http://www.doi.org/

DMOZ: Open Directory Project. http://dmoz.org

Doerr M., Hunter J., Lagoze C., "Towards a Core Ontology for Information Integration", *Journal of Digital Information,* 24(3): 75-92(2003).

Doerr M., Semantic Problems of Thesaurus Mapping. Journal of Digital Information, Special Issue on Networked Knowledge Organization Systems, Volume 1, issue 8, April 2001.

Doerr, M. (2001). Semantic problems of thesaurus mapping. In: Journal of Digital information, Volume 1 Issue 8. Article No. 52, 2001-03-26. http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/

DOLCE : a Descriptive Ontology for Linguistic and Cognitive Engineering. http://www.loa-cnr.it/DOLCE.html

Dublin Core Metadata Element Set, ISO 15386: The Dublin Core Metadata Element Set, http://www.niso.org/international/SC4/sc4docs.html

Dublin Core Metadata Initiative (DCMI) Registry. http://dublincore.org/dcregsitry/index.html
Education Group (MEG) Registry.
http://www.ukoln.ac.uk/metadata/education/regproj/

Efron, M., Elsas, J., Marchionini, G., Zhang, J. (2004). Machine Learning for Information Architecture in a Large Governmental Website. Proceedings of the Fourth ACM/IEEE Joint Conference on Digital Libraries, Tucson, AZ, June 7-11, 2004. pp.151-159

EIF (2004). European Interoperability Framework for pan-European eGovernment Services. http://europa.eu.int/idabc/en/document/3761

Environmental Data Registry.
http://www.epa.gov/edr

Extensible Markup Language (XML).
http://www.w3.org/XML/

Farquhar A., Fikes R. and Rice J. (1996), The Ontolingua Server: a Tool for Collaborative Ontology Construction, Proceedings of Tenth Knowledge Acquisition for Knowledge-Based Systems Workshop 1996

FAST: Faceted Application of Subject Terminology.
http://www.oclc.org/research/projects/fast/default.htm

FATHUM - A faceted classification for the humanities.
http://www.ucl.ac.uk/fatks/database.htm

Fensel, D., Bussler, C. (2002), *The Web Service Modeling Framework WSMF. Eletronic Commerce: Research and Applications*, Vol. 1. (2002). 113-137

Flickr. http://flickr.com

Foundation for Intelligent Physical Agents, Ontology Service Specification,
http://www.fipa.org/specs/fipa00086/

FRBR, Functional Requirements for Bibliographic Records (1998) Final Report. K. G. Saur München, 1998. http://www.ifla.org/VII/s13/frbr/frbr.pdf

Functional requirements for the broker system (D1.3). Use cases developed using the Unified Modelling Language (UML). http://www.renardus.org/about_us/deliverables/d1_3/titlePage.html

G. Karvounarakis, S. Alexaki, V. Christophides, D. Plexousakis & M. Scholl, RQL: A Declarative Query Language for RDF, Proceedings11th International Conference on the WWW, Hawaii, 2002.
Gal A. (1999), SI in Information Services: Experiencing with CoopWARE, ACM SIGMOD Record, Vol 28(1) March 1999, pp68-75

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., & Schneider, L.2002, Sweetening ontologies with DOLCE. in A. Gomez-Perez & V. R. Benjamins (eds.), *Knowledge Engineering and Knowledge*

Management. Ontologies and the Semantic Web, 13th International Conference, EKAW 2002, Siguenza, Spain, October 1-4, 2002, Proceedings. Lecture Notes in Computer Science 2473 Springer 2002, ISBN 3-540-44268-5: pp.166-181.

