Draft
Final version to appear in
Sebastian Ryszard Kruk and Bill McDaniel, eds.
Semantic Digital Libraries
Springer 2009

Digital Libraries and Knowledge Organization

Dagobert Soergel[1]

College of Information Studies, University of Maryland, College Park, MD
dsoergel@umd.edu

This chapter describes not so much what digital libraries are but what digital libraries with semantic support could and should be. It discusses the nature of Knowledge Organization Systems (KOS) and how KOS can support digital library users. It projects a vision for designers to make and for users to demand better digital libraries.

What is a *digital library*? The term "Digital Library" (DL) is used to refer to a range of systems, from digital object and metadata repositories, reference-linking systems, archives, and content management systems to complex systems that integrate advanced digital library services and support for research and practice communities. A DL may offer many technology-enabled functions and services that support users, both as information producers and as information users. Many of these functions appear in information systems that would not normally be considered digital libraries, making boundaries even more blurry. Instead of pursuing the hopeless quest of coming up with **the** definition of *digital library*, we present a framework that allows a clear and somewhat standardized description of any information system so that users can select the system(s) that best meet their requirements. Section 2 gives a broad outline; for more detail see the *DELOS DL Reference Model* [**?**].

1 A vision of digital libraries

At its best, a digital library

- integrates access to materials with access to tools for processing materials **(DL = materials + tools)**;
- supports individual and community information spaces through functionality for selection, annotation, authoring / contribution, and collaboration.

The remainder of this section elaborates on this vision, starting with a use case that illustrates advanced DL functionality.

1.1 Use case illustrating advanced DL functionality

Table 1. Use case/scenario. Writing a proposal for a drug prevention program

The director of a drug-free community coalition works on developing a prevention project and the funding for it. Signing on to the AOD (Alcohol and Other Drugs) DL, she begins by browsing the prevention section of the thesaurus hierarchy to get a structured overview of various prevention approaches (see Table 2). From the thesaurus scope notes, some approaches seem particularly applicable to her community, so she follows the links to more in-depth explanations. Back from the thesaurus she follows a link from JG10.4.6 *prevention through education* to a funding program announcement. She opens the guidelines for submitting proposals and copies a proposal template into her private space (shown in another window) and fills in some text and copies some text (which is transferred with its source). From the program announcement, she follows a link to projects funded previously and further to project reports and evaluations. She comes across the unfamiliar term *triangulation* and clicks on it to see the thesaurus entry, which gives an explanation and the hierarchical context. In another document she highlights the phrase *prevention program evaluation* to initiate a cross-database search in her own system and three external databases. She copies four references with abstracts to her private space and adds some (public) semantic tags that capture what these documents can contribute to the proposal. (Later she will return to these, select one for detailed reading (using a reading tool that lets her quickly indentify important paragraphs), and add more notes and quotes to her emerging proposal.) From the program announcement, she follows a link to relevant research, selects some articles, and tags them with elements of her proposal outline. One of the papers compares the effectiveness of several prevention curricula. She follows a link to the top-rated curriculum and from there finds further reviews and some discussion of resources required. She still needs demographics of her community; she uses a tool to query census data and produce a table with the data she needs. She also needs funding sources for the required local match, so she searches two external databases and incorporates the results into her proposal. Now she completes the first draft with annotations as to why a piece is included or why certain language is used. She emails two board members and a city staff member for comments, giving them access to her private space. Upon receiving their revisions (with changes tracked) and comments, she produces the final version and uploads a pdf to the agency's submission system.

Table 2. Prevention approach hierarchy. Excerpt, only selected annotations

JG	**prevention approach**
JG10	. **individual-level prevention**
JG10.2	. . individual- vs family-focused prevention
JG10.2.2	. . . individual-focused prevention
JG10.2.4	. . . family-focused prevention
JG10.4	. . prevention through information and education
	SN Information and education directed at individuals to influence their knowledge, beliefs, and attitudes towards AOD and their AOD use behavior.
	ST *prevention through persuasion*
	NT +JH2 health information and education
	BT +N communication, information, and education
	RT MP18.2.8.8 demand reduction policy
	+ND16.8 information event
	+T demographic characteristics
JG10.4.2	. . . social marketing prevention approach
	SN Using techniques from product marketing to influence the acceptability of a social idea or cause by the members of a group, or to influence attitudes, beliefs, (and behaviors such as drug use) with the goal of effecting social change.
	RT +JE6 prevention campaign
	+MI6 cultural sensitivity
	+MR2 marketing
	+JE6 mass media
JG10.4.4	. . . prevention through information dissemination
JG10.4.6	. . . prevention through education
	SN This approach aims to improve critical life and social skills (decision making, refusal skills, critical analysis, and systematic and judgmental abilities). Where appropriate, index the subject matter.
	ST *educational prevention approach*
	NT JH2.2.2 health promotion in the classroom
	BT +NF education and training
	RT MP18.2.8.16 harm reduction policy
	+NF16.2 AOD education
JG10.4.6.4	. . . . prevention through youth AOD education
JG10.4.6.6	. . . . prevention through parent AOD education
JG10.4.6.8	. . . . drinking and driving education
JG10.4.8	. . . peer prevention
JG10.8	. . prevention through spirituality and religion
JG10.10	. . prevention through public commitment
JG12	. **environmental-level prevention**
JG14	. . social policy prevention approach
JG12.4	. **multi-level prevention**

1.2 Challenges for digital libraries

To achieve the broad vision of enabling access to any data (information, knowledge, answer, digital object), digital libraries face many challenges, among them:

- Searching for text, images, sound, and composite objects – multimedia search.
- Semantically enhanced search to improve retrieval from free text and image content and to better exploit user-assigned tags.
- Search across many languages – multilingual search.
- Search across many systems – syntactic and semantic interoperability.
- Finding answers, not just documents; reasoning and inference.

Of course, there are also challenges of physical and semantic preservation (interoperability over time) and of hardware and software implementation, such as P2P and grid.

Another major theme in this vision of a comprehensive digital library or information space is *integration*:

- Integrate many presentation formats.
- Integrate libraries, archives, and museums; also databases and other information systems.
- Integrate reading/viewing/listening, database access, processing data, and authoring/creating.
- Integrate publishing and communication platforms.

This integration will result in a unified environment in which users can carry out all their work – work tasks and information / communication tasks, reading and authoring. The user need worry only about doing tasks, **not** about accessing different kinds of information formats and systems, selecting task-specific applications, or sharing information between applications [**?**].

1.3 Illustrative advanced DL functions

In advance of Section 3, Table 3 gives a glimpse of **advanced digital library functionality** organized into three major themes:

(1) **Document presentation**, interactive documents, learning objects
(2) **Tools for working with documents and data**, and
(3) **Tools for creating documents**

Table 3. Advanced digital library functionality

(1a) **Present documents in new ways**:

- Display the structure of hypertext documents as a graph.
- Show high-level overviews with drill-down to text, then to data and detail of methods; for example, graphical representation of the flow of ideas in a document, or concept maps showing the relationships between ideas in a document.
- Use moving video, sound, and animation to make ideas come alive.
- Let user control views of 3-D objects (rotate, select cross sections).
- Highlight named entities in the text, such as person names in a news reader.
- Integrate presentation approaches, multimedia.
- Provide alternate versions for different readers
  - by difficulty, such as an introduction to statistics with/without calculus;
  - by language (with automated translation as needed);
  - a spoken or braille version for the blind;
  - same data as text, table, or graphic — adapt to the reader's cognitive style.

