ISSN 1747-1524

DCC | Digital Curation Manual

Instalment on "Ontologies"

http://www.dcc.ac.uk/resource/curation-manual/chapters/ontologies/

Martin Doerr
Information Systems Lab,
Institute of Computer Science,
Foundation for Research and Technology - Hellas (FORTH)
http://www.ics.forth.gr/isl/cci.html

January 2008
Version 1.0

Legal Notices

Catalogue Entry

Title	DCC Digital Curation Manual Instalment on Ontologies
Creator	Martin Doerr (author)
Subject	Information Technology; Science; Technology--Philosophy; Computer Science; Digital Preservation; Digital Records; Science and the Humanities.
Description	Instalment on the role of ontologies within the digital curation life-cycle. Describes the increasingly important role of ontologies for digital curation, some practical applications, the topic's place within the OAIS reference model, and advice on developing institution-specific selection frameworks.
Publisher	HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils.
Contributor	Seamus Ross (editor)
Contributor	Michael Day (editor)
Date	29/01/2008 (creation)
Type	Text
Format	Adobe Portable Document Format v.1.3
Resource Identifier	ISSN 1747-1524
Language	English
Rights	© HATII, University of Glasgow

Citation Guidelines

Martin Doerr, (January 2008), "Ontologies", *DCC Digital Curation Manual*, S.Ross, M.Day (eds), Retrieved <date>, from
http://www.dcc.ac.uk/resource/curation-manual/chapters/ontologies

About the DCC

The JISC-funded Digital Curation Centre (DCC) provides a focus on research into digital curation expertise and best practice for the storage, management and preservation of digital information to enable its use and re-use over time. The project represents a collaboration between the University of Edinburgh, the University of Glasgow through HATII, UKOLN at the University of Bath, and the Council of the Central Laboratory of the Research Councils (CCLRC). The DCC relies heavily on active participation and feedback from all stakeholder communities. For more information, please visit www.dcc.ac.uk. The DCC is not itself a data repository, nor does it attempt to impose policies and practices of one branch of scholarship upon another. Rather, based on insight from a vibrant research programme that addresses wider issues of data curation and long-term preservation, it will develop and offer programmes of outreach and practical services to assist those who face digital curation challenges. It also seeks to complement and contribute towards the efforts of related organisations, rather than duplicate services.

DCC - Digital Curation Manual

Editors

Peer Review Board

Preface

The Digital Curation Centre (DCC) develops and shares expertise in digital curation and makes accessible best practices in the creation, management, and preservation of digital information to enable its use and re-use over time. Among its key objectives is the development and maintenance of a world-class digital curation manual. The *DCC Digital Curation Manual* is a community-driven resource—from the selection of topics for inclusion through to peer review. The Manual is accessible from the DCC web site (http://www.dcc.ac.uk/resource/curation-manual).

Each of the sections of the *DCC Digital Curation Manual* has been designed for use in conjunction with *DCC Briefing Papers*. The briefing papers offer a high-level introduction to a specific topic; they are intended for use by senior managers. The *DCC Digital Curation Manual* instalments provide detailed and practical information aimed at digital curation practitioners. They are designed to assist data creators, curators and re-users to better understand and address the challenges they face and to fulfil the roles they play in creating, managing, and preserving digital information over time. Each instalment will place the topic on which it is focused in the context of digital curation by providing an introduction to the subject, case studies, and guidelines for best practice(s). A full list of areas that the curation manual aims to cover can be found at the DCC web site (http://www.dcc.ac.uk/resource/curation-manual/chapters). To ensure that this manual reflects new developments, discoveries, and emerging practices authors will have a chance to update their contributions annually. Initially, we anticipate that the manual will be composed of forty instalments, but as new topics emerge and older topics require more detailed coverage more might be added to the work.

To ensure that the Manual is of the highest quality, the DCC has assembled a peer review panel including a wide range of international experts in the field of digital curation to review each of its instalments and to identify newer areas that should be covered. The current membership of the Peer Review Panel is provided at the beginning of this document.

The DCC actively seeks suggestions for new topics and suggestions or feedback on completed Curation Manual instalments. Both may be sent to the editors of the *DCC Digital Curation Manual* at curation.manual@dcc.ac.uk.

Seamus Ross & Michael Day.
18 April 2005

Table of contents

Biography of the author

Martin Doerr studied mathematics and physics from 1972-1978 and holds a PhD in experimental physics from the University of Karlsruhe, Germany. Since 1990 he has been Principal Researcher at FORTH in Heraklion, Crete, Greece, and heads the Centre for Cultural Informatics, an activity of FORTH. He has led the development of knowledge base middleware and multilingual thesaurus management systems. Since 1992 he has been involved in developing a series of applied cultural information systems in Greece and elsewhere in Europe, within funded research projects and private contracts. Since 1996 he has collaborated with the Committee for Documentation of the International Council of Museums (ICOM-CIDOC). Under his initiative, CIDOC established ISO 21127 through intensive interdisciplinary collaboration between 1996 and 2005. ISO21127 is a core ontology for the integration of heterogeneous data structures in archives and museums of all kinds, and is currently being extended by the librarian concepts formulated by the International Federation of Libraries Associations (IFLA). Dr Doerr's research interests are semantic interoperability, information integration and ontology engineering.

Introduction and scope

Ontologies are formal models relating to how we perceive the possible states of affairs in our domain of discourse.[1, 2] For example, we may state: "A digital library consists of objects. Those can be texts, images, or multimedia objects. There may be links between those objects". An ontology would formulate such statements in a language based on logic[3] and define the intended real world meaning of the entities referred to, such as 'text', 'object', and of the relationships referred to, such as 'consists of', 'link'. Ontologies provide an effective means whereby human and electronic agents can communicate unambiguously about concepts by electronic means, which typically lack the facility for the user to ask for further clarification.[4]

In our context, ontology-mediated communication may take place with regard to the form and the states of the digital curation processes, the formal structure of documents and metadata and, most importantly, the subject or knowledge contained in the documents themselves. Even at the design phase, with the help of ontologies, system designers, digital curators and content experts can agree on their functional requirements, define data and metadata structures, and identify mechanisms for achieving interoperability and integration between heterogeneous digital library systems. Ontologies are particularly useful for making explicit the common underlying meaning of data structures encoded in different forms and database paradigms, such as XML and E-R.[5, 6]

Today the most prominent application of ontologies is the indexing of content for resource discovery. Depending on the circumstances, data creators, curators or tools can index content using concepts that a user may include in a query in order to access the relevant content. Ontologies are particularly useful for increasing the probability that the indexer and the user arrive at the same representation of the respective subject matter. An ontology known to both the indexer and the user can be considered as replacing the traditional verbal dialogue between librarians and library users. The logical rigour makes it easier to relate uncontrolled language to standardised concepts and to correlate indexing terms from different indexing systems, so that user language and indexing language can be matched.[7, 8, 9, 10] Interoperability between various ontologies will also become increasingly important in enabling members of disparate communities to reuse and understand digital information over time.

We envisage new ontology-based services that will leverage content in order to support research tasks through knowledge extraction, knowledge integration and subsequent reasoning (for example, Ref. 11). An overview of knowledge extraction for metadata generation in digital libraries can be found in Ref. 12. Finally, Web Services are becoming increasingly important for distributed processing. This is creating a new demand for ontologies to communicate the functions and respective states of data sets in automated, distributed processes in Digital Curation.[13, 14]

Ontologies can be regarded as the technological successors to thesauri. Whereas a traditional thesaurus aims to identify the meaning and relationships of a set of established expert terms, an ontology would do quite the opposite: it would aim at analysing our conceptual frames of mind[3, 15, 16] and try to encode them in a formal language in a kind of engineering process. In the end, a thesaurus with a logically well-defined structure such as SKOS Core[17] may well be regarded as an ontology, and thesauri and ontologies are frequently confused. It appears that both approaches are justified depending on the application. Ontologies and thesauri must also be seen as a particular case of Knowledge Organisation Systems (KOS), which comprise not only the description of possible states of affairs but also factual knowledge for unambiguous reference, such as gazetteers and person name authorities. Most of what will be said in this chapter also holds for KOS in general. It should, however, be clear that plain vocabularies are not ontologies, since words out of context have no definite meaning.[18]

An ontology may be as generic as Ranganathan's Fundamental Categories[19] or as specific and culture-dependent as the terms 'Samurai' or 'Potlatch'. It may comprise merely a dozen or up to millions of terms.[20] The construction is a costly intellectual process involving smaller or larger teams of experts in different areas. The cost of construction and maintenance, the necessary degree of elaboration, the breadth of application and the potential benefit should be carefully evaluated in decisions regarding the deployment of an ontology. This chapter tries to analyse characteristic use cases in Digital Curation and to give preliminary advice and further references for the effective creation and deployment of ontologies and other KOS in Digital Curation.

Background and developments to date

Ontology is originally a domain of philosophy. The clarification of the different meanings of words and the classification of what can be said about something go back to the very early philosophers. Ontology can be seen as the study of the existence of all the kinds of entities that make up the world.[3] In Artificial Intelligence or, better, Knowledge Representation (KR) as a discipline of Computer Science, it became apparent towards the end of the twentieth century that the challenges of automated reasoning not only lie in the development of logic but also in the non-trivial, explicit formulation of how the perceived world is actually composed. Consequently, the application of knowledge representation formalisms to ontological studies led to the so-called formal ontology, which now draws on logic, philosophy (mainly of the twentieth century) and cognitive science.

