Systems of Knowledge Organization for Digital Libraries:

Beyond Traditional Authority Files

April 2000

by Gail Hodge

ISBN 1-887334-76-9

Published by:

The Digital Library Federation
Council on Library and Information Resources
1755 Massachusetts Avenue, NW, Suite 500
Washington, DC 20036

Additional copies are available for $15.00 from the address noted above. Orders must be prepaid, with checks made payable to the Council on Library and Information Resources.

The paper in this publication meets the minimum requirements of the American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials ANSI Z39.48-1984.

The Digital Library Federation

On May 1, 1995, 16 institutions created the Digital Library Federation (additional partners have since joined the original 16). The DLF partners have committed themselves to "bring together—from across the nation and beyond—digitized materials that will be made accessible to students, scholars, and citizens everywhere." If they are to succeed in reaching their goals, all DLF participants realize that they must act quickly to build the infrastructure and the institutional capacity to sustain digital libraries. In support of DLF participants' efforts to these ends, DLF launched this publication series in 1999 to highlight and disseminate critical work.

About the Author

Gail Hodge is a senior information specialist at Information International Associates, Inc. (IIa). She has worked in the information industry for more than 20 years, specializing in bibliographic database production systems, information systems planning and development, and standards. She conducts research on scientific information policy, technologies, and standards for the U.S. government and for international and commercial organizations. Recent projects include an analysis of the state-of-the-practice in digital archiving for the International Council for Scientific and Technical Information, and development of controlled vocabularies for the National Institute for Literacy, the Department of Energy's Environmental Management Science Program, and the National Biological Information Infrastructure. Before joining IIa, Ms. Hodge held positions at the Drexel University Library, BIOSIS, and the NASA Center for AeroSpace Information.

Acknowledgments

I would like to thank the following individuals for their input to this report:

Stan Blum, California Academy of Science
Bonnie C. Carroll, Information International Associates, Inc.
Michael Dadd, General Manager, BIOSIS UK
Linda Hill, Alexandria Digital Library Project, University of California at Santa Barbara
Ray Larson, University of California at Berkeley
Jessica Milstead, The JELEM Company
Douglas Yanega, Entomology Research Museum, University of California at Riverside

Contents

Foreword vi

Executive Summary 1

1. Knowledge Organization Systems: An Overview 3

- Common Characteristics of Knowledge Organization Systems 4
- Types of Knowledge Organization Systems 4
  - Term Lists 5
  - Classifications and Categories 6
  - Relationship Lists 6
- The Origin and Use of Knowledge Organization Systems 7
  - Abstracting and Indexing Services 8
  - Publishers 9
  - Trade, Professional, and Governmental Organizations 9
  - Internal Projects 9
- Summary 9

2. Linking Digital Library Resources to Related Resources 10

- Expanding Codes to Full Text 10
  - Linking Sequence Numbers to Biosequence Databanks 10
  - Linking Individual Industrial Codes to the Full Scheme 11
- Linking to Descriptive Record 11
  - Linking Organism Names to Taxonomic Records 12
  - Linking Chemical Names to Molecular Structures 13
  - Linking Personal Names to Biographical Information 14
- Linking Entity Names to Physical Specimens 15
- Summary 16

3. Making Resources Accessible to Other Communities 17

- Providing Alternate Subject Access 17
  - Indexing the Material with Multiple Schemes 17
  - Retaining Alternate Indexing from Contributions 18
  - Mapping Multiple Schemes 18
- Adding New Modes of Understanding to the Digital Library 19
- Providing Multilingual Access 21
- Expanding Free-Text Search Terms 21
- Summary 22

4. Planning and Implementing Knowledge Organization Systems in Digital Libraries 23

Planning Knowledge Organization Systems 23
Analyzing User Needs 23
Locating Knowledge Organization Systems 23
Planning the Infrastructure 24
Maintaining the Knowledge Organization System 26
Presenting the Knowledge Organization System to the User 26
Implementing Knowledge Organization Systems 27
Acquisition and Intellectual Property Issues 27
Making the Link 28
Summary 30

5. The Future of Knowledge Organization Systems on the Web 31

6. Conclusion: Enhancing Digital Libraries with Knowledge Organization Systems 34

References 35

Foreword

Access to digital materials continues to be an issue of great significance in the development of digital libraries. The proliferation of information in the networked digital environment poses challenges as well as opportunities. The Digital Library Federation is committed to fostering work that addresses these challenges and opportunities while also ensuring the timely dissemination of information about state-of-the-art initiatives.

The author reports on a wide array of activities in the field. While this publication is not intended to be exhaustive, the reader will find, in a single work, an overview of systems of knowledge organization and pertinent examples of their application to digital materials. Technological developments have made it possible to provide alternate subject access through the adoption and use of multiple knowledge organization schemes. The report offers extensive practical information for institutions embarking on digital library initiatives. In particular, the section on planning and implementing organization systems identifies methods for enhancing access to existing digital materials.