(1b) **Present documents interactively**:

- Make mathematical formulas and procedure descriptions live (executable). For example, a document may present an economic model with links to the software and economic data sets so that the reader can run the model.
- Make the reader an active participant:
  - Interactive fiction; presenting questions or problems to be solved, with the answer determining further presentation;
  - simulations involving the reader; for example, a simulation of pricing decisions in a business textbook or an adventure game.

(2a) **Provide tools for working with documents:**

- Links from text terms to a thesaurus show hierarchical context and definition.
- Fine-grained search of text databases; find specific passages or facts within documents, incl. the document under study (s.a. information extraction below).
- Annotation and social tagging tools. Private and public annotation.
- Authoring tools. Integrate reading/viewing/listening and authoring/creating.
- Tools for working with images and sound documents.

(2b) **Provide tools for working with data**:

- Tools for importing data from tables in text.
- Tools for information extraction: Extract statements from text and insert them into a database (entity identification, relationship extraction).
- Tools for reasoning over a database.
- Tools for statistical processing.
- Tools for running a model over a set of data (economic or biological simulation).
- Tools for analyzing large instrument-collected data sets (e.g., gene chip data).
- Sequencing individual modules into processing chains to be run repeatedly.

(3) **Provide new ways of creating documents**

- Support producing documents by combining text, image, and sound modules already in the digital library ("writing in the large", virtual documents produced by a script).
- Auto-compile personalized documents; for example, a personalized textbook on statistics, taking a reader with a given subject interest and mathematics background from her present state of knowledge to her desired state of knowledge. Such documents could be implemented as paths through a hypertext, using prerequisite strands of concepts, such as in the AAAS Atlas of Science Literacy (see Table 12).
- Documents from data:
  - Text generation, graphs, visualization. Extreme case: automatically analyze instrument-generated data and then compose a paper reporting the results.
  - Web pages live with a database

1.4 Some examples of digital libraries and digital library software

Table 4 shows some **examples of digital libraries** that among them illustrate some of the functions listed. The table is arranged from more conventional DLs that focus on facilitating access to documents to DLs with more functionality. Most of these have some kind of subject directory users can browse, and some of them have most of their content contributed by users.

There is a wide range of software systems supporting the creation and maintenance of digital libraries; the examples in Table 5 illustrate the range. Content management software offers much DL functionality with a focus on collaborative content production and versioning, often including semantic-based search. Enterprise search software is also in this general arena; it often comes with powerful features supporting semantic search, such as linguistic processing, entity and relationship extraction, and automatic classification.

2 Characteristics of a digital library

Many ask "What is a digital library?", but the more important question is "What combination of system components and features best supports a user's work and other needs?". Rather than giving a definition of "digital library", this section discusses some typical characteristics of digital libraries and information systems in general, arranged by

- collection,
- user community served,
- purpose,
- specific functions and services.

Table 4. Examples of digital libraries (DLs)
Arranged roughly from more conventional DLs to DLs with more functionality

ACM DL ScienceDirect	Many professional associations (here the Association of Computing Machinery) and publishers (here Elsevier) have a DL of their journals, books, and reference works. Free access to bibliographic data, paid access to full text. portal.acm.org/dl.cfm sciencedirect.com
ICDL	International Children's Digital Library Focuses on digitizing children's books from around the world, making them findable through child-centered criteria, and facilitating online reading. icdlbooks.org
The Shoah VHA	52,000 videotaped interviews with Holocaust survivors, thesaurus of 4,000 subjects and 45,000 names of places, periods, people, etc. usc.edu/schools/college/vhi click Archive > About The Archive > The Visual History Archive
NSDL	National Science Digital Library (US). Support for education & collaboration nsdl.org
Connexion	A user-created DL of educational material; small knowledge chunks (modules) that can be organized as courses, books, reports. cnx.org
Wikipedia	Collaboratively constructed collection of anonymous encyclopedia articles wikipedia.org
Louvre	A museum Web site seen as a digital library containing both images and text, often with interactive features louvre.fr
Perseus	A rich network interconnecting places and sites, buildings, art objects (all represented by images), people, texts, words, ... Virtual walks through historical places. perseus.tufts.edu
Tufts University	An interesting array of DL-related tools uit.tufts.edu/at/?pid=24, uit.tufts.edu/at/?pid=24, dca.tufts.edu/tdr/pr/index.html

The characteristics are multi-faceted and often measured on a continuum. Any digital library or other information system can be described by a profile expressed in terms of these characteristics.

At the core of a digital library is a **collection**:

1. Typically a collection of digital objects that are of interest in their own right (primarily for reading, listening, viewing by people, but also for use by programs) rather than merely pointing to other objects. Examples:
   - a collection of digitized books (as opposed to just an online catalog),
   - a collection of biographies (as opposed to a personnel database),
   - a collection of oral histories,
   - a collection of software modules (on the margin of what many would consider a DL).

Table 5. DL software systems

Focusing on **digital repository** functions	**DSpace** (dspace.org)
	Fedora (fedora.info), see Tufts University, Table 4
Wider spectrum of DL functionality	**Greenstone** (greenstone.org)
	OpenDLib (opendlib.com/home.html)
	DELOS DLMS (Digital Library Management System, see delos.info, search for DLMS, more at dbis.cs.unibas.ch/delos_website/delosdlms.html). A software environment for integrating many tools
Content management See list in Wikipedia	**IBM DB2 Content Manager** (www-306.ibm.com/software/data/cm/cmgr/mp)
	Oracle Content Management SDK (oracle.com/technology/products/ifs/index.html)
	Documentum (documentum.com) **Vignette** (vignette.com)
	DRUPAL (drupal.org), **Joomla!** (joomla.org) (open source, managing Web sites)
Enterprise search	**IBM's OmniFind** (www-306.ibm.com/software/data/enterprise-search)
	Verity (verity.com)
	Convera's RetrievalWare (convera.com/solutions/retrievalware/default.aspx)
	Endeca (endeca.com), one of the growing breed of facet-based search engines

2. Typically, a collection for which items are carefully selected and acquired. Selection implies weeding, as some objects become less useful with age.
3. Typically, a collection which is curated; minimally, objects are preserved.

The collection of a DL can be described along many dimensions, among them

- Types of information objects included (text, images, sound recordings, learning objects);
- Origin of information objects by place and time;
- Content coverage of the information objects
  - Language of text objects
  - Subject domain
  - Place and time coverage.

The types of materials in the collection can also be characterized with respect to their suitability for given user groups and purposes (see below).

A DL that provides just a collection with access to known items is called a *digital repository*. Repository functions include document acquisition, safe storage and preservation, version control, finding known documents, and document presentation.

A digital library must manage composite documents, and the functionality it provides here is a DL characteristic. Composite documents can be quite complex, often including components in several media (multimedia documents), where the components are information objects in their own right. Components may be annotations. A composite object could be an entire database. (With native XML databases, the boundary between "document" and "database" has become completely blurred.) So a DL may offer, within one integrated environment, search for documents, search within documents, and search within one or more databases (as these terms are usually understood). A repository needs a *document model* [**?**, **?**] to manage these complexities, such as the XML-based Document Object Model (DOM) (see Wikipedia), or the Fedora Digital Object Model [**?**].