Garrod, Peter (2000). Use of the UNESCO Thesaurus for Archival Subject Indexing at UK NDAD. In: Journal of the Society of Archivists Vol.21 (1), 2000

GEMET. http://www.eionet.eu.int/gemet/

GeoXwalk Gazetteer Project. http://hds.essex.ac.uk/geo-X-walk/

Gibbons N., Harris S. and Shadbolt N. (2003), Metadata for Agent based Semantic Web services, Proceedings of the 12th World Wide Web Conference, Budapest, May 2003. http://www2003.org/cdrom/papers/refereed/p455/p455-gibbins.html

GovTalk. http://www.govtalk.gov.uk

Guarino N. (1998), "Formal Ontology and Information Systems", pp. 3-15, Proceedings of the first International Conference Formal Ontology in Information Systems (FOIS '98), June 6-8 1998, Trento, Italy, Ed. Guarino N., IOS Press

Guarino N., Carrara M., Giaretta P. (1994), Formalizing Ontological Committment, In Proceedings of National Conference on Artificial Intelligence (AAAI-94), Seattle, Morgan-Kaufman: 560-567

Guarino N., *Formal Ontology and Information Systems*. In N. Guarino (ed.), *Formal Ontology in Information Systems*. Proc. of the 1st International Conference, Trento, Italy, 6-8 June 1998. Amsterdam, IOS Press, pp. 3-15.

Guha R., McCool R., Miller E. (2003), Semantic Search, Proceedings WWW Conference 2003, May 2003, Budapest, Hungary

Heery R., Patel M. (2000), Application profiles: mixing and matching metadata schemas, Ariadne Issue 25, September 2000, http://www.ariadne.ac.uk /issue25/app-profiles/intro.html

Hendler J. (2001), Agents and the Semantic Web, IEEE Intelligent Systems Vol 15(2) 2001, pp30-37

High-Level Thesaurus Project. http://hilt.cdlr.strath.ac.uk/ (with pilot terminologies server)

Hill L., Buchel O., Janée G. (2002): Integration of Knowledge Organization Systems into Digital Library Architectures (Position Paper for 13th ASIS&T SIG/CR Workshop, November 17, 2002), http://alexandria.sdc.ucsb.edu/~lhill/paper_drafts/KOSpaper7-2-final.doc

Hill, L. et al. (2002). Integration of Knowledge Organization Systems into Digital Library Architectures. ASIST SigCR. http://www.lub.lu.se/SEMKOS/docs/Hill_KOSpaper7-2-final.doc

HILT (2001). High-Level Thesaurus Project: Final Report To RSLP & JISC December 2001. 104pp. http://hilt.cdlr.strath.ac.uk/Reports/FinalReport.html

HILT (High-Level Thesaurus) Project, UK. http://hilt.cdlr.strath.ac.uk/

HILT lists: http://hilt.cdlr.strath.ac.uk/hilt2web/Sources/vocabulary.html, http://hilt.cdlr.strath.ac.uk/hilt2web/Sources/thesauri.html

Hodge, G. M. (2000). Best Practices for Digital Archiving : An Information Life Cycle Approach. D-Lib Magazine, January 2000, Volume 6 Number 1. http://www.dlib.org/dlib/january00/01hodge.html

Hodge, Gail (2000). Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. CLIR Pub91. April 2000. http://www.clir.org/pubs/abstract/pub91abst.html

Hunter J. (2001) *Adding Multimedia to the Semantic Web - Building an MPEG-7 Ontology*, in Proc. of International Semantic Web Working Symposium (SWWS) Stanford, July 30 - August 1, 2001

Hunter J. (2002) "Combining the CIDOC CRM and MPEG-7 to Describe Multimedia in Museums", Museums on the Web 2002, Boston, April 2002

IEEE Learning Object Metadata (LOM), http://www.ischool.washington.edu/sasutton/IEEE1484.html

ISO/IEC 11179 standard for data registries.
http://www.iso.ch/iso/en/CombinedQueryResult.CombinedQueryResult?queryString=11179

ISO/IEC JTC 1/SC 29/WG 11/N3966 (2001) "Text of 15938-5 FCD Information Technology – Multimedia Content Description Interface – Part 5 Multimedia Description Schemes", Singapore, 2001

ISO/IEC, *ISO/DIS 21127 -- Information and documentation -- A reference ontology for the interchange of cultural heritage information*, 2004

Java Database Connectivity, JDBC. http://java.sun.com/products/jdbc/

Jenkins, C. et al. (1998), "Automatic Classification of Web Resources using Java and Dewey Decimal Classification", Computer Networks & ISDN Systems, vol. 30, pp. 646-648.