Around 1990, Computer Science undertook a series of large-scale experiments to integrate multiple, heterogeneous databases.[21, 22, 23] The experiments revealed that database integration must ultimately be based on explicit knowledge representation of the underlying common meaning rather than on formal manipulation of the data structures involved. With the work of Thomas Gruber[2] and others,[1, 3, 24] by 1998 the extraordinary importance of ontology for the design and operation of information systems was widely

recognised, and people started to see a series of formerly disparate fields in this new light – fields such as automated Natural Language translation and semantic networks, conceptual modelling and subject indexing in information science. Now, formal ontology has also become the term for the discrete product of an *ontology engineering* process, i.e. a particular set of concepts and their relationships declared in a logical form, with the new plural *ontologies*.

Apart from this scenario, information science recognised the need to refine library indexing systems or *indexing languages* under the pressure of the rapidly increasing volume of information world-wide. Between 1925 and 1965 Ranganathan developed the faceted classification system, a formal method for dynamically combining classification terms into new concepts.[19] Classification systems were supplemented with more flexible keyword systems or subject headings. The increasing physical distance of the cataloguer from the user required new measures to understand a user request. The first measure for obtaining a better match between keywords used by indexers and by users was vocabulary control, i.e. a reduction in the permitted synonyms. Since this does not reduce the inherent ambiguity of words, more elaborate knowledge organisation systems were developed. From about 1950 onwards, so-called thesauri[25, 26] were employed to associate terms with rich definitions ('scope notes'), synonyms and semantic relationships, which found their most characteristic expression in ISO2788.[27]

With the growing popularity of formal ontologies, it rapidly became clear that thesauri and ontologies have similar goals and use partially equivalent constructs. In particular, the basic structural element of a thesaurus to declare *generalisation hierarchies*, the *broader term generic* (BTG) defined in ISO2788, can be identified with the fundamental *IsA* relationship in KR. For example, 'bridge' BTG 'building' is equivalent to 'bridge' IsA 'building'. This is the key to merging the two methods. The specific contribution of thesauri is the detailed description of synonyms, the linguistic nature and cultural context of terms that allows for the relating of texts to concepts. However, ISO2788 still fails to make a clear distinction between terms and concepts ('senses', e.g. 'school', can be an institution and a building), whereas formal ontologies used to deal only with concepts. This incompatibility has been overcome in the schema of WordNet,[28] SKOS Core and many recent implementations. The specific contribution of formal ontologies is the detailed, logical definition of the possible semantic relations between entities, e.g. 'Person – *participates in* – Activity', 'Object – *was used in* – Activity' etc. In Digital Curation, all these features are needed, but not necessarily all together for the same use case. It will be explained later how these features relate to Digital Curation use cases.

The most prominent notation for ontologies is the *Terminological Logics* or *Description Logics* ($\mathcal{DL}$),[3, 29, 30, 31] a subset of First Order Logic specifically tailored to derive new concepts from more primitive ones in a generalisation hierarchy. For instance, in $\mathcal{DL}$ one may not only relate: 'Food – *is suitable for* – Living Being' and 'Food – *is made from*

– Thing', but by virtue of these relations (called 'roles' in $\mathcal{DL}$) one can classify Food into: 'Food *which is suitable for* Sharks' and 'Food *which is made from* Sharks'. In faceted classification systems, the term 'shark + food' would be ambiguous. This example demonstrates the utility of $\mathcal{DL}$ for classification tasks. The relations in ontologies, however, have two functions. First, they are categorical statements, such as 'a Shoemaker *makes* Shoes', describing how entities are generally related in our domain of discourse, equivalent to the *Related Term* (RT) in ISO2788. Secondly, they can be used as *conceptual schemata* to classify factual data, such as: 'John Smith *make*s my wedding shoes', where John Smith *is* a Shoemaker, and my wedding shoes *are* Shoes, and the described act is an example of '*makes*'".[32] Consequently, $\mathcal{DL}$ *systems* (e.g., CLASSIC, FaCT) support various kinds of automated reasoning such as automated classification of declared facts and control of the logical consistency of declarations.[33]

This function is distinct from that of thesauri, and it is the key to the use of ontologies for information integration at the schema level, as started in the early nineties (see above). A *manually* created ontology describes the intended meaning of the schema elements to be integrated, the shared underlying concepts and the relations between the respective concepts. From this description, an *automated* procedure is derived, which allows for transformation or merging of the data from the source systems[34] or transformation of global queries to queries understood by the source systems (so-called *mediation*[21]). The Digital Libraries community would talk about 'metadata crosswalks'. The necessary ontologies are small, very general in meaning and rich in relationships. For instance, ISO 21127 is a core ontology for schema integration in cultural heritage, containing 80 classes and 130 relationships.

After data are integrated under a common physical or virtual schema, the terminology in the data or in user requests may still not match between different systems. Typical terminological systems can contain between thousands and millions of terms. Information science has approached the problem by declaring different types of exact and inexact *equivalences* between terms (ISO5964),[7, 35, 36] which either guide the user from his terminology to that used in the target system or allow for automatic replacement of source terms by target terms.[4] ISO5964 could now be rewritten using Description Logic expressions. However, the cost of the manual labour necessary to create logical *thesaurus correlations* could often be prohibitive. Furthermore, the concepts behind terms may not be well defined and the actual associations between different classification terms may be *very vague* but nevertheless useful to increase *recall*. Therefore, there have been many attempts to automate thesaurus and vocabulary correlation. Fairly successful are the *statistical methods* from information retrieval, and neural network methods, which use training cases of parallel classification. *Machine learning* and *ontology learning* methods try to automatically recognise logical relations between terms from the context of use – a rapidly developing discipline but still more on the experimental side.

Computer Science research has been using a lot of similar knowledge representation languages, all suitable for ontologies. However, only recently has the *Resource Description Framework (RDF)*, a W3C recommendation, reached the status of standard of industrial relevance (http://www.w3.org/RDF). It allows for effectively defining concept hierarchies, conceptual schemata and semantic networks of factual knowledge. OWL, a successor to RDF, is now being promoted as the new standard for the formulation of ontologies (http://www.w3.org/2004/OWL). It combines the features of RDF with the concept derivation mechanisms of $\mathcal{DL}$. The old thesaurus standards have in recent years been undergoing revision in America and Europe. *SKOS Core*, one of the most promising initiatives for Digital Curation, replaces ISO2788 and ISO5964 with a clear-cut logic.[17] It allows for encoding multilingual and distributed terminological systems under an RDF schema. *Topic Maps*, on the other hand, seem to leave too much of the structural definitions up to the user. Finally, the TEI standards for printed thesauri do not play a practical role for Digital Curation. For other KOS, such as gazetteers and person lists, specific XML schemata may be the most effective encoding (see below).

How the topic applies to Digital Curation

The most important uses of ontologies in Digital Curation are:

- as an indexing language;
- in system level activities, such as schema design, configuration, integration.

Because the intended use affects the content, structure and methodology of ontology creation and maintenance, we shall deal with these two cases separately.

Ontologies as indexing language

Indexing or classification of digital objects can be seen as part of the metadata generation process. The cataloguer relates the object to certain categories based on its *content* or on its *context of creation and use* as access points or finding aids for the potential re-user of the object. Characteristically, these categories are assigned to metadata elements, such as 'subject', 'keywords', 'object type', 'creator role', 'format' etc. Alternatively, they may only be associated in a retrieval system index. These categories are the better the higher the probability is that the user will find one of the categories when he/she tries to solve a problem for which the object is relevant. Categories may be simple concepts such as 'wind' or complex expressions such as 'calculation of the stability of bridges under wind using finite elements'. They may include factual items, such as 'women emancipation in Goethe's work'. The relevance may be a piece of information from its content, the use of an image as illustration or decoration, or even a selection criterion for disposal. Furthermore, the category should relate to *all* relevant objects ('recall') and *only* to relevant objects ('precision'). From these three quality criteria, the most important requirements for ontologies are derived.

First, the number of categories and admissible syntax must be restricted in an indexing language. A very precise

expression that the user won't think of is of no use. Also, terms that are used only once in an index are not cost-effective. This is the reason for the great success of faceted classification. It reduces the vocabulary to a set of base terms and reduces complex relationships to one kind of relation. This method may find its limitations in ambiguous cases such as 'shark+food', and in cases where combinations seem to be far-fetched, such as 'grinding+factory' instead of 'mills'.[4] Not all terms that experts use are good indexing terms. Scientists and scholars used to develop detailed classifications to distinguish between possible contextual factors, for instance style and social status. Such terms are contestable, fuzzy and yield poor recall. Only stable, indisputable expert terms should be used.

Secondly, once the language is restricted, users must learn which of the terms they think of is in use. The user cannot be expected to be naturally aware of the expression in use or to continue searching until the correct term is selected. The rich response of search engines to simple keywords only conceals the poor level of recall due to the huge amount of available material. More targeted queries and exhaustive results are simply impossible. The problem of *learning* the indexing language is often a serious barrier for the user, though not so much for the cataloguer. The naïve idea that a user will effectively find a concept by navigating down a deep specialisation tree of terms does not work in practice. Therefore, the ontology should contain:

1. All reasonable **generalisations** of a concept. For instance, *carmine* is a *pigment*, a *dye*, an *animal product* and a *red colorant*.
2. All reasonable **contextual associations** of a concept. For instance, one may prefer to look for *bridge construction* via *bridge* rather than via *engineering* disciplines.
3. All reasonable **synonyms** and **lead-in terms** of a concept, even partial synonyms. For instance, *taxonomy* may be referred to as *classification, nomenclature, taxonomic system, categorisation, grouping, arrangement, organisation.*

Scarcely any current ontology fulfils all these criteria. Thesaurus editors have particular difficulties with requirement 3, because synonyms are not readily to hand. Natural Language processing techniques should be used to collect *candidate* synonyms and associations, which should be manually validated later. Rich synonyms allow for effective guidance of the user to a declared concept, but also for hiding the ontology and silently replacing non-preferred search terms with valid ones, as many search engines do today. Several user studies in the past claim that users do not like to use thesauri as search aids. We attribute these results rather to the inadequate organisation of the thesauri used and not to psychological barriers on the user side.[37] If an ontology is small enough to be displayed completely on one screen, users seem to be more attracted to using it.