Even DIALOG (dialog.com), a service that provides search of over 600 databases, qualifies as a DL: it provides access to bibliographic databases that link to the full text of documents, full-text databases, and substantive databases with data on companies, chemical compounds, etc.

A **DL can be characterized by the user community it serves** along any number of demographic characteristics, such as age, level of education, subject specialty, or membership in an organization. Users can be both consumers of services and contributors if the DL allows.

A **DL can be characterized by the broad purposes it serves** or domain in which it operates, such as **scholarship, education, e-government, e-commerce (B2B or B2C), entertainment**, and more specific purposes, such as providing job-related information, supporting students with homework, supporting the internal work of an organization, supporting clients of an organization, supporting communication among users, etc.

Both user and purpose characteristics can also be used to characterize the types of materials in the collection with respect to their suitability for these users and purposes.

A **DL can be characterized by the functions it serves and the services it provides**.

3 Functions of digital libraries and beyond

To give the reader a sense of what one should expect from a DL and what is involved in establishing and maintaining a DL, this section gives an overview of functions an ideal DL would provide. An actual DL provides a subset of these functions, each at a given level of sophistication. The list provided here can serve a framework for describing the functionality of a DL.

This section is based on four premises for advanced DLs discussed in Section 1:

1. A DL has many functions and should integrate support for information seeking, users' work tasks, information production, and collaboration.
2. A DL links many types of information objects in many formats (including documents and databases) in all media into a complex structure.
3. Users both use and create information, and the processes of using and creating information are closely intertwined. The old distinction between producers of information (the few) and users of information (the many, the people, the masses) is rapidly fading away. Power to the people!
4. Digital libraries must interoperate.

Table 6 gives an illustrative list of DL functions.

Table 6. Illustrative functions of/tools provided by digital libraries and beyond

1 Search and other user-system interaction functions
. KOS (thesaurus or classification or ontology) related tools.
. . KOS use interface.
. . KOS creation and maintenance tool.
. Search and browse.
Search starts from a search element.
A search criterion specifies how the search targets relate to the search element.
The search element can be a term, a text passage, a whole document,
or a symbol,image, or sound bite (e.g. a note sequence), or a longer
sound passage, or a whole musical work, or another entity.
The search criterion specifies the relationship: the targets sought should
be about the search element, contain the search element, be similar
to search element, ...;
about could be limited (like a definition of a term) or broad.
. . Search at different depths.
. . . Search on catalog or metadata records (including social tags).
. . . Search on the full content of the documents (text, images,
sound) or of a database.

. . Input query.
. . . Invoke search from multiple places: Input search element or click on a search element found by navigating a hierarchy or in a text, or click on a page representing a document, person, or project to bring up a search form pre-filled to search for similar objects.
. . . Search form for the specification of field values and search logic.
. . . Save search form as an information object.
. . . Search or browse for previously used search forms, public or private.
. . Expand a search term or other search element using a thesaurus (possibly from a concept or terminology server on the Web).
. . Search multiple databases.
. . . Map query to several databases.
. . . Map results to local format (based on Z39.50, for example).
. . . Detect duplicates.
. . Browse.
. . Arrange search output in a meaningful way.
. Graph browser for browsing any kind of structure.
. . Hierarchy browser.

2 Reading tools and document creation and editing tools.
. Viewers for many types of information objects: text, images, 3-D objects, sound, multimedia objects, composite information objects. This includes many functions:
. . Handle and exploit many different document models and templates.
. . Assemble documents "on the fly"; for example, insert links and/or annotations stored in multiple separate locations.
. .Handle documents that are structured into different layers (e.g., appearance layer and OCR text layer, multivalent.sourceforge.net).
. . Manipulate objects, e.g., rotate a 3D object or view it in cross-section.
. An object viewer may be integrated with an object editor (see below).
. Reading, annotation, and linking tool.
. Sense-making tool. Assist users in creating structured representations (concept maps, templates) of data extracted from documents (by the user or automatically).
. Data stream processing tool.
. (Collaborative) authoring tool supporting the use of templates, version control, and authentication. (S. also "writing in the large", Table 3).
. Specialized tools used in creating text documents.
. . Bibliography generation tool.
. . Document glossary generation tool, using definitions from thesaurus
. . Document index and concordance generation tool.
. Access status management tool.
. Print tool. The Creation of print documents (and PDF files) from documents in many formats (esp. XML-tagged documents or several subdocuments linked together).

3 Document processing and link generation tools
- Tool for finding a full text document from metadata (wherever the full text may be).
- Document quality assessment tool.
- Document acquisition tool. Acquire documents in any format and convert them to the system format for easy processing by other tools.
- Document segmentation and tagging tool.
- Citation link and citation index tool (extract citations from text).
- Bibliographic citation parser to convert bibliography entries into bibliographic records.
- Metadata generation tool, incl. automatic/computer-assisted categorization and indexing.
- Summarization tool.

4 Collaboration tools.
- email support from within the DL, email documents and annotations.
- Online meeting support in all modalities, joint viewing of objects (e.g., book discussion forum).
- Archive of email, meetings, fora, with links to documents being discussed.
- (Collaborative authoring, see above).

5 DL management tools.
- Collection manager.
  - Collection analysis tool.
  - Selection and acquisition manager.
  - Weeding manager.
- Policy manager.
- Preservation manager.
  - Overall preservation program analysis.
  - Preservation monitoring of individual objects.
- Digital rights manager.
- External communications manager: Harvesting data and responding to requests from other systems, using protocols such as Z39.50 or OAI-PMH.
- Usage tracking.

4 Knowledge Organization and Knowledge Organization Systems

What is knowledge organization? This entails two questions: What is knowledge? and What is organization? For our purposes, knowledge is any representation of what is or will be or could be or should be or what is believed or asserted by some person or device, whether true or false; knowledge encompasses what some like to distinguish as data, information, and knowledge. Knowledge serves many purposes: planning, decision making, and action; satisfying curiosity; entertainment; healing (as in bibliotherapy). To be used, knowledge must be embodied in a person or device that can actuate it, reason with it, act on it, use it to govern the behavior of devices. To be useful,

knowledge must be organized. We organize information – in our minds and in information systems – in order to collect and record it, retrieve it, evaluate and select it, understand it, process and analyze it, apply it, and rearrange and reuse it. Somewhat tautologically, we can define organization as the arrangement of elements into a structure.

In a DL, knowledge organization comes into play in several closely interrelated ways:

1. Organization of information in substantive databases;
2. Organization of information within documents;
3. Organization of information about documents and databases (metadata);
4. Organization of information about any type of subject treated in documents (needed to support finding documents and other digital objects);
5. Information about concepts and terms and their relationships; organization of ontological and lexical information. Knowledge Organization Systems in the core sense.

There are **two important principles in applying knowledge organization** in digital libraries:

1. Use KOS **behind the scenes** to assist users and improve search and processing results.
2. When it is beneficial for users to interact with a knowledge organization system, provide user-friendly displays and interaction that **guide users in making sense** of what they see.

Underlying all systems for the representation and organization of knowledge is entity-relationship (E-R) representation. Table 7 gives an example of different kinds of data stored in a DL organized in an E-R representation; we will refer to this example throughout this section.