KESI Workshop, Semantic Interoperability in Digital Library Systems, September 2004,
http://www.delos.info/eventlist/semint_bath.html

Koch, T. (1994), 'Experiments with Automatic Classification of WAIS Databases and Indexing of WWW', In Internet World & Document Delivery World International 94, London, May 1994, pp. 112-115.

Koch, T. Controlled vocabularies, thesauri and classification systems available in the WWW. http://www.lub.lu.se/metadata/subject-help.html

Koch, T., and Ardö, A. (2000), "Automatic classification", (11 February 2000), (DESIRE II D3.6a, Overview of results), http://www.lub.lu.se/desire/DESIRE36a-overview.html

Koch, Traugott, Neuroth, Heike and Day, Michael (2001). Renardus: Cross-browsing European subject gateways via a common classification system (DDC). In: "Subject Retrieval in a Networked Environment". Proceedings of the IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing and the IFLA Section on Information Technology, 14-16 August 2001, Dublin, OH, USA. UBCIM Publications - New Series Vol. 25, Muenchen 2003. pp25-33. Manuscript at: http://www.lub.lu.se/~traugott/drafts/preifla-final.html

Koch, Traugott, Neuroth, Heike and Day, Michael (2001). Renardus: Cross-browsing European subject gateways via a common classification system (DDC). In: Subject Retrieval in a Networked Environment, Proceedings of the IFLA Satellite Meeting sponsored by the IFLA Section on Classification and Indexing and the IFLA Section on Information Technology, 14-16 August 2001, Dublin, OH, USA, 25-33. München: UBCIM Publications New Series Vol. 25, 2003.

Koffina I., Serfiotis G., and Christophides V (2004) Foundations for Information Integration: A State of the Art, DELOS Deliverable WP2T2, 01/11/2004.

Lakoff, G."Women, Fire, and Dangerous Things: What Categories Reveal about the Mind", University of Chicago Press, Chicago (1987)

Ledsham, Ian (1999). HARMONICA : Accompanying Action on Music Information in Libraries : HARMONICA II : Deliverable 1.3.3- Recommendations for the Use of Indexing Systems. http://www.svb.nl/project/harmonica/Deliverables/D133.htm

Lindholm, J., Schönthal, T., and Jansson, K. (2003), "Experiences of Harvesting Web Resources in Engineering using Automatic Classification", Ariadne, no. 37, http://www.ariadne.ac.uk/issue37/lindholm/.

MAchine-Readable Cataloging, MARC. http://www.loc.gov/marc/

MARC Code List for Relators, Sources, Description Conventions (of MARC 21). (Part III: Classification Sources and Part IV: Subject/Index Term Sources). http://www.loc.gov/marc/relators/relahome.html

Martin M., Agnew G., Kuhlman D., McNair J., Rhodes W., Tipton R. (2002), Federated Digital Rights Management: A proposed DRN solution for Research and Education, D-Lib Magazine, July/august 2002. http://www.dlib.org/dlib/july02/martin/07/martin.html

Martines J. (Ed.) "ISO/IEC JTC1/SC29/WG11N5525 – MPEG-7 Overview (version 9)", International Organisation for Standardisation (ISO) http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm, Pattaya, March 2003

Masolo C., Borgo S., Gangemi A., Guarino N., Oltramari A. (2003), "Wonderweb Deliverable D18 – Ontology Library", 2003