Expressions

Concept derivation rules, such as faceted classification and $\mathcal{DL}$, allow for a dramatic reduction of the base vocabulary, which becomes easier to

learn, to maintain and to apply. However, this holds only if the combination rules are simple and obvious, and if there is only one expression that best fits an intended concept. When using $\mathcal{DL}$, the number of admissible *roles* should be minimal. There are some drawbacks, except when all acceptable combinations are explicitly listed ('*pre-coordinated*', antonym '*post-coordinated*'). The drawbacks are as follows:

- It is not easy to exclude nonsensical combinations such as 'sea + winter sports + resorts.'[38]
- There are still no browsers for the dynamic concept combinations on the market.
- Intelligent lead-in systems that would, for example, lead from *mills* to *grinding* + *factory* are also missing.

Digital curators and their consultants should make their choices carefully, taking into account the rapid technological progress.

Ontology correlation

If an indexing language is ideal and established for one system, it is still not ideal for access to distributed collections. The latter may span *different social or language groups* using other concept systems. Collections from a highly *specialised domain* use indexing concepts that are too specialised for a general collection. Since an indexing language must be continuously updated to follow the evolution of the world, every collection using it may employ a *different edition*. This leads to the requirement to correlate concepts from different groups, domains and versions. The correlations themselves are a huge intellectual investment. Therefore, an indexing language *must not* be developed in isolation. Any new concept *must be related* to the most similar concepts of some more widely used ontology and to the previous version. *Obsolete* concepts must continue to be accessible. Access systems must be able to exploit these correlations in order to mediate between different indexing languages. Only by observing these measures does semantic interoperability become scalable. The numerous architectures and methods of *thesaurus correlation* and *ontology mapping* are beyond the scope of this chapter (e.g. Refs 4, 39, 40, 41).

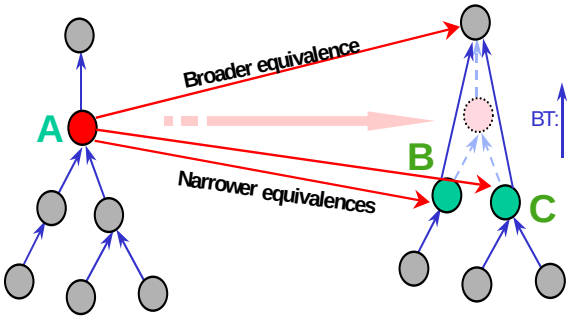

Figure1: Approximation of a missing concept in a target ontology by partial equivalences.

A factor that is often overlooked, and which complicates ontology correlation, is the intended context of use. Ideally, classification would mean that the classified object is an instance of the assigned class. This is not always what is meant. For instance, classification by subject can have different meanings.[42] '*Mars*' would characterise books dealing with Mars. '*Astronomy*' may characterise books containing knowledge produced by the scientific discipline astronomy. Making *Mars* a specialisation of astronomy is good practice in a subject catalogue but ontologically it is nonsense: a rigid body is not a kind of activity. Many

classification systems and thesauri cannot be correlated or merged because of such problems. The effect puzzles digital curators and researchers alike[43] and merits more research. As our need for global communication and information reuse increases, we can only recommend explicitly declaring the real relationship in such cases, for example Astronomy *studies* Mars or defining the concept in a fitting way: '*astronomy of Mars*'. Large thesaurus integration projects have adopted the approach of declaring more specific relationships than the one recommended in ISO2788.[20, 44, 45]

Indexing tools

Concepts of an indexing language can be used as data in metadata elements, such as subject, object type, format type, creator type, and other fields requiring categories. These data may be added manually by cataloguers or even by the creator of the object. Digital curators may need the assistance of domain experts to classify certain objects. Increasing the number of different persons involved increases the risk of different use of concepts. Digital curators should consider using tools that compare the content of a newly classified object with others of the same class. This can significantly improve consistency and completeness. Automatic indexing tools can analyse free text with regard to the occurrence of concepts from a thesaurus. Ontology editors should take this use into account. Such an ontology should also be rich in synonyms and in addition contain information about the linguistic characteristics of each term. Similarly, ontologies are being used increasingly for summarisation tools and data mining applications.

Indexing language creation and maintenance

It is possible to distinguish three different processes of manual ontology creation:

1. By knowledge engineering from an expert team in discussions and interviews.
2. By engineering from a vocabulary or other corpus.
3. As the result of submissions from individuals.

For an indexing language, process 2 seems to be the most effective, and process 3 for keeping the ontology up-to-date. Dagobert Soergel describes in his tutorials a well-established multi-step procedure for turning a vocabulary into an ontology. There is much literature dealing with methodologies for ontology creation (see, for example, Refs 45, 46, 47, 48). We recommend the reader to be critical, if the authors do in fact draw on a broad enough empirical base. In particular, the introduction of hierarchical and associative semantic relationships into a collection of concepts tends to reveal inconsistencies. Theoretical methodologies often ignore the fact that this leads to cyclic revisions of concepts and relationships that eventually converge.

Frequently, a new ontology is started by merging older ones. This may introduce unequal coverage of topics. A balanced coverage of topics is important and can be achieved by starting with a good corpus. Natural language processing has been quite successful with efficient term extraction tools that should be used to create initial vocabularies. Automatic thesaurus construction and recently ontology learning are still at a more

experimental stage, but we regard these methods as helpful for speeding up the creation process by proposing *candidate relations*. Furthermore, it appears that the upper level of an ontology, i.e. the more general concepts, is more fundamental for information integration.[1] It therefore seems reasonable to concentrate the expensive intellectual effort on the upper level, whereas automatic methods may be used more for the semantic organisation of lower-level terminology. Also, some of the better automatic methods need upper-level ontologies as input.

Ontology creation is extraordinarily labour intensive and error prone. For instance, the subsumption relationship (IsA) is frequently a case of confusion,[49] as in the example with Mars above. Ontology editors may easily get caught up in particular views motivated by the temporary context in which they are acting. Therefore ontologies should be developed collaboratively by teams comprising varied expertise and should be reviewed by several people. Library tradition assumes that there must be a small authoritative team of editors that controls the development in order to achieve the required intellectual quality. English Heritage and other organisations have developed good practice workflows for the processing of submissions to a growing ontology by an editorial team.

The authoritative model has the obvious drawback that it is not scalable and the normalisation of concepts always lags behind a rapidly developing world. Wikipedia and other Web fora show that an engaged, active public can also be a mechanism for enforcing quality. We recommend that the editors of indexing languages:

1. agree on common upper-level ontologies across domains and disciplines in order to guarantee interoperability at the fundamental and functional level. These ontologies can be small and stable;
2. develop specialised terminology in a highly distributed manner by collaborating with teams of varied expertise, where each team works on a smaller portion. The teams should not only develop the concepts but also their relations to the upper level and to *all* related concepts of other relevant teams;
3. elicit active feedback from a wide public to control the quality of concept definitions and to monitor the adherence of the participating teams to a common methodology. Automated reasoning tools will play an increasing role in detecting inconsistencies in large networks of correlated concepts, such as cyclic specialisation. Ontology learning tools may suggest revisions of manually created hierarchies from evidence of use.

The major challenge is to manage distribution, scalability and immediate updating. The high cost of ontology creation, and then of correlation with other ontologies, suggests intensive reuse. Therefore effective exchange formats should be used for individual portions of ontologies and not only for the whole. Each concept, and each related concept, must be globally identified because it may be used out of context. Exactly like species definitions in biology, the concept acquires identity through the definition (or ‘scope note’) and the context of creation (creator or edition of the ontology and date), and not just through its name. Respective

identification schemes have only recently become good practice, for example the Alexandria gazetteer ids, Universal Resource Identifiers for concepts in RDF etc.

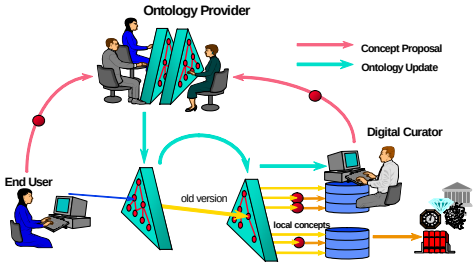

Figure 2: Ontology maintenance by concept submission.

Indexing language deployment

Indexing languages are shared resources. They should be accessible as Web Services for searching concepts by certain characteristics, browsing and downloading of portions of, or of complete, ontologies. Concepts must be accessible in previous and current versions to guarantee referential integrity with the information systems in use. An information system may enrich the indexing language locally with concepts before these or equivalent ones have been integrated into a new version. For this reason, the local system should maintain links to the closest concepts in the indexing language. Furthermore, an information system may maintain local copies of the concepts it actually uses, in order to provide users with only valid concepts for their search. The Web Service should maintain a specific information service regarding the changes in each new version, so that local copies of concepts can be efficiently updated. There should be tools to support the workflow of the updating process. If an indexing language is enriched locally without maintaining correlations with the new versions, updating at a later stage may become economically impossible. Digital curators should be aware of this.