A DL contains many kinds of data. Data consist of statements (propositions, assertions), where a statement consists of a relationship binding together two or more entities. (In the Web context, esp. in RDF, entities are called resources and relationships are called properties; in topic maps entities are called topics.) Statements can be conceived as relationship instances. In a statement, one entity can be put in focus, and we can say that the statement is about the entity. For example, we can say that the statement

P15 <*runBy*> Drug-Free Community Coalition (DFCC)

is a statement about the project identified as P15; but is equally a statement about DFCC if we put DFCC in focus. In that case, we may want to write the same statement as

DFCC <*runs*> P15

Many relationships have two arguments (binary relationships, the easiest and most common case), but often relationships with three or more arguments are needed to express reality.

Table 7. Statements in a DL illustrating relationship types

P15	<*isa*>	Project
P15	<*hasTitle*>	Drinking is not cool
P15	<*runBy*>	DFCC (Drug-Free Community Coalition)
P15	<*hasCollaborator*>	ACS (Alay City Schools)
P15	<*fundedBy*>	NIAAA
P15	<*startDate*>	2006
P15	<*endDate*>	2009
P15	<*hasBudget*>	$1,200,000
P15	<*addressesProblem*>	Alcohol abuse
P15	<*hasTargetAudience*>	Adolescent girls
P15	<*usesApproach*>	Prevention through youth AOD education
ACS	<*isa*>	School system
ACS	<*isa*>	City government organization
D40	<*isa*>	Document
D40	<*publishedBy*>	DFCC
D40	<*hasTitle*>	AOD curriculum for teen girls
D40	<*dealsWith*>	prevention through youth AOD education
D40	<*hasComponent*>	D43
D43	<*isa*>	Image
D40	<*hasAnnotation*>	D58
D58	<*authoredBy*>	"Joe Smith"
D40	<*hasAccessRight*>	(Anna Cole, Modify)
D43	<*depicts*>	Girl
Girl	<*depictedNextTo*>	(Boy, D43)
(Alay, Houston, train)	<*hasTravelTime*>	3 hrs
City government organization	<*isSubclassOf*>	Local govt. organization
Local govt. organization	<*isSubclassOf*>	Government organization
Image	<*isSubclassOf*>	Document
prevention through youth AOD education	<*isSubclassOf*>	prevention through education
<*hasAnnotation*>	<*isSubrelationOf*>	<*hasComponent*>

When a user has a simple question, such as *What projects are run by DFCC?*, the system first checks its database for a directly matching answer statement. If none is found, the system tries inference. If that fails, the system tries to find a text source (or an image) that contains the answer and gives the user that source or, even better, extracts the answer from it. An answer to the following question can be found through a chain of inference combining several statements from the database:

What projects does NIAAA support that target adolescent girls in school?

We are now ready to discuss the different ways of knowledge organization in a DL.

4.1 Organization of information in substantive databases

Database organization is fundamental for DLs, since DLs can be seen as a special form of database. But more specifically, as we saw above, a DL can and often should include access to one or more substantive databases that can be queried in simple and in complex ways to provide immediate answers to users. Conceptually, all database structures are based on entity-relationship representation. The sample data from Table 7 could be stored as relational tables or as objects. The entity types and relationship types in an E-R data model provide the basis for defining tables or object classes.

Ideally, databases in similar domains would use a common E-R schema as the conceptual basis and similar schemas of tables or object classes. Of course, this does not happen, so interoperability requires schema mapping or schema integration — thorny problems in the database field.

4.2 Organization of information in documents

It is helpful to the reader if documents belonging to a given genre, such as project reports or descriptions of visual works or recipes, follow a common structure laid out in a *document template*. This idea can be implemented, for example, by using XML schema. The relationship types in Table 7 provide guidance for the project description template shown in Table 8.

Table 8. Document template for a project report (derived from Table 7, not complete)

<title>
<organizations involved>
<funder>
<time period>
<description>
<problem>
<targetAudience>
<approach>
<budget>

A DL or an organization that produces multiple types of documents must create and maintain a hierarchical system of document templates and the tags they use. Again, templates should be standardized, at least within communities of practice, to provide interoperability. A document model (see Section 2) is a highly abstract and general document template.

Some widely used document templates (document schemas) are

- TEI (Text Encoding Initiative, tei-c.org).
- MPEG (Moving Picture Experts Group) standards, especially MPEG-7, a schema for encoding both the structure of and metadata about multimedia documents.
- SCORM 2004 (Sharable Content Object Reference Model) for learning objects.

For a long list of specialized markup languages, see Wikipedia, List of XML markup languages, and the references to document models in Section 2.

A closely related concept is markup of text to identify specific types of information, such as people or standard names or dates; see, for example, the Orlando project (http://orlando.cambridge.org/).

Table 9. Dimensions for the analysis of metadata

Metadata can be analyzed along the following dimensions:
(1) the **purpose** for which the metadata are used and
(2) the **kind of information** given about a resource.
Some kinds of metadata are used for only one purpose, others for several purposes.

Dimension 1. The purpose for which the metadata are used

1.1 Resource (information) seeking and use, by stage in the information-seeking process
  1.1.1 Resource discovery: retrieval and selection of resources, specifically information objects, that are useful for a given purpose (are about a topic, illuminate an abstract theme, assist in performing a task, etc.).
  1.1.2 Dealing with a known resource: use and interpretation
1.2 Manage a resource (*administrative metadata*), in particular
  1.2.1 Manage the preservation of a resource (*preservation metadata*).

Dimension 2. The kind of information given about a resource.
(Categories overlap.)

2.1 Information about the intrinsic nature and the context of the resource.
  2.1.1 Information about identity and formal characteristics, including physical description(*descriptive metadata*).
  2.1.2 Information concerning what the resource is about and what it is relevant for (*subject metadata*).
  2.1.3 Information about the history, future disposition, and other features of the context of the resource (*contextual metadata*). Includes *provenance* (which in turn includes authorship) (also considered part of 2.1.1), *use history* and *relation to other resources*.
2.2 Information about how one can use the resource.
  2.2.1 Information on how to gain legal access to the resource (*access and use rights metadata*).
  2.2.2 Information on how to gain technical access to the resource (what machinery and software is needed to access the resource for a given purpose, such as assimilation by a person or processing by a computer program)(related to 2.1.1 physical description)(*technical metadata*).
2.3 Information about the status of a resource (past, present, and future), in particular
  2.3.1 Information about the preservation status of a resource.

4.3 Organization of information about documents. Data about data (metadata)

A piece of data is *used as* metadata if it is used for the purpose of discovering and using information objects which then give the ultimate data wanted; metadata are used to manage, find, interpret, and/or use other data or a source of such data. Note we said *used as* metadata. The "metadata-hood" of an information object does not reside in the information object, but in its relationship to another information object and, more specifically, in its use. The same piece of data may fill the ultimate need of one user and be used as metadata by another: A dean may use a bibliographic database to count the number of publications by a faculty member; she uses authorship data for her ultimate purpose, not as metadata. But we use authorship data as metadata if we use them to find a book from which we then learn what we need to know or to assess the authority of a book. In common usage today, data in a library catalog are considered metadata because they are most often used that way. By extension, similar data in other databases, such as a product catalog, are called metadata, even though they do not lead to other data. Metadata can be analyzed along a number of dimensions, as shown in Table 9.