Masolo C., Borgo S., Gangemi A., Guarino N., Oltramari A., "Wonderweb Deliverable D18 – Ontology Library", 2003

Matthews B., Miles A., Wilson M., "Modelling Thesauri for the Semantic Web", http://www.w3c.rl.ac.uk/SWAD/papers/thesaurus/swdbpapers.html

McGuiness D., F. van Harmelen (2004) *OWL Web Ontology Language Overview*,

McGuinness D., Pinheiro da Silva P. (2003),Registry-Based Support for Information Integration, Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03), August 9 - 10, 2003 Acapulco, Mexico

Miller, P. (2000). I Say What I Mean, But Do I Mean What I Say? In: Ariadne, 23, 2000. http://www.ariadne.ac.uk/issue23/metadata/into.html

Minutes of the 2nd Meeting on FRBR/CRM Harmonization, Heraclion (Crete) 22-25 March 2004

Mitra P., and Wiederhold G., An Algebra for Semantic Interoperability of Information Sources, Proceedings 2nd Annual IEEE International Symposium on Bioinformatics and Bioengineering Conference, 2001

Möller, G. et al. (1999), 'Automatic classification of the WWW using the Universal Decimal Classification', In: McKenna, B. (ed), Proceedings of the 23rd International Online Information Meeting. London, 7-9 Dec 1999, pp. 231-238.

Motta, E., Domingue, J., Cabral, L., Gaspari, M. (2003), *IRS-II: A Framework and Infrastructure for Semantic Web Services*, in: Fensel, D., Sycara, K., Mylopoulos, J. (volume eds.): The SemanticWeb - ISWC 2003. Lecture Notes in Computer Science, Vol. 2870. Springer-Verlag, Heidelberg, pp. 306–318
MPEG-7, Coding of Moving Pictures and Video. http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

Mylopoulos J., Borgida A., Matthias J., Koubarakis M. (1990), TELOS: Representing Knowledge About Information Systems, ACM Transactions on Information Systems, Vol. 8(4), 1990

Napster. http://www.napster.com/

National Health Information Knowledgebase, http://www.aihw.gov.au/knowledgebase/index.html

Nicholson, D., Ali Shiri, Emma McCulloch:
HILT: High-Level Thesaurus Project Phase II Final Report. A Terminologies Server for the JISC Information Environment. Final Report To JISC. 44pp.
http://hilt.cdlr.strath.ac.uk/hilt2web/finalreport.htm

Nicholson, D., Wake, S., & Currier, S. (2001). High-Level Thesaurus project: Investigating the problem of subject cross-searching and browsing between communities. In C.-C. Chen (Ed.), Global Digital Library Development in the New Millennium: Fertile ground for distributed cross-disciplinary collaboration (pp. 219-226). Beijing: Tsinghua University Press.

NKOS taxonomy (2000): NKOS: Taxonomy of Knowledge Organization Sources/Systems (draft) (June 2000). http://nkos.slis.kent.edu/KOS_taxonomy.htm

NKOS: Networked Knowledge Organization Systems/Services. http://nkos.slis.kent.edu/

Nordic WAIS/World Wide Web Project, (14 February 1995), (Lund University Libraries), http://www.lub.lu.se/W4/

Noy, N.F., Musen, M.A. (2001). Anchor-PROMPT: Using non-local context for semantic matching. Workshop on Ontologies and Information Sharing at the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-2001), Seattle, WA, 2001.
http://smi-web.stanford.edu/pubs/SMI_Abstracts/SMI-2001-0889.html

NSF Post Digital Libraries Futures Workshop: Wave of the Future, Chatham, Massachusetts, June 15-17 2003

OBSERVER (Ontology Based System Enhanced with Relationships for Vocabulary hEterogeneity Resolution). http://sid.cps.unizar.es/OBSERVER/

OCLC Metadata Switch Project. http://www.oclc.org/research/projects/mswitch)