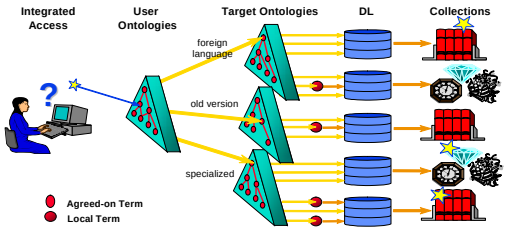

Figure 3: Integrated access by correlated ontologies.

The correlations with other indexing languages are either part of an indexing language itself or a resource in their own right. In the latter case, a Web Service may only describe correlations. Digital curators should agree on international collaboration to maintain a network of Web Services with correlated indexing languages. This implies common protocols, URIs for concepts, conventions and strategies to maintain referential integrity between those services.

Ontology presentation

As mentioned earlier, ontologies may be concealed from the user in indexing tools and tools transforming user requests into system-internal representations. However, ontology editors, cataloguers and users of search aid ontologies need effective presentation mechanisms to internalise the meaning of the declared concepts and relations. There are two competing requirements: The necessity to provide an overview, an orientation on a global map, and the avoidance of information overload. Whereas tree-like conceptual hierarchies can be reasonably presented

with indented lists or graphs, fully developed ontologies with many semantic relations easily become very messy.

Fortunately, indexing languages tend to be dominated by hierarchical relations. Nevertheless, many user interfaces for browsing ontologies do not show an adequate semantic environment of a concept. In this case, the user loses orientation and has difficulties in deciding if the selected concept is the correct one, the best one, or the only one that fits the search. Editors have difficulties recognising contradictory definitions in adjacent branches of a hierarchy etc. The issue must be regarded as an open research problem. Probably a multitude of presentation mechanisms and layout algorithms should be employed simultaneously, depending on user preferences and graph complexity. Graphical presentations ('conceptual maps') are gaining increasing popularity. There are even experiments with three-dimensional layout.

In their use as a search aid, and as a support for cataloguers, browsing and search facilities for concepts should be combined, i.e. a multitude of access methods is needed to ensure the quickest possible success rate, even from a poorly specified request. Since the indexing language is a reduced one, the users will in general not find 'what they are looking for', but only some more or less good match. This is frequently overlooked by system designers. A search facility should respond with available terms, choices of similar spellings, synonyms and other associations, and never leave the user with a blank screen or no answer.

Ontologies for schema integration

The second important application of ontologies in Digital Curation is the interoperability between heterogeneous metadata schemata.[50] We would like to make clear that a metadata schema, such as Dublin Core or METS, is neither a vocabulary nor an ontology, in spite of numerous references of this kind. A metadata element is a schema element, which expresses primarily a *relationship*, and secondarily a *constraint* on the referred *value*. For example, the element 'creator' means the described *object was created by an agent*, which is referred to by the value. If it were an element of a vocabulary, it would classify the referred value as 'creator' independent of context, an attribute making little sense at all. Nor does it qualify as an element of an ontology, since it mingles the type of the value with the name of the relationship. The difference is essential in order to understand the added value of an ontology, namely making the meaning underlying a metadata schema explicit.

An ontology for schema integration should define domain, relationship and value range explicitly, and declare the meaning of domain and range independently of the context as metadata of a digital object. The ontology may even analyse a relationship by more primitive elements. For instance, ISO21127[51, 52] recommends modelling events explicitly. In this formalism, the Dublin Core element DC:creator is explained as a path connecting three classes (entities) with two properties (relationships): Information Object – was created by – Creation Event – was carried out by – Actor.[53] The explicit declaration of the creation event allows for implementation of data integration

systems, which can consistently merge different information elements about one creation event from the same or different sources, such as the date, the place of creation (which is impossible in DC), the tools used, other participants and witnesses, the influences and motivations, even if the event is only implicit in the sources. There are at least three possibilities for using such *mappings* of metadata schemata to an ontology:

1. The data warehouse approach: Data from multiple sources are transformed into valid instances of the ontology, merged and then stored together in a suitable storage format.
2. The mediation approach: Queries are formulated in terms of the ontology, transformed into queries in terms of the source metadata, and only the query answers are integrated.[21]
3. The crosswalk approach: The mapping of two different metadata schemata to a common ontology is used to generate an algorithm for direct transformation of data from one schema to the other.

In all three cases, the mapping to a suitable ontology ensures the correct interpretation of the intended meaning of a metadata schema, and in the case of multiple transformations it avoids the need to declare separate transformations between all schemata, which frequently become inconsistent with each other. The application of these ontologies lies in the collaboration between digital curators and IT developers. They are common languages that domain experts and IT developers can understand and interpret, and which allow for formulating valid requirements for the configuration of information systems. They should be developed by digital curators and domain experts, perhaps taking advice from IT experts.

Ontology creation for schema integration

The methodology for creating ontologies for schema integration differs from that for an indexing language, as the focus is on the relationships a schema declares between entities. Normally, these entities are very generic. Their origin is not a vocabulary but the question for all items that may carry a relationship (or *property* in RDF) expressed in a schema. For instance, in Digital Curation the values of 'creator' may be people or organisations that have the potential to create the kinds of objects we are talking about, i.e. immaterial Information Objects. In other contexts, one may include robots, animals or natural processes. There may be no established terms in our language that group all these things, and we have to reinterpret terms like *Agent*, *Actor*, *Resource*, or *coin* terms like *Information Object* to describe the intended meaning.

Ontologies for schema integration should be developed by abstracting from a set of selected source schemata, by an interdisciplinary team that covers the intended range of application. These ontologies are generic, small ('*core ontologies*') and stable over time. Scientific disciplines and cultural differences play a minor role. The need for updating comes from the wish to extend the scope rather than from a changing world. They become a kind of *interlingua*. Therefore it is recommended to seek harmonisation with similar efforts from the outset, or to

agree on common standards such as ISO21127.

Frequent development mistakes are overgeneralisation and overspecialisation. Overgeneralisation, such as extending creation to natural processes, or regarding a book as equivalent to a theatre performance, can prevent a development team for a long time from understanding the real relationships of a domain and identifying the precise matches with other conceptualisations. The problems typically become apparent when the information integration is extended to new domains. Overspecialisation is a waste of development time and frequently an indication that a useful abstraction has not been recognised. For instance, one may argue that the notion of a *publisher* is overspecialised with respect to the intended scope of Dublin Core, and that it conceals the more generic case of multi-step production processes.

Core ontologies and indexing languages should fit. If the concepts of an indexing language appear in metadata fields that have been explained by a core ontology, then it must be possible to formulate the indexing language as an *extension* of the core ontology. The major reason for enforcing this compatibility is the integration of information in schemata with different levels of detail. For instance, schema A may declare one field 'object type', with values such as 'music CD'. Schema B may declare two fields: 'carrier', with value 'CD', and 'content type', with value 'music'. Only an integrated ontology for both schemata and values of the indexing language allows for translating data consistently in this seemingly trivial case. Few indexing languages used in digital curation have been investigated from this perspective.

Other kinds of use
Increasingly, ontologies will be used to define the functions, commands and parameters of the communication with Web Services.[13, 14] Furthermore, ontologies may play a role as parts of guides to good practice for any activity in Digital Curation, in particular for:

- appraisal and selection criteria;
- business models;
- certification;
- legal restrictions and obligations;
- workflows;
- institutional infrastructure models.

Some examples of these use cases may be found in the respective chapters of this manual.

Topic in action

There is an immense number of ongoing projects and initiatives that create or modify ontologies, deal with ontologies or claim to use ontologies. We shall present here some projects that are noteworthy for their approach, size or outcomes and constitute an interesting empirical base as advice for further development.

Thesaurus integration:
CENL, the Consortium of European National Libraries, has declared the strategic goal of correlating the general-purpose subject headings of all European languages (see CoBRA+, http://www.bl.uk/information/cobra.html) in order to support cross-language searching in European libraries, but leaving the autonomy of each indexing language untouched. The investment in

the existing indices of library content is so high that a change of indexing language is prohibitive. The pilot project called MACS[54, 55] aims to correlate the American LCSH, the French RAMEAU and German SWD. Worthy of note in this project is the commitment to validate manually all the links between hundreds of thousands of terms. Of course, the proposal of associations is tool assisted. The main argument is that the lifecycle of these vocabularies is longer than a hundred years. In such timeframes the investment in high quality pays off. On the other hand, CENL is in a particular situation. The vocabularies have a long history and stability due to the long-term standardisation efforts and collaboration in these organisations. Modern activism tends to define very short-term projects, which frequently reinvent similar things. Digital curators should be aware of the long potential lifetime of a good ontology and should consequently value stability and reuse more highly than the satisfaction of local tastes and preferences.