Metadata schemas are usually represented as a set of tags that form a schema or template for a metadata record (a simple version of an object class). Each tag corresponds to a relationship type in an explicit or implied E-R schema. Metadata schemas are usually adapted to the type of information object and the user requirements. We list a few examples

- **Bibliographic metadata .** A widely used, but for many purposes overly simplistic, schema is the unqualified Dublin Core (DC, Table 10), but DC has many extensions (dublincore.org). The MARC format (loc.gov/marc) is a much more complete and fine-grained schema that covers many types of documents. RFC 1807 defines a bibliographic format for technical reports (ukoln.ac.uk/metadata/desire/overview/rev_19.htm). These three formats are specified in the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (openarchives.org). FRBR/CRM, the relatively new approach of dealing with bibliographic data, has its conceptual basis in E-R modeling (loc.gov/cds/FRBR.html, cidoc.ics.forth.gr/scope.html).

Table 10. A few of the 15 Dublin Core elements

Relationship type		**Dublin core element**
Document <*hasTitle*>	Text	<*dc:title* >
Document <*hasCreator*>	Person	<*dc:creator* >
Document <*dealsWith*>	Subject	<*dc:subject* >
Document <*publishedBy*>	LegalEntity	<*dc:publisher* >
Document <*publishedIn*>	Date	<*dc:date* >

- **Archival metadata.** Encoded Archival Description (EAD) (loc.gov/ead).
- **Metadata for learning objects** (instructional materials) has several standards

  The Gateway to Educational Materials (GEM)
  (thegateway.org/about/documentation/metadataElements)

  The Learning Technology Standards Committee of the IEEE
  (ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf)

  IMS learning resource metadata information model (imsproject.org/metadata)

  The DCMI Education Working Group. dublincore.org/groups/education

  The CRP Henri Tudor-CITI: Training Exchange Definition: TED.
  (xml.org/xml/schema/8dbca03a/trainingExchangeDefinition.pdf)
- **Metadata for geospatial data sets**.
  ISO 19115:2003 Geographic information – Metadata iso.ch/iso/en
  fgdc.gov/metadata/geospatial-metadata-standards

4.4 Organization of information about any type of subject treated in documents (needed to support finding documents)

Section 4.1 dealt with access to substantive databases to obtain immediate answers. The nature of the data referred to here is the same, but the use is different: People often look for documents (texts, images, music) based on their content or the circumstances under which they were created. Such queries require a database that can deal with any type of content. To find portraits of physicists, we need a biographic database where we can find physicists so we can then find their portraits; to find descriptions of medieval houses, we need a database of buildings from which we can find medieval houses (each identified by a building identifier) so we can then find texts that describe any of these buildings. In the same way we could find descriptions of buildings designed by the famous architect Pei. The richer this database of content, the richer the possibilities for the user to make connections in finding documents. This is the basic idea behind Topic Maps, a standard for specifying relational databases that are optimized for this retrieval support function. Examples of objects, concepts, ideas of particular importance in this context include

People	Concepts, ideas
Organizations	Problems and proposed solutions
Events	Computer programs
Places	Mathematical theorems
Dates	

Examples of tools that fall under here are *gazetteers*, databases about places and place names, and biographic databases. The tools discussed in

Sections 4.5 - 4.9 are a subset: databases of concepts, ideas, and their names (terms), or, put differently, systems that organize ontological and lexical information (information about concepts and terms and their relationships); such systems are called *Knowledge Organization Systems (KOS)*.

4.5 What is a Knowledge Organization System (KOS)? A first look

People often search for subjects, concepts, or ideas, and these are the most difficult searches. Often the user just wants to type in some words, but those words do not always express what the user has in mind. Other times, the user is at a loss for words. **DLs must support the user in this quest for meaning**. This is one of the chief roles of Knowledge Organization Systems. In some information systems such searches are supported by manual, computer-assisted, or automatic subject cataloging or indexing, using a controlled vocabulary of terms or other concept designations stipulated in the KOS. For an example of hierarchy that epitomizes what a KOS is, see Table 2. There are many types of KOS, some of them are listed in Table 11. Table 12 gives pointers to illustrative KOS.

Table 11. The many forms of Knowledge Organization System

Dictionaries, glossaries	Concept maps
Thesauri, subject heading lists	Classification schemes, taxonomies
Topic maps	Ontologies with rich semantic relationships

A **dictionary** is a listing of words and phrases, giving information such as spelling, morphology and part of speech, senses, definitions, usage, origin, and often equivalents in other languages.

A **thesaurus** manages the complex relationships between terms and concepts and provides conceptual relationships, ideally through an embedded classification. A thesaurus may specify descriptors authorized for indexing and searching. These descriptors form a controlled vocabulary (authority list, index language). A monolingual thesaurus has terms from one language, a multilingual thesaurus from two or more languages.

A **classification** is a structure that organizes concepts into a hierarchy, possibly in a scheme of facets. The term **taxonomy** was originally used for the classification of living organisms, then expanded to any kind of classification. The term **typology** is used for small classifications, often in the context of research studies. The term **ontology** is often used for a shallow classification of basic categories or a classification used in linguistics, data element definition, or knowledge management or (increasingly) for any classification. In AI-related contexts, an ontology is a classification with a rich set of semantic relationships that support reasoning.

Table 12. Examples of Knowledge Organization Systems

AOD	**Alcohol and Other Drug Thesaurus.** Universal, semi-faceted etoh.niaaa.nih.gov/AODVol1/Aodthome.htm
MeSH	**Medical Subject Headings** hierarchical, available in many languages nlm.nih.gov/mesh/meshhome.html nlm.nih.gov/mesh/MBrowser.html
UMLS	**Unified Medical Language System** over 100 biomedical KOS in one database nlm.nih.gov/research/umls/umlsmain.html, umlsinfo.nlm.nih.gov
NCI	**National Cancer Institute Thesaurus** nciterms.nci.nih.gov/NCIBrowser/Dictionary.do
AAT	**Art and Architecture Thesaurus** getty.edu/research/tools/vocabulary/aat/index.html
AGROVOC	AGROVOC (agriculture. fisheries, forestry), in many languages fao.org/agrovoc
ERIC	**Education Resources Information Center Thesaurus**. searcheric.org
LCSH	**Library of Congress Subject Headings** for alphabetic subject access loc.gov/cds/lcsh.html Search http://authorities.loc.gov/
LCC	**Library of Congress Classification** for shelving / directory loc.gov/catdir/cpso/lcco/
DDC	**Dewey Decimal Classification.** Semi-faceted, intended for shelving or directory oclc.org/dewey/about/default.htm
Yahoo	**Yahoo classification**. Semi-faceted, Web subject directory yahoo.com
ASL	**Atlas of Science Literacy** project2061.org/publications/atlas/default.htm
CYC	**CYC Ontology** cyc.com/cycdoc/vocab/merged-ontology-vocab.html
GO	**Gene Ontology** geneontology.org
WN	**WordNet** A rich dictionary database with a built-in classification cogsci.princeton.edu/∼wn, search notredame.ac.jp/cgi-bin/wn

4.6 The many functions of Knowledge Organization Systems

A KOS can serve many functions (see Table 13). Understanding this simple truth is of paramount importance if one wants to maximize the return on the large investment required to construct a good KOS.

One of the most important, but often ignored, functions of an index language that supports retrieval is to make sure that documents are indexed or tagged with concepts that reflect user interests so users can actually formulate queries that express their interest and find what they are looking for. The principles of this request-oriented (user-centered) approach to indexing are summarized in Table 15, with some examples given in Table 16.

Table 13. Functions of a KOS

Semantic road map to individual fields and the relationships among fields.
- Map out a concept space, relate concepts to terms, and provide definitions.
- Clarify concepts by putting them in the context of a classification.
- Relate concepts and terms across disciplines, languages, and cultures.
- Many specific functions build on this foundation.