Omalyenko, B., Crubezy, M., Fensel, D., Benjamins, R., Wielinga, B., Motta, E., Musen, M., Ding, Y. (2003), *UPML: The language and Tool Support for Makiing the Semantic Web Alive*, In: Fensel, D. et al. (eds.): Spinning the Semantic Web: Bringing the WWW to its Full Potential. MIT Press, pp. 141–170

OntoBroker. http://www.ontoprise.de/products/ontobroker_en

OntoGov (2004). Ontology-enabled e-Gov Service Configuration, EU Project.
http://www.ontogov.com/

Open Digital Rights Language Initiative. http://odrl.net/

Open Knowledge Base Connectivity, OKBC. http://www.ai.sri.com/~okbc/spec.html

Osthaus, Britta, University of Exeter, http://www.ex.ac.uk/~bosthaus/

Ouksel A.M. and Sheth A. (1999) Semantic Interoperability in Global Information Systems, ACM SIGMOD Record, Vol 28(1) March 1999, pp 5-12

OWL - Web Ontology Language. http://www.w3.org/2004/OWL/

OWL Web Ontology Language Use Cases and Requirements. W3C Recommendation 10 Feb 2004. Jeff Heflin, ed. http://www.w3.org/TR/webont-req/

OWL-S Coalition (2003), *OWL-S 1.0 Release*, http://www.daml.org/services/owl-s/1.0/

Pan J., Cranefield S. and Carter D. (2003), A Lightweight Ontology Repository, Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems, 2003

Patel M. and Duke M. (2004), Knowledge Discovery in an Agents Environment,
The Semantic Web: research and applications, Proceedings First European Semantic Web Symposium, ESWS 2004, Heraklion, Crete, Greece, May 10-12, 2004

Persistent Uniform Resource Locator. http://purl.oclc.org/

Pinker, S. "The Language Instinct", New York: W. Morrow and Co (1994)

Qin, Jian & Paling, Stephen (2001) "Converting a controlled vocabulary into an ontology: the case of GEM" Information *Research*, **6**(2) Available at: http://InformationR.net/ir/6-2/paper94.html

Renardus Browse by Subject. http://www.renardus.org/

Renardus Home Page. http://www.renardus.org/.

Renardus Project Archive and Associated Research and Development, 2002.
http://www.renardus.org/about_us/project_archive.html

Resource Description Framework, http://www.w3.org/RDF/

Resource Description Framework, RDF. http://www.w3.org/RDF/

SCHEMAS Registry, http://www.schemas-forum.org/registry/

SchemaWeb. http://www.schemaweb.info/default.aspx

SEMKOS use cases, unpublished. http://www.lub.lu.se/SEMKOS/

Shadbolt N. et al. (2003). Advanced Knowledge Technologies, Interdisciplinary Research Collaboration, Mid-term Review, Scientific report, September 2003.
http://www.aktors.org/publications/Mid-Term%20Scientific%20Review.doc

SIMILE use cases: Semantic Interoperability of Metadata and Information in unLike Environments http://simile.mit.edu/documents/useCases/useCases.html (OAI repositories, institutional repositories)

Simple Object Access Protocol, SOAP 1.0. http://www.w3.org/TR/soap/

SKOS Simple Knowledge Organisation System. http://www.w3.org/2001/sw/Europe/reports/thes/

SKOS, SWAD-Europe Thesaurus Activity, http://www.w3.org/2001/sw/Europe/reports/thes/
Smith J.R. (2004), MPEG-21 REL: Enabling Interoperable Digital Rights Management, IEEE Multimedia, Oct-Dec 2004

Soergel, D. (1985). Organizing Information. San Diego, CA: Academic Press.