Current subject headings are almost 'flat' vocabularies, i.e. only small clusters of terms are related by hierarchical relationships. This does not satisfy the needs of a vast international public for modern, computer-aided information access. It would be a gigantic task to order half a million terms of LCSH hierarchically. On the other hand, library *classification systems* developed for 'shelving' books[42] by librarians have a deep and successful semantic structure. The aspects of classification must be rigorous to allow for a unique placement of each book in a subject space, therefore, by necessity other valid aspects of indexing have been neglected in favour of the unique ordering principle. Consequently, current library practice may use either subject headings or a classification system or both. OCLC has started a project to associate LCSH terms with Dewey (DCC)[56] classes in order to exploit both indexing methods in integrated access systems. The semantics of these associations can be fairly diverse. Telescopes, for example, can be associated with disciplines like optics, astronomy, geodesy, bird watching, warfare etc. The idea is to use a kind of probability measure for a subject heading to be associated with a Dewey class. Evidence for these probabilities is taken from co-occurrence of terms and classes in library indices. This probabilistic approach is characteristic of a series of projects for providing a quick solution to searching across deeply incompatible indexing systems, e.g. for cross-disciplinary searching.[41, 56, 57, 58]

The multilingual thesaurus attached to the European HEREIN project intends to offer a terminological standard for national policies dealing with architectural and archaeological heritage. Beyond just correlating concepts from different languages, as CENL does, the project decided to create for each language a new generalisation-specialisation hierarchy and to harmonise concepts manually. However, they did not preserve the concepts as found in other sources or link to them. We regard this as problematic, as an opportunity for interoperability seems to have been thrown away unnecessarily.[59]

Whereas the CENL subject headings from different countries are alternatives with basically the same thematic coverage, the situation is different if

complementary ontologies are to be integrated to make use of the combined potential. The gigantic integration of five million medical-pharmaceutical terms in UMLS to about one million concepts has clearly shown the necessity to explicitly declare dozens of different relationships that were mistaken for hierarchical relationships or undistinguished associations in the original thesauri, in order to achieve a more or less consistent overall intellectual structure. UMLS has defined about sixty relationships, which can be regarded as specialisations of the RT relationship in ISO2788. The experience of UMLS shows that thesauri produced by different methodologies can only be integrated into an intellectually coherent whole by creating an ontology (a 'Metathesaurus', as UMLS calls it). This Metathesaurus is produced by the automated alignment of machine-readable versions of some hundred source vocabularies, followed by human review and editing by subject experts. Obviously, manual methods would fail in such an endeavour. Furthermore, the innovation rate in this domain is so high that purely manual methods would most probably not be sufficient to keep up with it. Another interesting aspect is that for good reasons UMLS preserves the original structure of the source thesauri and just overlays the new semantic structure of the Metathesaurus.[20, 44] This *semiautonomous* status of both the overall ontology and the linked-in sources is exemplary for managing the collaborative aggregation of ontological resources. In other words, rather than creating yet another independent resource, all parties respect the referential integrity of the whole in their practice of updating the individual parts. This does not restrict the individual choices and requires little other than good will.

Similar to the approach taken by UMLS, the Food and Agriculture Organisation (FAO) of the United Nations is about to re-engineer its major thesaurus AGROVOC into an ontology with elaborate relationships between concepts. FAO maintains Digital Libraries in more than ten languages on the subject of agricultural knowledge, indexed by partially overlapping thesauri, such as the FAO Terminology, the Aquatic Sciences and Fisheries Abstracts (ASFA) Thesaurus, the Fisheries Glossary and the OneFish Glossary. The intention is to integrate these thesauri in the long term. Re-engineering of the thesaurus with the widest coverage into an ontology with elaborate relationships seems to be the only viable way to overcome the heterogeneity of interpretation of the classical thesaurus relationships BT, UF and RT in the different thesauri, such that an intellectually homogeneous access layer can be presented.[45]

Formats, Standards:

The computer science and KR community produces one ontology format after another, causing considerable confusion for the user. A reason for these variations is the fact that these formats come from research and are designed to demonstrate the utility of advanced features rather than become industrial standards. Another reason is the bias between the expressive power to make more and more sophisticated statements and the capability to run such systems in scalable environments. Few of the more advanced research formats can be implemented in such a way as to deal efficiently with a hundred thousand

concepts. The W3C Consortium is the most important initiative that is trying to produce viable standards for practical use, such as RDF and OWL. If digital curators are offered more advanced KR systems, they should ask for a demonstration of scalability according to the expected size of their ontologies. On the other hand, the pure thesaurus formats are lagging decades behind recent developments.

Recently, revisions of the 20-year old ISO2788 have been discussed, frequently ignoring all the needs of a networked environment. Significantly, SKOS Core[17] recognises the need to link any particular concept to concepts in other ontologies. It combines a modern interpretation of the semantic links of ISO2788 and ISO5694 with the extensibility features of RDFS to describe systems such as UMLS and AGROVOC. It contains practical recommendations on how to use concept identifiers in RDF-encoded metadata elements. SKOS Core is published and maintained by the W3C Semantic Web Best Practices and Deployment Working Group. It is still under development and another extension to add basic linguistic characteristics is under discussion, which would bridge the current gap between text processing and ontologies even better. We expect SKOS Core to become an important standard for encoding ontologies as indexing languages.

Factual KOS:

The most prominent applications of knowledge organisation systems for factual knowledge are the gazetteers and person name authorities. Of the gazetteers, the most advanced format seems to be that of the Alexandria Digital Library Gazetteer from the University of California. It not only provides an interactive Web interface but an exemplary, fully fledged Web Service protocol to search for possible place names by synonyms, wider geographic units and geographic coordinates. With more than five million place names, the gazetteer offers good general coverage of the world. Note that gazetteers should always represent coordinates as an area enclosing a place, however larger than the described place it may be, and not represent the centres (‘centroids’) only, so as to be able to narrow down search spaces by using coordinate boxes. Current shortcomings of the Alexandria project are that there is no wider collaboration as yet for systematic maintenance, no use of national alphabets (Unicode), and historical place names are missing.

Systematically maintained is the equally large GEOnet Names Server (GNS) of the National Geospatial-Intelligence Agency (NGA) and the US Board on Geographic Names.[60] The Thesaurus of Geographic Names by the Getty Research Institute lists more than a million place names in a less elaborate format, but it includes most of the known historical names of each geographic unit. EDINA is currently producing a comprehensive UK national gazetteer based on the Alexandria format and protocol. However, it doesn’t seem to be planning to grant free access so far, which will reduce its value for interoperability. Obviously, a larger application would want to use all three of these resources together in order to improve coverage, but none of the three seems to plan to correlate their concepts with the others. This leaves the job up to the user, to re-detect again and again

which places are identical between these resources. Finally, there is a plan for a comprehensive and distributed European gazetteer: "the Dutch- and German-speaking Division of the United Nations Group of Experts on Geographical Names (UNGEGN) initiated the project EuroGeoNames (EGN)**,** the vision of a distributed multi-lingual geographical names data network for Europe. EGN will be a distributed multi-lingual Internet service linking geographical names from official sources across Europe. Names searches in the EGN network will be possible for all official European languages including the officially recognized minority languages."[61]

Probably the best methodology for maintaining correlated KOS has been described by the European LEAF project in the case of person name authorities.[62] It implies OAI harvesting and semi-automatic concept consolidation with natural language processing techniques from different resources, and maintains full records of the source data and their relation to the consolidated records. Even though the consortium is about to fill the resource massively with data after the project has ended, the TEL (The European Library) project[63] is now adopting the LEAF methodology, and one can expect the emerging resource to produce a considerable impact. The only thing missing in LEAF is a Web Service interface. Even though museums, site and monuments records and archaeologists have been creating very large corpora and databases about objects, there is no general concept emerging of KOS in relation to physical objects. In parallel, IFLA is discussing a general model for authority files, FRAR.[64] 'Thesauri' relating to historical periods and events pose theoretical challenges due to the complex spatiotemporal and causal relationships, and are in the process of being discussed.[65]

The NKOS Forum is an informal international grouping that maintains a common mailing list and organises regular workshops on research issues of Networked Knowledge Organisation Systems in the framework of international conferences.

Core Ontology engineering:

There have been a few important ontology engineering activities for achieving interoperability of metadata structures for Digital Libraries. On the one hand, the <indecs> project[66, 67] was aimed at the massive integration of multimedia metadata for tracing intellectual property rights in the music industry. Being supported by experts on legal issues, they came up with an event-centric core ontology that was later developed into the ABC model in the ABC Harmony project,[68] an international collaboration funded by DSTC, JISC and NSF from 1999 until 2002 to investigate a number of the key issues in describing complex multimedia resources in digital form. It tested applications of the ABC model in general digital library projects and in more specific cultural heritage applications. Being very compact, it had a distinct theoretical impact on several research projects. In practical applications, such as an RDF metadata schema, the decision to model both events and the states between the events turned out to be rather unwieldy. ABC also integrated some fundamental concepts from FRBR, the Functional Requirements for Bibliographic Records,[69] an initiative of IFLA to

introduce the new notions of Work and Expression into cataloguing practice.

On the other hand, between 1996 and 2005 the International Committee for Documentation of the International Council of Museums (ICOM) developed the CIDOC Conceptual Reference Model (CRM), a core ontology aimed at the interoperability of object descriptions in all kinds of collections. This would become ISO21127 in early 2006. Both models, CIDOC CRM and ABC, are very similar, with CIDOC CRM being considerably more detailed. After a longer harmonisation effort by both sides, the CIDOC CRM now basically covers the ABC concepts.[70] Ongoing collaboration between IFLA and CIDOC aims to extend the CIDOC CRM to model completely the concepts of FRBR and its continuation, the FRAR,[64] following the CIDOC CRM methodology. A first version of this common model is expected in the summer of 2006. Currently, the CIDOC CRM covers nearly all of the meanings encoded in metadata schemata for Digital Libraries and can be regarded as the best core ontology for Digital Curation. Note, however, that it is not a metadata schema itself, but rather a language to describe the common meaning of different metadata schema elements.