Rebecca Graham
Research Associate

Executive Summary

Librarians are increasingly called upon not only to collect information in electronic form but also to organize it into digital libraries. The materials may be created and held locally, or they may be created and accessed in a distributed fashion as a virtual library. Digital libraries can provide material on a variety of topics, from children's games to high-energy physics. Their scope may be local, national, or even international; the audience may be a small group with specialized interests or the broader public. Essential to the successful implementation and use of any digital library is the organization of that library, either directly or indirectly, by one or more knowledge organization systems (KOS).

The term *knowledge organization systems* is intended to encompass all types of schemes for organizing information and promoting knowledge management. Knowledge organization systems include classification and categorization schemes that organize materials at a general level, subject headings that provide more detailed access, and authority files that control variant versions of key information such as geographic names and personal names. Knowledge organization systems also include highly structured vocabularies, such as thesauri, and less traditional schemes, such as semantic networks and ontologies. Because knowledge organization systems are mechanisms for organizing information, they are at the heart of every library, museum, and archive.

The knowledge organization system used in a particular situation may be borrowed from the library tradition, such as the Library of Congress Classification Schedule, or from commerce, such as the Yahoo categories, or it may be developed locally. The KOS may be applied to metadata records for each resource, embedded in metatags, or separated from the digital library resources as part of the access mechanism. Regardless of its location with respect to the resource, its origin, or its type, the KOS has a single purpose: to organize content to support retrieval of relevant items from a digital library collection.

The first section of this report defines the general characteristics of KOSs, with emphasis on their connection to a particular view of the world. The historic origins and uses of KOSs, in libraries and in other information management environments, are described. Various types of KOSs are discussed.

Section 2 provides examples of how knowledge organization systems can be used to enhance digital libraries in a variety of disci-

plines. It describes how a KOS can be used to link a digital resource to related material. The KOS can be used directly or indirectly to provide more descriptive records for entities in the digital resource. Finally, the KOS can provide access not only to a descriptive record but also to location information about a relevant physical object.

Section 3 discusses how KOSs can be used to provide disparate communities with access to digital library resources. This can be done by using a KOS to provide alternate subject access, to add a new mode of access to the digital library (for example, visual or geographic in addition to textual), to provide multilingual access, or to support free-text searching.

The report concludes with a discussion of issues to consider when using KOSs with digital libraries. It provides a framework for the design, planning, implementation, and maintenance of KOSs in digital library environments.

Knowledge Organization Systems: An Overview

1

The term *knowledge organization systems* is intended to encompass all types of schemes for organizing information and promoting knowledge management[1]. Knowledge organization systems include classification schemes that organize materials at a general level (such as books on a shelf), subject headings that provide more detailed access, and authority files that control variant versions of key information (such as geographic names and personal names). They also include less-traditional schemes, such as semantic networks and ontologies. Because knowledge organization systems are mechanisms for organizing information, they are at the heart of every library, museum, and archive.

Knowledge organization systems are used to organize materials for the purpose of retrieval and to manage a collection. A KOS serves as a bridge between the user's information need and the material in the collection. With it, the user should be able to identify an object of interest without prior knowledge of its existence. Whether through browsing or direct searching, whether through themes on a Web page or a site search engine, the KOS guides the user through a discovery process. In addition, KOSs allow the organizers to answer questions regarding the scope of a collection and what is needed to round it out.

All digital libraries use one or more KOS. Just as in a physical library, the KOS in a digital library provides an overview of the content of the collection and supports retrieval. The scheme may be a traditional KOS relevant to the scope of the material and the expected audience for the digital library (such as the Dewey Decimal System or the INSPEC Thesaurus), a commercially developed scheme such as the Yahoo or Excite categories, or a locally developed scheme for a corporate intranet.

The decision of what knowledge organization system to use is central to the development of any digital library. The KOS must be applicable, either automatically or by human catalogers, to the resources included in the digital library. Once the material is included in the collection, the KOS must be meaningful to its users.

This section outlines the characteristics of KOSs, describes the common types, and discusses their origins and traditional uses.

---

1. The term *knowledge organization systems* as used in this report was coined by the Networked Knowledge Organization Systems Working Group at its initial meeting at the ACM Digital Libraries '98 Conference in Pittsburgh, Pennsylvania.

Common Characteristics of Knowledge Organization Systems

It is often said that humans are inherent organizers. From an early age, children play sorting and matching games. We cope with our ever-changing world by comparing new objects or experiences with those with which we are familiar, identifying patterns and categorizing what is new into our existing frame of reference. The emphasis on developing comprehensive KOSs can be seen in the writings of our earliest philosophers, many of whom continue to influence our view of the world. For example, Aristotle's effort to categorize knowledge into groups (such as physics, politics, or psychology) is reflected in our language, our education, and our science. The original classification scheme of the Library of Congress, used between 1800 and 1814, was based on the philosophical works of Sir Francis Bacon and inherited from the English tradition. Beginning in 1814, the influence of Thomas Jefferson can be seen on the Library of Congress collection. Jefferson, who reclassified the library, reflected a more humanist philosophy (Lesk 1997).