Soergel, D., Lauser, B., Liang, A., Fisseha, F. (March 2004).
Reengineering Thesauri for New Applications: The AGROVOC Example. In: Journal of Digital Information, Volume 4 Issue 4, Article No. 257, 2004-03-17.
http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Soergel/

St. Pierre M., LaPlant W. (1998) Issues in Crosswalking Content Metadata Standards (NISO whitepaper), http://www.niso.org/press/whitepapers/crsswalk.html

Stam, Deirdre C. (1990). The Quest for a Code, or a Brief History of the Computerized Cataloging of Art Objects. In: Petersen, Toni and Molholt, Pat (eds.). Beyond the Book : Extending MARC for Subject Access Boston: G.K. Hall, 1990. pp. 117-143.

Subramanian, S., and Shafer, K.E. "Clustering", (1998), (OCLC Publications),
http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003409

Suguri H., Kodama E., Mivazaki M. and Nunokawa H. (2001), Implementation of FIPA Ontology Service, Proceedings of the Workshop on Ontologies in Agent Systems, 5th International Conference on Autonomous Agents, 2001

SUO WG, the "Standard Upper Ontology Working Group" (IEEE P1600.1) web site. http://suo.ieee.org/

SWAD-E, Thesaurus links. http://www.w3.org/2001/sw/Europe/reports/thes/thes_links.html

Swad-E: 2003-11-13: Use Cases for a Thesaurus Service. http://www.w3.org/2001/sw/Europe/200311/thes/Use_cases_Thes_Service.html

SWAD-Europe Thesaurus Activity, http://www.w3.org/2001/sw/Europe/reports/thes/

SWAD-Europe Thesaurus Activity: SKOS API, http://www.w3.org/2001/sw/Europe/reports/thes/skosapi.html

SWWS Consortium (2003), *Report on Development of Web Service Discovery Framework. October 2003*, http://swws.semanticweb.org/public_doc/D3.1.pdf

Taxonomy Warehouse. http://www.taxonomywarehouse.com/

Technorati. http://www.technorati.com/tag

Tempich C., Staab S., Wranik A. (2004), REMINDIN': Semantic Query Routing in Peer-to-Peer Networks based on Social Metaphors, Proceedings WWW Conference 2004, May 2004, New York, USA

The ANSI/NISO Z39.50 Protocol. http://www.niso.org/z39.50/z3950.html

The BFO Ontology, http://ontology.buffalo.edu/bfo/

The Cyc web site. http://www.cyc.com/

The Digital Millennium Copyright Act, 1998. http://www.copyright.gov/legislation/dmca.pdf

The DOLCE Ontology. http://www.loa-cnr.it/DOLCE.html

The OCHRE Ontology. http://www.infomis.unileipzig.de/Research/pubs/forthcoming/ki2003epaper.pdf

The SDK WSMO working group, *Web Services Modeling Ontology*, http://www.wsmo.org/

Troncy R. (2003) *Integrating Structure and Semantics into Audio-visual Documents*, In Proc. of the 2nd International Semantic Web Conference, 2003

Tsinaraki C., Fatourou E., Christodoulakis S. (2003) *An Ontology-Driven Framework for the Management of Semantic Metadata describing Audiovisual Information*, in Proc. of CAiSE, Velden, Austria, 2003, pp 340-356

Tsinaraki C., Polydoros P., Christodoulakis S. (2004b) *Integration of OWL ontologies in MPEG-7 and TVAnytime compliant Semantic Indexing*, in Proceeding of CAiSE 2004

Tsinaraki C., Polydoros P., Christodoulakis S. (2004c) *Interoperability support for Ontology-based Video Retrieval Applications*, in Proceedings of CIVR 2004

Tsinaraki C., Polydoros P., Kazasis F., Christodoulakis S. (2004a) *Ontology-based Semantic Indexing for MPEG-7 and TV-Anytime Audiovisual Content*, Special issue of Multimedia Tools and Applications Journal on Video Segmentation for Semantic Annotation and Transcoding, 2004

Tudhope D., An overview of technological solutions to terminology services, http://www.ukoln.ac.uk/events/jisc-terminology/programme.html