Nicola Guarino and others developed in the WonderWeb project[71] a methodology to monitor whether generalisation-specialisation ('IsA') relations logically hold, based on elaborate distinctions between essential properties, linked to the existence of an entity, and accidental properties.[13] The method may be regarded as one of the most advanced theories of subsumption ('IsA') with practical applications, exemplary for reasoning in ontology engineering. It was applied to clean up the upper level of the WordNet ontology, one of the largest freely available ontologies of categories from common language.[49] The group also compiled what they call *foundational ontologies*[71] – ready-made, systematic logical elaborations of complementary and alternative notions of occurrence, substance and parthood among others. Even though these elaborations are quite useful for understanding and selecting ontological options for specific problems, the formulation in First Order Logic currently precludes their application by the majority of practitioners.

Theory and practice show that the interoperability of different ontologies is greatly improved if they share the same core concepts. Consequently, IEEE has engaged in a large-scale effort to standardise a set of core concepts for data interoperability, information search and retrieval, automated inferencing, and natural language processing in general, under the title Standard Upper Merged Ontology (SUMO).[72] It even integrates WordNet (http://www.ontologyportal.org). The initiative has yet to prove its practical utility. Already the WonderWeb project has shown the immense number of reasonable 'flavours' of foundational concepts. As we have stated above, automated communication needs a reduction in the possible concepts to those needed in a certain application context. This reduction can only come from an explicit relation of each selected concept to the intended functionality, as is the case for instance with the CIDOC CRM. In the case of SUMO, WordNet or the WonderWeb libraries, the user or

ontology engineer still has to select the concepts needed for the specific application and to justify their utility.

Furthermore, the necessary variation of a concept may still not be in the resource and needs to be constructed. Common-sense meaning, as captured by WordNet, refers to relevant distinctions in daily discourse. Classification and information integration may need quite different generalisations. It may be necessary to cover exceptions, non-typical or rare cases irrelevant to daily discourse. However, we highly recommend anyone engaging in ontology engineering tasks to study these resources in order to acquire an understanding of *ontological choices*. If a fitting concept can be found, it is always better to refer to it than to reinvent it.

Next steps

What you can do to promote effective use of ontologies depends on the tasks in which you are involved. We shall now present some specific suggestions. Let us assume:

a) You use your own indexing language and you are responsible for it.

- If you use a flat vocabulary, consider developing a (poly-) hierarchical semantic structure and enriching it with synonyms. If you employ different metadata elements for indexing different levels of genericity (such as 'classification, object category, object name'), control their consistency with the concept hierarchy.
- Relate your ontology to at least one more general and more widely used ontology. Adopt, if possible, the upper-level structure of a suitable, more general, ontology rather than developing your own.
- Set up a development team and take advice from well-tested development methodologies. Consider employing a consultant experienced in ontology development. A good domain expert is not automatically a good ontology engineer. Do not create 'your own ontology'.
- Consider assistance from natural language technologies or machine learning tools to enrich your vocabulary and to develop semantic hierarchies.
- Consider developing factual KOS, in particular in relation to persons and places (gazetteers).
- Consider the use of URIs wherever possible for concepts and ontologies. Follow standard rules and consider the use of registration services for your URIs (such as DOI, http://www.doi.org/).
- Do not try to hide your ontology or to *make money* out of it. Only a widely used ontology has value. Any obstacle to its take-up reduces your other returns from the investment.

b) You intend to employ a third-party indexing language.

- Check its quality according to the criteria referred to in this chapter and in the literature. Keep in mind, however, that it is more important to benefit from an already widely used and well-maintained ontology than to elaborate your own, even if

it should become more sophisticated than existing ones.

- **Do not** take the lack of some concepts that you need as the occasion to develop your own and better ontology. Rather, find a way to relate your local concept to the more standard indexing language, so that a request by a more general, standard concept will also return the objects indexed with your local concepts.
- Try to collaborate with the maintenance team of the selected ontology regarding your local requirements. Others may have the same needs as you and simply not yet organised collaboration.
- Install procedures to take part in all updates of the selected ontology and maintain referential integrity with your local concepts.

c) You plan to integrate legacy systems.

- If legacy systems use different or obsolete metadata schemata, consider mapping the semantics of all metadata schemata to a common core ontology. Begin with a suitable core ontology, for instance ISO21127, and extend it if necessary.
- Take into account the concepts of the upper levels of the indexing languages used in the metadata elements of your legacy systems and describe how they map to, or fit under, the core extended ontology (see above).
- Consider developing an integrated schema compatible with the extended core ontology in case you wish to integrate the legacy data physically without loss of meaning. Consider providing access by one or more standard metadata schemata by mapping them to the integrated schema.[73]
- Alternatively, the mapping of all involved schemata to the core ontology will help to define consistent crosswalks to the new schema to be used. This approach is only recommended for simple cases.
- Collaborate with your IT support team on the above. Make sure the data transformations according to the mappings you provide can be implemented and are effective. If necessary, advise your IT support team to acquire the respective know-how.
- If legacy systems employ different or obsolete indexing languages, you may consider reclassification of the data records. In many cases, this is not economically feasible. It is preferable to define cross-correlations between the concepts in the indexing languages that are evaluated at query time, similar to the local concepts described under b) above. Once cross-correlations are found, reclassification may be undertaken at any later time and in manageable steps.
- Consider assistance by automated tools to find candidate cross-correlations between the indexing languages.

d) You plan to provide integrated access to third-party digital repositories.

- Since the source systems are not under your control, mapping of metadata schemata and indexing languages is the only solution.

- You should choose a global schema, better defined by use of a core ontology as described above. Mappings are not implemented as data transformations, but as query transformations to the local schemata (so-called *schema mediation*).
- Indexing languages should be cross-correlated. The user may be offered one or more global indexing languages for information access, and the opportunity to try out local concepts.

e) You are using an indexing language as search aid.

- Bear in mind that any feedback about the quality of an indexing language is extremely valuable for the digital curator and will help the next user.

Future developments

We expect that the general knowledge about ontologies and their usefulness will become more widespread, with the effect that currently disparate methods such as controlled vocabularies, thesauri and ontologies will increasingly be integrated into resources combining sound logical definition with linguistic properties and other contextual knowledge. We do not expect the more advanced forms of logic discussed in research to acquire a wide practical importance in the near future, but the elaboration of different semantic relationships should soon become standard, both as a means for structuring ontologies and, increasingly, as resources in their own right for structuring information.

The development of ontologies of relationships or relationship-centric ontologies should have a strong impact on the creation of large KOS of factual knowledge[74] and on the formulation of virtually all engineering and management problems. We expect KOS of factual knowledge (places, people, events, periods, objects) to acquire a similar importance for indexing resources to that of ontologies at the present time.

We expect a rapid evolution and transfer of methods from natural language processing, such as term extraction, knowledge extraction, and summarisation, into the practice of ontology engineering, as well as from neural network and machine learning methods. But at least for the next few decades, we expect mainly semi-automatic methods to bring about real progress in the practice of Digital Curation, i.e. methods that combine automated reasoning with manual control and correction and that learn from human decisions.

Finally, the future of ontology creation, maintenance and use can only be seen in distributed collaborative work. The notion of ‘authority control’ by small expert teams will be replaced by a scientific and interested public collaborating in different roles. The challenges are two-fold:

1. To manage a system of open roles according to the capabilities of the partners so that they interact like the parts of an organism.
2. To organise the units of work and the relations between stages of development so that the curated knowledge converges in a more

consistent state, even though it is not controlled by a single mind.

Associated with these developments, we expect considerable shifts in the system of scholarly and scientific roles and the awarding of scholarly and scientific work.

Conclusions

Ontologies are formal models relating to how we perceive the possible states of affairs in our domain of discourse. They are engineering constructs that give a precise, logical account of the intended meaning of terms, data structure elements and other engineering models about the real world to which they refer or relate. They enable machines to process information provided by human agents in a manner that is consistent with the intended meaning. As such, they play an important role in mediating between users and information systems, but also between domain experts and IT experts, and in guiding and justifying system design decisions.

Information Science and Computer Science have come together in the development of indexing languages under the term ‘ontology’. The merging of methods and know-how from both sides is about to begin. Factual KOS, ontologies and linguistic resources are about to become facets of a continuous spectrum of resources. To fully exploit this potential, there is a need for detailed mutual understanding of the communities involved and there is still a lot of work to do. Practitioners in Digital Curation need to be systematically informed about ontologies and offered training in principles of ontology engineering. On the other hand, academic results from artificial intelligence may need to be more tailored to practice. Tools are essential to make ontology engineering and ontology use more efficient. Domain experts should overcome their reservations about (semi)automatic methods and help IT engineers to integrate such tools efficiently into their workflow.

It is difficult to create an ontology, and it is even more difficult and time consuming to create a proven one. A domain expert is not necessarily a good ontology engineer and vice versa, and a single individual is unlikely to have all the relevant concepts and exceptions readily at hand to define a good ontology by his- or herself. Ontology creation results from systematic collaboration between experts of different disciplines. Even if someone defines a perfect ontology, it is of very limited use if it does not relate to the ontologies of other groups. Only if we learn to overcome the current social and technical isolation of expert groups and their ontologies and form wide collaboration networks will we have an opportunity to realise the expectations of ontologies as a tool for precise, global information access. On the other hand, there is no real alternative to ontologies in order to make information precise, and a proven ontology has a long period of validity. Therefore, a well-planned investment in ontologies will always pay off.

References

[1] Guarino, N., "Formal Ontology and Information Systems", in: N. Guarino (ed.), Formal Ontology in Information Systems, Proceedings of the 1st International Conference, Trento, Italy, 6-8 June 1998, IOS Press.