Improve communication. Support learning & assimilating information.
- Conceptual frameworks for learners. Help learners ask the right questions.
- Conceptual frameworks for the development of instructional materials.
- Assist readers in understanding text by giving the meaning of terms.
- Assist writers in producing understandable text by suggesting good terms.
- Support foreign language learning.

Conceptual basis for the design of good research and implementation.
- Assist researchers and practitioners with problem clarification.
- Consistent data collection, compilation of (comparative) statistics.

Classification for action. Classification for social and political purposes.
- Classification of diseases for diagnosis; of medical procedures for billing; of commodities for customs.

Support information retrieval and analysis. Retrieval of goods and services for e-commerce.
- Support searching, esp. knowledge-based support for end-user searching:
  - assistance in clarifying the search topic;
  - (automatic) synonym expansion and hierarchic expansion (see Table 14)
- Support indexing, especially request-oriented (user-centered) indexing.
- Facilitate the combination of or unified access to multiple databases.
- Support meaningful, well-structured display of information.
- Support document processing after retrieval.

Ontology for data element definition. Data element dictionary.

Conceptual basis for knowledge-based systems. Example:

Reading instruction $<hasDomain>$ Reading AND Reading ability $<hasDomain>$ Reading AND Reading ability $<supportedBy>$ Perception $\Longrightarrow$ inference: Reading instruction $<shouldConsider>$ Perception

Do all this across multiple languages

Mono-, bi-, or multilingual dictionary for human use.

Lexical knowledge base for natural language processing (NLP).

Table 14. Query term expansion. Example

A search for Drug use by teenagers formulated as *teenage AND drug* will find
Drug Use Rises for Teenagers **but miss** Adolescent Drug Abuse Treatment Outcome, KCEOC Substance abuse/youth program, and Smoking still increasing among teens
To find these, use automatic manual query term expansion: OR synoyms and narrower terms and their synonyms, as follows (illustrative only):
(teenage OR teen OR youth OR adolescent OR "high school") AND (drug OR substance OR nicotine OR smoking OR cigarette OR cocaine OR crack)

Table 15. User-centered indexing / request-oriented indexing principles

Construct a **classification/ontology from** actual and anticipated **user queries and interests**.
Thus provide a **conceptual framework** that organizes user interests into a meaningful arrangement and communicates them to indexers.
Index materials from users' perspectives: Add need-based retrieval clues beyond those present in the document. Increase probability that retrieval clues helpful to users are available.
Index language as checklist. Indexing = judging relevance against user concepts. Judging relevance goes beyond just determining aboutness.
Implementation: Knowledgeable indexers. Expert system using syntactic & semantic analysis & inference. **Social tagging**: tags based on user's own interests.

Table 16. Request-oriented indexing. Examples

Document	User concept
The drug was injected into the aorta	*Systemic administration*
Children of blue-collar workers going to college	*Intergenerational social mobility*
CSF studies on alcoholism and related behaviors	*Biochemical basis of behavior*
Drug use among teenagers (read methods section)	Longitudinal study
Image	Good scientific illustration
Image	Useful for fundraising brochure
Image	*Appealing to children*
Image	*Cover page quality*

4.7 The structure of KOS

The structure of a comprehensive KOS consists of two levels:

Level 1: Concept-term relationships (Section 4.7.1)

Level 2: Conceptual structure (Section 4.7.2)

- 2.1 Semantic analysis and facets (Section 4.7.2.1)
- 2.2 Hierarchy (Section 4.7.2.2)
- 2.3 Interaction of hierarchy and facets (Section 4.7.2.3)
- 2.4 Differentiated (refined) concept relationships (Section 4.7.2.4)

Some KOS focus on only one of these two levels. For example, many KOS that are labeled ontologies focus on the concept level and do not worry about terms.

4.7.1 Concept-term relationships

Table 17 gives some examples of concept-term relationships, making clear the need for vocabulary control either at the point of indexing (controlled vocabulary) or at the point of searching (query term expansion), especially in searching based on free text and user-assigned (social) tags.

Table 17. Concept-term relationships (Terminological structure)

Controlling synonyms (one concept - many terms)		**Disambiguating homonyms** (One term - many concepts)
Term	***Preferred term***	administration 1 (management)
		administration 2 (drugs)
teenager	adolescent	
teen	adolescent	discharge 1 (from hospital or program)
youth (person)	adolescent	discharge 2 (from organization or job)
pubescent	adolescent	Preferred synonym: Dismissal
		discharge 3 (medical symptom)
alcoholism	alcohol dependence	discharge 4 (into a river)
drug abuse	substance abuse	discharge 5 (electrical)

4.7.2 Conceptual structure

The key to a KOS that fulfills the functions listed in Table 13 is the **conceptual organization**. There are two interacting principles of conceptual structure: **facet analysis** (componential analysis, feature analysis, aspect analysis, semantic factoring) and **hierarchy**.

Table 18. A facet frame for prevention projects (derived from Table 7).

Relationship	Facet	Sample facet value
<*addressesProblem*>	*Problem, disorder, disease*	**Alcohol abuse**
<*hasTargetAudience*>	*Target audience, population*	**Adolescent girls**
<*usesApproach*>	*Approach*	**Prevention through youth AOD education**

4.7.2.1 Semantic analysis and facets

Facet analysis is best understood through examples. From the entity-relationship schema in Table 7, we can see three of the facets needed to analyze and describe a prevention project, repeated in the *facet frame* in Table 18.

Each facet describes one aspect of the project. Facet analysis is a great way to conceptualize a search, hence the increasing popularity of facet-based search [**?**, **?**]. Each facet (or slot in the facet frame) has an associated set of values (slot fillers); Table 2 gives a hierarchy of values of the *Approach* facet.

For *disorders,* carrying facet analysis further leads to atomic (or elemental) concepts, see Table 19. Table 20 gives general facet principles and Table 21 gives more examples.

Table 19. More facet examples. Facet frame for disorders

alcohol abuse		**alcoholic liver cirrhosis**	
Pathologic process:	**substance abuse**	*Pathologic process*:	**inflammation**
Body system:	**not specified**	*Body system*:	**not specified**
Cause:	**not specified**	*Cause*:	**chem.induced**
Substance/organism:	**alcohol**	*Substance/organism*:	**alcohol**
hereditary alcohol abuse		**hepatitis A**	
Pathologic process:	**substance abuse**	*Pathologic process*:	**inflammation**
Body system:	**liver**	*Body system*:	**liver**
Cause:	**genetic**	*Cause*:	**infection**
Substance/organism:	**alcohol**	*Substance/organism*:	**hepatitis A virus**

4.7.2.2 Hierarchy

Table 2 gives an example of a concept hierarchy. For information retrieval and user orientation the main purposes of hierarchy are

(1) hierarchic query term expansion and

(2) organizing concepts into a structure that can be easily understood.

For (1) we can define broader concept (usually called Broader Term, abbreviation BT) pragmatically as **Concept B falls under broader concept A if any search for A should find everything on B as well** (B BT A or conversely, A has narrower concept B, A NT B). For (2) create suitable headings to structure the hierarchy. Table 22 shows another example of

a pragmatic hierarchy. Reasoning requires a more formal definition: B <*isa*> A, which means that instances of concept B have all the characteristics of concept A and at least one more.