Tudhope, D. and Binding, C. (2004). A Case Study of a Facetted Approach to Knowledge Organisation and Retrieval in the Cultural Heritage Sector. In: Resource Discovery Technologies for the Heritage Sector. DigiCULT Thematic Issue 6, June 2004. pp. 28-33. http://www.digicult.info/pages/pubpop.php?file=http://www.digicult.info/downloads/digicult_thematic_issue_6_lores.pdf

TV-Anytime Forum web-site. http://www.tv-anytime.org

UDDI Consortium (2000), *UDDI Specification*, http://www.uddi.org/specification.html

Uniform Resource Identifier. http://www.ietf.org/rfc/rfc2396.txt

Uniform Resource Name. http://www.ietf.org/rfc/rfc2141.txt

Visser, U., Stuckenschmidt, H., Wache, H. (2003). Ontology-Based Information Integration. Tutorial. IJCAI 2003, 18th International Joint Conference on Artificial Intelligence, Acapulco, Mexico. http://www.cs.vu.nl/~heiner/IJCAI-03/Tutorial-OntologyBasedIntegration-FINAL.ppt

Visser, U., Stuckenschmidt, H., Wache, H. (2004). Ontology-based Information Integration for the Semantic Web. Tutorial Handout. Lund, April 2004.

Visser, U., Stuckenschmidt, H., Wache, H. (2004). Ontology-based Information Integration for the Semantic Web. Tutorial Handout. Lund, April 2004.

Vizine-Goetz D. (2001). NKOS Registry - draft proposal for KOS-level metadata. http://staff.oclc.org/~vizine/NKOS/Thesaurus_Registry_version3_rev.htm

Vizine-Goetz D., Hickey C., Houghton A., and Thompson R. (2004). Vocabulary Mapping for Terminology Services. In: Journal of Digital Information, Volume 4 Issue 4. Article No. 272, March 2004, http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/, http://www.oclc.org/research/projects/termservices/default.htm

VocML. http://xml.coverpages.org/vocML.html

Volz R., Oberle D., Staab S. and Motik B. (2003), KAON Server -A Semantic Web Management System, Proceedings of the 12th WWW Conference, 2003

W3C (2003), *SOAP 1.2, W3C Recommendation*, http://www.w3.org/TR/soap12-part0/

W3C Recommendation 04 February 2004, http://www.w3.org/TR/REC-xml/

W3C Recommendation, 2004, http://www.w3.org/TR/owl-features/

Wallis, J., and Burden, P. (1995), "Towards a Classification-based Approach to Resource Discovery on the Web", (1995), http://www.scit.wlv.ac.uk/wwlib/position.html

Web Ontology Language, OWL. http://www.w3.org/2004/OWL/

Wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Main_Page

Wolverhampton Web Library. http://www.scit.wlv.ac.uk/wwlib/newclass.html

Wordnet web site, http://www.cogsci.princeton.edu/~wn/

XML Document Type Definition, Extensible Markup Language (XML) 1.0 (Third Edition), XrML. http://www.xrml.org/

Zeng, M.L., Chan, L.M. (2004). Trends and Issues in Establishing Interoperability Among Knowledge Organization Systems. In: Journal of the American Society for Information Science and Technology, 55(5): 377-395

Zisman, A., Chelsom, J., Dinsey, N., Katz, S. and Servan, F. (2002) "Using Web Services to Interoperate Data at the FAO". Proc. International Conference on Dublin Core and Metadata for e-Communities (Firenze UP), pp. 147-156

ZThes: A Z39.50 Profile for Thesaurus Navigation. http://zthes.z3950.org/profile/current.html

Acknowledgements

This work was funded by the DELOS -Network of Excellence on Digital Libraries (EU 6. FP IST, G038-507618).

As mentioned in the overview, we are very grateful for comments from those who participated in the forum discussions during the workshop associated with this deliverable [KESI Workshop, Sept. 2004].