[2] Thomas R. Gruber, "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", in: Formal Ontology in Conceptual Analysis and Knowledge Representation, edited by Nicola Guarino and Roberto Poli, Kluwer, 1994.

[3] Sowa, John F., "Knowledge Representation: logical, philosophical, and computational foundations", Brooks/Cole, Pacific Grove, 2000, ISBN 0 534-94965-7.

[4] Doerr, M., "Semantic Problems of Thesaurus Mapping," *Journal of Digital Information*, Special Issue on Networked Knowledge Organization Systems, Vol. 1, Issue 8, April 2001.

[5] Ron Weber, "Conceptual Modelling and Ontology: Possibilities and Pitfalls", in: Lecture Notes In Computer Science; Vol. 2503, Proceedings of the 21st International Conference on Conceptual Modeling, 2002, pp.1-2, ISBN 3-540-44277-4.

[6] Nicola Guarino and Luc Schneider, "Ontology-Driven Conceptual Modelling," in: Lecture Notes In Computer Science, Vol. 2503, Proceedings of the 21st International Conference on Conceptual Modeling, 2002, p. 10, ISBN 3-540-44277-4.
Soergel, D., "Organizing Information", San Diego, CA: Academic Press, 1985.

[7] M. Doerr and I. Fundulaki, "SIS – TMS: A Thesaurus Management System for Distributed Digital Collections," Proceedings of the 2nd European Conference, ECDL'98, September 1998, Heraklion, Crete, Greece.

[8] M. Doerr, "Effective Terminology Support for Distributed Digital Collections", in: Sixth DELOS Workshop, Preservation of Digital Information, Tomar, Portugal, 17-19 June 1998, ISBN 2-912335-06.

[9] "Indexing languages and thesauri: construction and maintenance", Dagobert Soergel, Los Angeles, CA: Melville, 1974. 632 pp., 72 fig., ca 850 ref. (Wiley Information Science Series).

[10] D. Soergel, "SemWeb: Proposal for an open, multifunctional, multilingual system for integrated access to knowledge about concepts and terminology," *Advances in Knowledge Organization*, 5, pp. 165-173, 1996.

[11] Sanjeev Thacker, Amit P. Sheth and Shuchi Patel, "Complex Relationships for the Semantic Web, Spinning the Semantic Web", pp. 279-315, 2003.

[12] Les Carr, Tim Brody, Nicholas Gibbins, Liz Lyon, Ann Chapman and Michael Day, "Study to determine the requirements for and usage of extracted knowledge", DELOS Network of Excellence on Digital Libraries – deliverable 5.1.2, February 2005, http://delos-wp5.ukoln.ac.uk/project-outcomes/southampton/final-delos-task-5_1_2.doc

[13] Dieter Fensel and Christoph Bussler, "The Web Service Modeling Framework WSMF", *Electronic Commerce Research and Applications*, 1(2), pp. 113-137, 2002, Gazetteers.

[14] Asunción Gómez-Pérez, Rafael González-Cabero and Manuel Lama, "Development of Semantic Web Services at the Knowledge Level," ECOWS 2004 http://www.informatik.uni-trier.de/~ley/db/conf/ecows/ecows2004.html#Gomez-PerezGL04, pp. 72-86.

[15] M. Minsky, "A Framework for Representing Knowledge. The Psychology of Computer Vision", P.H. Winston (ed.), McGraw-Hill, 1975.

[16] Gilles Fauconnier and Mark Turner, "The Way we Think: Conceptual Blending and the Mind's Complexities", Basic Books, New York, 2002.

[17] Miles, A., Matthews, B., Wilson, M. and Brickley, D., "SKOS Core: Simple Knowledge Organisation for the Web," International Conference on Dublin Core and Metadata Applications: Vocabularies in Practice, 2005. Available at http://purl.org/dcpapers/2005/Paper01.

[18] Ludwig Wittgenstein, „Philosophische Untersuchungen", Werkausgabe, Band 1, Suhrkamp Taschenbuch, Frankfurt, 1984, ISBN 3-518-28101-1.

[19] S.R. Ranganathan, "A descriptive account of Colon Classification", Bangalore: Sarada Ranganathan Endowment for Library Science, 1965.

[20] US National Library of Medicine, National Institutes of Health (2004), "Fact Sheet, UMLS Semantic Network", 19 February 1998, http://www.nlm.nih.gov/research/umls/about_umls.html Last updated 19 July 2004.

[21] Gio Wiederhold, "Mediators in the Architecture of Future Information Systems", in: *IEEE Computer*, March 1992.

[22] R.J. Bayardo *et al.*, "InfoSleuth: Agent-Based Semantic Integration of Information in Open and Dynamic Environments", MCC Technical Report, MCC-INSL-088-96, October 1996.

[23] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman and J. Widom, "The TSIMMIS Project: Integration of Heterogeneous

Information Sources", in: Proceedings of the IPSI Conference, pp. 7-18, Tokyo, Japan, October 1994.

[24] Uschold, M. and Gruninger, M., "Ontologies: principles, methods and applications," *The Knowledge Engineering Review*, 11(2), pp. 93-136, 1996.

[25] Vickery, B.C., "Thesaurus – A new word in documentation", *Journal of Documentation*, 16(4), pp. 181-89, 1960.

[26] D.J. Foskett, "Thesaurus", in: Readings in Information Retrieval, eds K. Sparck Jones and P. Willet, Morgan Kaufmann, 1997.

[27] ISO 2788-1986: Documentation – Guidelines for the establishment and development of monolingual thesauri, International Organization for Standardization, Ref. No. ISO 2788-1986, 1986.

[28] Miller, A. George, Beckwith, Richard, Fellbaum, Christiane, Gross, Derek and Miller, Katherine, "Introduction to WordNet: An On-Line Lexical Database", 1993.

[29] F. Baader, H.-J. Burckert, J. Heinsohn, B. Hollunder, J. Muller, B. Nebel, W. Nutt and H. Profitlich, "Terminological knowledge representation: a proposal for a terminological logic", DFKI Report, DFKI, Saarbrucken (1992).

[30] A. Borgida, "Description logics in data management," *IEEE Transactions on Knowledge and Data Engineering*, 7(5), pp. 671-682, 1995.

[31] More information can be found at:
http://www.ida.liu.se/labs/iislab/people/patla/DL/index.html

[32] Bechhofer, S. and GobleE, C.A., "Classification Based Navigation and Retrieval for Picture Archives", IFIP WG2.6 Conference on Data Semantics, DS8, Rotorua, New Zealand, 1999.

[33] S.K. Bechhofer, C.A. Goble, A.L. Rector, W.D. Solomon and W.A. Nowlan, "Terminologies and Terminology Servers for Information Environments".

[34] Diego Calvanese, Giuseppe De Giacomo, Maurizio Lenzerini, Daniele Nardi and Riccardo Rosati, "Description logic framework for information integration", in: Proceedings of the 6th International Conference on the Principles of Knowledge Representation and Reasoning, KR'98, pp. 2-13, 1998.

[35] ISO5964-1985: Documentation – Guidelines for the establishment and development of multilingual thesauri, International Organization for Standardization, Ref. No. ISO 5964-1985, 1985.

[36] "Guidelines for Forming Language Equivalents: A Model Based on the Art & Architecture Thesaurus", International Terminology Working Group, Getty Information Institute, 1996 (for copies, contact Murtha Baca, mbaca@getty.edu).

[37] Srinivasan, Ramesh and Huang, Jeffrey, "Fluid Ontologies for Digital Museums", *Journal of Digital Libraries*, special issue on Digital Museums, May 2005.

[38] Yannis Tzitzikas, Anastasia Analyti and Nicolas Spyratos, "The Semantics of the Compound Term Composition Algebra", Proceedings of the 2nd International Conference on Ontologies, Databases and Applications of Semantics, ODBASE'2003, Catania, Sicily, Italy, November 2003.

[39] Ralf Kramer, Ralf Nikolai and Corinna Habeck, "Thesaurus federations: loosely integrated thesauri for document retrieval in networks based on Internet technologies," *International Journal on Digital Libraries* (ISSN: 1432-5012), Vol. 1, No. 2, pp. 122-131, September 1997.

[40] Ralf Nikolai, Ralf Kramer, Marc Steinhaus, Bruno Felluga and Paolo Plini, "GenThes: A General Thesaurus Browser for Web-based Catalogue Systems," in: Proceedings of the Third IEEE Meta-Data Conference, 6-7 April 1999, Bethesda, Maryland.

[41] Y. Kalfoglou and M. Schorlemmer, "Ontology mapping: the state of the art", *The Knowledge Engineering Review*, 18(1), pp. 1-31, January 2003.

[42] Chris Welty and Jessica Jenkins, "Formal Ontology for Subject", *Journal of Knowledge and Data Engineering*, 31(2), pp. 155-182, September 1999, Elsevier.

[43] E. Schulten, *et al*., "The E-Commerce Product Classification Challenge," *IEEE Intelligent Systems*, Vol. 16, No. 4, pp. 86-89, 2001.

[44] US National Library of Medicine, "2001 UMLS Metathesaurus", 12 January 2001, Section 2, http://www.nlm.nih.gov/research/umls/META2.HTML

[45] Dagobert Soergel, Boris Lauser, Anita Liang, Frehiwot Fisseha, Johannes Keizer and Stephen Katz, "Reengineering Thesauri for New applications: the AGROVOC Example", *Journal of Digital Information*, Vol. 4, Issue 4, Article No. 257, 17 March 2004.