Table 20. General facet principles

A facet groups concepts that fill the same role:

- concepts that fall under the same aspect or feature in the definition of more complex concepts;
- concepts that can be answers to a given question;
- concepts that can serve as fillers in one frame slot;
- concepts that combine in similar patterns with other concepts.

Elemental concepts as building blocks for constructing compound concepts:

- Reduces the number of concepts in the KOS, leading to conceptual economy.
- Facilitates the search for general concepts, such as searching for the concept *dependence* (in medicine, psychology, or social relations).

Facets can be defined at high or low levels in the hierarchy; see Table 21.

Table 21. More facet examples

Top-level facets	**Low-level facets**
pathologic process	**route of administration**
organism	. by scope of drug action (local/topical or systemic)
body part	. by body site
chemical substances by function	. by method of application (injection, rubbing on, etc.)
chemical substances by structure	

A Area of ability		combines with	**B Degree of ability**	
A1	psychomotor ability		B1	low degree of ability, disabled
A2	senses		B2	average degree of ability
A2.1	. vision		B3	above average degree of ability
A2.1.1	. . night vision		B3.1	. very high degree of ability
A2.2	. hearing			
A3	intelligence			
A4	artistic ability			

Examples
A2.1B1 visually impaired
A2.2B1 hearing impaired
A3B1 mentally handicapped
A3B3 intellectually gifted

Table 22. Hierarchy example

groups at highrisk of drug use
. *at high risk of drug use due to family background*
. . persons from unstable or low-cohesion families
. . children of alcoholic or other drug-abusing parents
. . children of single teenage mothers
. *at high risk of drug use due to abuse or neglect*
. . persons subjected to abuse/neglect by parents
. . . latchkey children
. . persons subjected to abuse/neglect by spouse
. *at high risk of drug use due to internal factors*
. . suicidal or physically or mentally disabled
. . gateway drug users
. . persons engaged in violent or delinquent acts
. *at high risk of drug use due to external circumstances*
. . single teenage mothers
. . school dropouts or those at risk of dropping out
. . unemployed or in danger of being unemployed
. . economically disadvantaged
. . homeless
. . . runaway youth

4.7.2.3 Interaction of hierarchy and facets

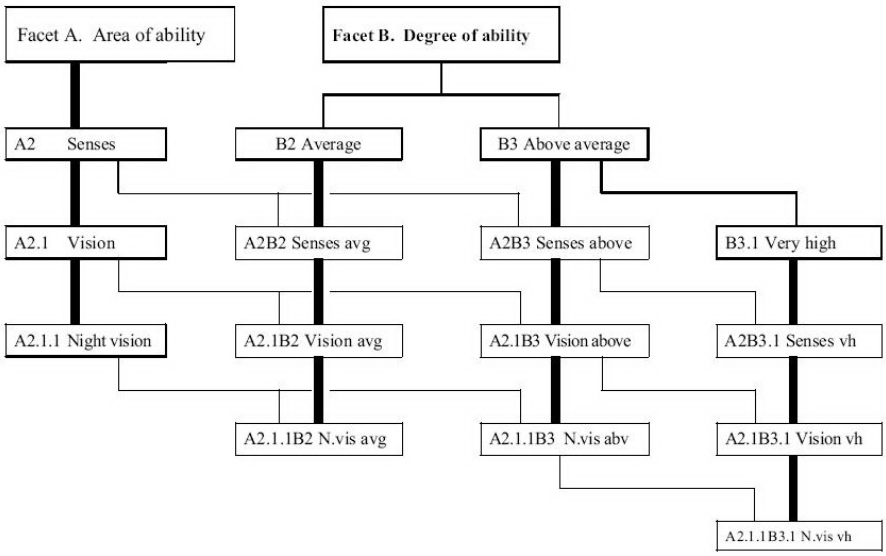
Table 21 illustrates how concepts from two facets can be combined to form compound concepts. Combination of concepts may be done just in searching (postcombination) or ahead of search in indexing (precombination):

(1) **Postcombination** (usually, but poorly, called postcoordination). Assign elemental (atomic) concepts as descriptors and combine descriptors in searching. In the example this would not work well, since a person could have below average vision and above average hearing.

(2) **Precombination**. (usually, but poorly, called precoordination). Assign compound concepts as precombined descriptors. These precombined descriptors can be enumerated in the classification schedule ready for use by the indexer, or the indexer must construct precombined descriptors as needed.

The compound concepts form a hierarchy. To find all people who are *above average in vision*, we must also find people who are *above average in night vision* or who are *very high in vision*. Table 23 shows a small hierarchy that results from combining the concepts from two facets (not the facet heads) and shows it both in a graphical representation and as a linear arrangement with cross-references. Many concepts have two broader concepts (polyhierarchy) as shown through the cross-references. As a more efficient alternative to cross-references, one can set up a system that finds compound descriptors in terms of their elemental components (descriptor-find index).

Table 23. Hierarchy from combining two facets



Linear arrangement 1: Facet A primary

A Facet A. Area of ability
. A2 Senses
. . A2B2 Senses average NT A2.1B2; BT B2
. . A2B3 Senses above average NT A2.1B3; BT B3
. . . A2B3.1 Senses very high NT A2.1B3.1; BT B3.1
. . A2.1 Vision
. . . A2.1B2 Vision average NT A2.1.1B2; BT A2B2
. . . A2.1B3 Vision above average NT A2.1.1B3; BT A2B3
. . . . A2.1B3.1 Vision very high NT A2.1.1B3.1; BT A2B3.1
. . . A2.1.1 Night vision
. . . . A2.1.1B2 Night vision average BT A2.1B2
. . . . A2.1.1B3 Night vision above average BT A2.1B3
. . . . . A2.1.1B3.1 Night vision v. high BT A2.1B3.1

B Facet B. Degree of ability
. B2 Average ability NT A2B2
. B3 Above average ability NT A2B3
. . B3.1 Very high ability NT A2B3.1

Linear arrangement 2: Facet B primary not shown

4.7.2.4 Differentiated (refined) conceptual relationships

Precise query term expansion and reasoning to find answers require more precise concept relationships than provided in a traditional thesaurus, as illustrated in Tables 24, 25, and 26 (Tables 25 and 26 adapted from [**?**]).

The example in Table 24 illustrates the use of refined relationships for more precise query expansion. Assume a user is interested in finding documents on *reading and reading materials*. In the ERIC Thesaurus she could either search for just *Reading* or use all RTs for query term expansion, i.e., *Reading OR all RTs of Reading* – a very imprecise search. Using the refined ontology, the user could instead specify a much more precise selection of search terms:

Reading OR all concepts X for which Reading <*isAppliedTo*> Concept X.

Table 24. Refined relationships for more precise query expansion: ERIC Thesaurus - A sample of the RT relationships under **Reading**

ERIC Thesaurus	**Refined Ontology**
READING	
RT Advance organizers	<*facilitatedBy*> Advance organizers
Bibliotherapy	<*usedIn*> Bibliotherapy
Context clues	<*facilitatedBy*> Context clues
Readability	<*facilitatedBy*> Readability
Reading ability	<*supported/hinderedBy*> Reading ability
Reading assignments	<*isAppliedTo*> Reading assignments
Reading attitudes	<*supported/hinderedBy*> Reading attitudes
Reading games	<*isAppliedTo*> Reading games
Reading materials	<*isAppliedTo*> Reading materials
Reading motivation	<*supported/hinderedBy*> Reading motivation
Reading readiness	<*supported/hinderedBy*> Reading readiness
Reading skills	<*supported/hinderedBy*> Reading skills

Table 25 and Table 26 give examples of inferences that rely on the detailed semantic relationships given in an ontology. But the ERIC thesaurus gives us only some poorly defined broader term (BT) and related term (RT) relationships. These relationships are not differentiated enough to support inference.