[46] Cuel R. and Cristani M., "Ontology Methodologies: a survey", *International Journal on Semantic Web and Information Systems*, pp. 49-69, 2005.

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

[48] Óscar Corcho, Mariano Fernández-López and Asunción Gómez-Pérez, "Methodologies, tools and languages for building ontologies: Where is their meeting point?", Data Knowl. Eng. 46(1), pp. 41-64, 2003.

[49] Gangemi, Aldo, Guarino, Nicola, Masolo, Claudio and Oltramari, Alessandro, "Sweetening WordNet with DOLCE", *AI Magazine*, 24(3), pp. 13-24, 2003.

[50] Manjula Patel, Traugott Koch, Martin Doerr, Chrisa Tsinaraki, Nektarios Gioldasis, Koraljka Golub and Doug Tudhope, "Semantic Interoperability in Digital Library Systems", DELOS Network of Excellence on Digital Libraries – deliverable 5.3.1, June 2005.

[51] Doerr, M., "The CIDOC CRM – An Ontological Approach to Semantic Interoperability of Metadata", *AI Magazine*, 4(1), 2003.

[52] ISO/FDIS 21127: Information and documentation – A reference ontology for the interchange of cultural heritage information, 17 October 2005.

[53] Martin Doerr, "Mapping of the Dublin Core Metadata Element Set to the CIDOC CRM", Technical Report FORTH-ICS/TR-274, July 2000.

[54] Patrice Landry, "The MACS Project: Multilingual Access to Subjects (LCSH, RAMEAU, SWD)", 66th IFLA Council and General Conference, 13-18 August 2000.

[55] Martin Kunz, "Subject retrieval in distributed resources: a short review of recent developments", 68th IFLA Council and General Conference, 18-24 August 2002.

[56] Diane Vizine-Goetz, Carol Hickey, Andrew Houghton and Roger Thompson, "Vocabulary Mapping for Terminology Services", Journal of Digital Information, Volume 4 Issue 4, Article No. 272, 11 March 2004. http://journals.tdl.org/jodi/rt/printerFriendly/jodi-128/113

[57] Hsinchun Chen, J. Martinez, T.D. Ng and B.R. Schatz, "A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System", *Journal of the American Society for Information Science*, Vol. 47, No 8, August 1996, http://ai.bpa.arizona.edu/papers/wcs96/wcs96.html

[58] S. Amba, N. Narasimhamurthi, K.C. O'Kane and P.M.Turner, *Automatic linking of thesauri*, in: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996, pp. 181-187, Konstanz: Hartung-Gorre.

[59] Thesaurus: European Heritage Network, http://www.european-heritage.net/sdx/herein/thesaurus/introduction.xsp

[60] GEOnet Names Server (GNS) of the National Geospatial-Intelligence Agency (NGA) and the US Board on Geographic Names, http://earth-info.nga.mil/gns/html/index.html

[61] The EuroGeoNames (EGN) project, http://www.eurogeographics.org/eng/03_projects_EuroGeoNames.asp

[62] Max Kaiser, Hans-Jörg Lieder, Kurt Majcen and Heribert Vallant, "New Ways of Sharing and Using Authority Information: The LEAF Project", *D-Lib Magazine*, 9(11), 2003. http://www.dlib.org/dlib/november03/lieder/11lieder.html

[63] "The European Library Fact Sheet for Publishers", February 2003, http://www.europeanlibrary.org/pdf/fact_sheet%20for%20publishers2.pdf

[64] Glenn E. Patton, "FRAR: Extending FRBR Concepts to Authority Data", World Library and Information Congress: 71th IFLA General Conference and Council, August 2005, Oslo, Norway, http://www.ifla.org/IV/ifla71/papers/014e-Patton.pdf

[65] Martin Doerr, Athina Kritsotaki and Stephen Stead, *Which Period is it? A Methodology to Create Thesauri of Historical Period*, Computer Applications and Quantitative Methods in Archaeology Conference, CAA 2004, 13-17 April 2004, Prato, Italy. Available at: http://www.ics.forth.gr/isl/publications/paperlink/caa2004_period.pdf

[66] INDECS Home Page: Interoperability of Data in E-Commerce Systems, http://www.indecs.org

[67] Godfrey Rust and Mark Bide, "The <indecs> metadata framework Principles, model and data dictionary", WP1a-006-2.0, June 2000, http://www.indecs.org/pdf/framework.pdf

[68] Carl Lagoze and Jane Hunter, "The ABC Ontology and Model", DC-2001, International Conference on Dublin Core and Metadata, Tokyo, October 2001, http://metadata.net/harmony/lagoze_hunter_dc2001.pdf

[69] "Functional Requirements for Bibliographic Records," International Federation of Library Associations and Institutions, March 1998, http://www.ifla.org/VII/s13/frbr/frbr.pdf

[70] Martin Doerr, J. Hunter and Carl Lagoze, "Towards a Core Ontology for Information Integration", in: *Journal of Digital information*, Vol. 4, Issue 1, April 2003.

[71] Masolo, Claudio, Borgo, Stefano, Gangemi, Aldo, Guarino, Nicola and Oltramari, Alessandro, The WonderWeb Library of Foundational Ontologies and the DOLCE ontology (WonderWeb (EU IST Project 2001-33052), Deliverable D18, 2001, http://wonderweb.semanticweb.org/deliverables/documents/D18.pdf

[72] Pease, A. and Niles, I., "IEEE Standard Upper Ontology: A Progress Report", *Knowledge Engineering Review*, Special Issue on Ontologies and Agents, 17, pp. 65-70, 2002.

[73] Amato G., Gennaro, C., Rabitti, F. and Savino, P., "Milos: A Multimedia Content Management System for Digital Library Applications", ECDL 2004, pp. 14-25. http://www.informatik.uni-trier.de/~ley/db/conf/ercimdl/ecdl2004.html#AmatoGRS04

[74] Amit P. Sheth, "From Semantic Search to Analytics and Discovery on Heterogeneous Content: Changing Focus from Documents and Entities to Relationships", SWDB 2003, p. 7.

Terminology

Controlled Vocabulary

A controlled vocabulary is a list of terms whose meanings are specifically defined by organised editorial control. The purpose is to improve technical communication by ensuring that everyone is using the same word to mean the same thing. This consistency of terms is one of the most important concepts in technical writing, where effort is expended to use the same word throughout a document instead of slightly different ones to refer to the same thing. (Adapted from: en.wikipedia.org/wiki/Controlled_vocabulary)

Indexing Language

An artificial language consisting of a formal vocabulary selected to facilitate information retrieval by serving as access points in a catalogue or index, including any lead-in vocabulary and rules governing form of entry and syntax. (Adapted from: ODLIS – Online Dictionary for Library and Information Science, by Joan M. Reitz, http://lu.com/odlis/odlis_i.cfm)

Ontology

An ontology is *a logical theory* accounting for the intended meaning of a formal vocabulary, i.e. its ontological commitment to a particular conceptualisation 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 conceptualisation) by approximating these intended models.[1]

Schema Mediation

Schema mediation is a technique for issuing queries and receiving answers simultaneously from multiple heterogeneous databases. Based on the definition of a schema mapping from a global schema to each source schema, queries are transformed to fit each source schema, issued to the source databases, and then each answer set is transformed back into the format of the global schema. The answer sets from the individual databases may be integrated or not.

An annotated list of key external resources

IFLA, International Federation of Library Association and Institutions, maintains a Web site with information from their recent discussions and developments with regard to indexing and other topics.
http://www.ifla.org/

The NKOS Forum is an informal international grouping that maintains a common mailing list and organises regular workshops on research issues relating to Networked Knowledge Organisation Systems in the framework of international conferences.
http://nkos.slis.kent.edu/

The Laboratory of Applied Ontologies in Trento is one of the leading institutions dealing with issues of ontology engineering.
http://www.loa-cnr.it/

The CIDOC CRM Special Interest Group maintains the home page of ISO21127, with links to technical papers and applications.
http://cidoc.ics.forth.gr/

The Getty Research Institute has developed important knowledge organisation systems for cultural heritage, the Art & Architecture Thesaurus (AAT), the Thesaurus of Geographic Names (TGN) and the Union List of Artist Names (ULAN).
http://www.getty.edu/research/institute/

Leonard Will maintains a rich portal with information about thesauri.
http://www.willpowerinfo.co.uk/

The Alexandria Digital Libraries Project of the University of California has created exemplary applications of knowledge organisation systems.
http://www.alexandria.ucsb.edu/

The GEOnet Names Server (GNS) provides access to the National Geospatial-Intelligence Agency (NGA) and the US Board on Geographic Names.
http://earth-info.nga.mil/gns/html/index.html

There are several interesting conference series on the topic, such as:

ER2006, 25th International Conference on Conceptual Modeling
http://adrg.eller.arizona.edu/ER2006/

ISKO2006, 9th International Conference of the International Society for Knowledge Organization
http://isko.univie.ac.at/papers/openconf.php

ISWC2006, 5th International Semantic Web Conference
http://iswc2006.semanticweb.org/

ESWC2006, 3rd European Semantic Web Conference
http://www.eswc2006.org/

ICDL2004, International Conference on Digital Libraries
http://www.teriin.org/events/icdl/

ECDL2006, European Conference on Research and Advanced Technology for Digital Libraries
http://www.ecdl2006.org/

NKOS Workshops on ECDL and ICDL
http://nkos.slis.kent.edu/

More theoretically oriented is FOIS 2004, International Conference on Formal Ontology in Information Systems
http://fois2004.di.unito.it/