For another example, consider the ontological relationships and rules we could formulate with these relationships in an example taken from the AGROVOC thesaurus in Table 26. From the statements and rules given in the ontology, a system could infer that *Cheddar cheese* <*containsSubstance*> *milk fat* and, if cows on a given farm are fed mercury-contaminated feed, that *Cheddar cheese* made from milk from these cows <*mayContainSubstance*> *mercury*. But the present AGROVOC Thesaurus gives only NT/BT relationships without differentiation.

In both examples, many of the relationships are based on statements "about the world" rather than just conceptual definitions or terminology, blurring the distinction between KOS and the systems discussed in Sections 4.1 and 4.4.

Table 25. Refining relationships for inference: ERIC thesaurus

Eric Thesaurus	**Refined ontology**: Statement and rules
reading instruction	**reading instruction**
BT instruction	<*isSubclassOf*> instruction
RT reading	<*hasDomain*> reading
RT learning standards	< *governedBy*> learning standards
reading ability	**reading ability**
BT ability	<*isSubclassOf*> ability
RT reading	<*hasDomain*> reading
RT perception	<*supportedBy*> perception
	Rule 1: Instruction in a domain should consider ability in that domain: **X *shouldConsider* Y** **IF** X <*isSubclassOf*> instruction **AND** X <*hasDomain*> W **AND** Y <*isSubclassOf*> ability **AND** Y <*hasDomain*> W yields: (The designer of) *reading instruction* should consider *reading ability.* **Rule 2** **X *shouldConsider* Z** **IF** X <*shouldConsider*> Y **AND** Y <*supportedBy*> Z yields: (The designer of) *reading instruction* should consider *perception.*

Table 26. Refining relationships for inference: AGROVOC Thesaurus

AGROVOC	**Refined Ontology**
milk	**milk**
NT cow milk	<*includesSpecific*> cow milk
NT milk fat	<*containsSubstance*> milk fat
cow	**cow**
NT cow milk	<*hasComponent*> cow milk
Cheddar cheese	**Cheddar cheese**
BT cow milk	<*madeFrom*> cow milk
	Rule 1 **Part X <*mayContainSubstance*> Substance Y** **IF** Animal W <*hasComponent*> Part X **AND** Animal W <*ingests*> Substance Y **Rule 2** **Food Z <*containsSubstance*> Substance Y** **IF** Food Z <*madeFrom*> Part X **AND** Part X <*containsSubstance*> Substance Y

4.8 Interoperability. KOS standards

Syntactic interoperability requires that system A know how to read system B data and recognize a character string as a tag or relationship or term. *Semantic interoperability* requires that system A know also the *meaning* of the tag, relationship, or term in system B. Semantic interoperability is much harder. Searching or combining data across languages, cultures, disciplines, or time can be seen as an interoperability problem. There are two ways to approach interoperability: Standards, so system A and B use the same syntax and even the same semantics, and mapping (or cross-walks). In practice, both are used. Standards are easier on the syntactic level than on the semantic level. KOS standards are predominantly syntactic, unless one wants to consider widely used KOS, such as Dewey Decimal Classification as (de facto) standards, at least in a limited domain (such as public libraries in the US).

KOS standards serve three main functions:

- Input of KOS data into programs and transfer of data between programs.
- Querying KOS by people and programs and viewing results.
- Identifying specific terms/concepts in specific KOS, e.g., a unique URI (Universal Resource Identifier) for every concept and term to enable cross-KOS concept relationships and use of such URIs in metadata.

A KOS standard must specify the types of information to be included about each concept and term (relationship types, data fields, and standard symbols for them), as well as, for example, information needed to render a hierarchical display in outline form (with meaningful arrangement of concepts at the same level) or a graphical display, such as a concept map.

Unfortunately, there is no unifying standard for all types of KOS but rather a bewildering array of standards for different types of KOS (see Table 27). As a consequence, KOS management software is also splintered, making it almost impossible for an organization to develop and maintain the type of integrated multi-functional KOS that would be most cost-effective.

4.9 Unification: Ontologies

The relationships between document components in a document model, the tags in a document template or a metadata schema, the table structures in a relational database (or the object structures in an object-oriented database), and the relationships between concepts can all be traced back to (or defined in terms of) an entity-relationship model (possibly with added features to increase expressiveness). Such a model is an ontology, so all structures in a digital library can (and should) be conceived as subsets of an overarching ontology. This ontology can be used to make sure that all structures within the DL are consistent and that new structures, such as a template for a new type of document, can be developed easily and consistently. Ideally, the design of a new DL would start with an ontology.

Table 27. KOS Standards

Dictionaries, glossaries good ISO standards:
ISO 12200:1999, Computer applications in terminology–Machine Readable Terminology Interchange Format (MARTIF)–Negotiated Interchange
ISO 12620:1999, Computer applications in terminology–Data Categories
Many ISO terminology standards
Thesauri
ISO 2788-1986(E) / ANSI/NISO Z39.19-2005 (*niso.org) Poor and backwards
BS8723, Structured vocabularies for information retrieval. Good
Simple Knowledge Organisation Systems (SKOS) RDF name space see w3.org/2004/02/skos/ Restricted in expressiveness but widely used
ISO 5964-1985(E) (multilingual)
USMARC format for authority data (lcweb.loc.gov/marc/authority/ecadhome.html)
Topic maps (reference works, encyclopedias) (topicmaps.org/about.html)
ISO/IEC 13250:2000 Topic Maps
XML Topic Maps (XTM) 1.0 (topicmaps.org/xtm/1.0/)
Concept maps
Classification schemes
MARC for classification data lcweb.loc.gov/marc/classification/eccdhome.html
Ontologies
OWL Web Ontology Language, an extension of RDF (w3.org/TR/owl-ref)
Knowledge Interchange Format (KIF) (meta2.stanford.edu/kif/dpans.html)
Generic standards for knowledge structures, entity-relationship models
Resource Description Framework (RDF) (w3.org/RDF/)
The Topic Map standard belongs here as well.

Table 7 gives some examples of statements using entity types and relationship types that would be defined in an ontology; subsequently, the project description template in Table 8 was defined guided by some of these relationships. Likewise, the facet frame for prevention projects in Table 18 was derived from the ontology illustrated in Table 7. Concepts and statements about concepts are conventionally considered to be part of the ontology.

The ontological basis also supports interoperability: Mapping between the ontologies of two DLs, though by no means easy, allows one to derive mappings between the templates, schemas, and KOS used by the two DLs.

Good starting points for finding information about ontologies are the Web sites of Ontolog (ontolog.cim3.net/) and Barry Smith Web site (ontology.buffalo.edu/smith/).

5 Conclusion

Digital libraries with powerful semantic support (1) for complex searches for documents and immediate answers across system, language, cultural, and disciplinary boundaries and (2) for document creation and collaboration have the potential to transform how work is done by individuals and by groups and to evolve into a true “information commons”.

Acknowledgments

The discussions in the DELOS Digital Library Reference Model Working Group (delos.info) were very helpful in preparing this chapter.