There is no single knowledge classification scheme on which everyone agrees. Michael Lesk speculates that while a single KOS would be advantageous, it is unlikely that such a system will ever be developed. Culture may constrain the knowledge classification scheme so that what is meaningful to one culture is not necessarily meaningful to another (Lesk 1997). Therefore, we live in a world of multiple, variant ways to organize knowledge.

Despite their diversity, KOSs have the following common characteristics that are critical to their use in organizing digital libraries.

- The KOS imposes a particular view of the world on a collection and the items in it.
- The same entity can be characterized in different ways, depending on the KOS that is used.
- There must be sufficient commonality between the concept expressed in a KOS and the real-world object to which that concept refers that a knowledgeable person could apply the system with reasonable reliability. Likewise, a person seeking relevant material by using a KOS must be able to connect his or her concept with its representation in the system.

Types of Knowledge Organization Systems

A review of some typical knowledge organization systems shows their scope and applicability to a variety of digital library settings. While there are specific definitions for many of these KOSs in the computer science and information science literature, and even in standards documents, there is debate over these definitions. Terms are often used, particularly in the popular press and in the book trade, in nonstandard ways. Reflecting the scope of this practice, a recent National Information Standards Organization (NISO) work-

shop on electronic thesauri emphasized the need to improve the definitions of "terminology relating to terminology" (NISO 1999).

The descriptions given here provide an overview of possible systems for organizing digital libraries. The descriptions are based on characteristics such as structure and complexity, relationships among terms, and historical function. The list is not comprehensive; nor are the definitions of these terms contained in specific standards documents. They are grouped into three general categories: term lists, which emphasize lists of terms often with definitions; classifications and categories, which emphasize the creation of subject sets; and relationship lists, which emphasize the connections between terms and concepts.

Term Lists

Authority Files. Authority files are lists of terms that are used to control the variant names for an entity or the domain value for a particular field. Examples include names for countries, individuals, and organizations. Nonpreferred terms may be linked to the preferred versions. This type of KOS generally does not include a deep organization or complex structure. The presentation may be alphabetical or organized by a shallow classification scheme. A limited hierarchy may be applied to allow for simple navigation, particularly when the authority file is being accessed manually or is extremely large. Examples of authority files include the Library of Congress Name Authority File and the Getty Geographic Authority File.

Glossaries. A glossary is a list of terms, usually with definitions. The terms may be from a specific subject field or from a particular work. The terms are defined within a specific environment and rarely include variant meanings. Examples include the Environmental Protection Agency (EPA) Terms of the Environment.

Dictionaries. Dictionaries are alphabetical lists of words and their definitions. Variant senses are provided where applicable. Dictionaries are more general in scope than are glossaries. They may also provide information about the origin of a word, variants (by spelling and morphology), and multiple meanings across disciplines. While a dictionary may also provide synonyms and through the definitions, related words, there is no explicit hierarchical structure or attempt to group them by concept.

Gazetteers. A gazetteer is a list of place names. Traditional gazetteers have been published as books or have appeared as indexes to atlases. Each entry may also be identified by feature type, such as river, city, or school. An example is the U.S. Code of Geographic Names. Geospatially referenced gazetteers provide coordinates for locating the place on the earth's surface. The term *gazetteer* has several other meanings, including an announcement publication such as a patent or legal gazetteer. These gazetteers are often organized using classification schemes or subject categories.

Classifications and Categories

Subject Headings. This scheme type provides a set of controlled terms to represent the subjects of items in a collection. Subject heading lists can be extensive and cover a broad range of subjects; however, the subject heading list's structure is generally very shallow, with a limited hierarchical structure. In use, subject headings tend to be coordinated, with rules for how they can be joined to provide concepts that are more specific. Examples include the Medical Subject Headings (MeSH) and the Library of Congress Subject Headings (LCSH).

Classification Schemes, Taxonomies, and Categorization Schemes. These terms are often used interchangeably. Although there may be subtle differences from example to example, these types of KOSs all provide ways to separate entities into "buckets" or broad topic levels. Some examples provide a hierarchical arrangement of numeric or alphabetic notation to represent broad topics. These types of KOSs may not follow the rules for hierarchy required in the ANSI NISO Thesaurus Standard (Z39.19) (NISO 1998), and they lack the explicit relationships presented in a thesaurus. Examples of classification schemes include the Library of Congress Classification Schedules (an open, expandable system), the Dewey Decimal Classification (a closed system of 10 numeric sections with decimal extensions), and the Universal Decimal Classification (based on Dewey but extended to include facets, or particular aspects of a topic). Subject categories are often used to group thesaurus terms in broad topic sets that lie outside the hierarchical scheme of the thesaurus. Taxonomies are increasingly being used in object-oriented design and knowledge management systems to indicate any grouping of objects based on a particular characteristic.

Relationship Lists

Thesauri. Thesauri are based on concepts and they show relationships among terms. Relationships commonly expressed in a thesaurus include hierarchy, equivalence (synonymy), and association or relatedness. These relationships are generally represented by the notation BT (broader term), NT (narrower term), SY (synonym), and RT (associative or related term). Associative relationships may be more detailed in some schemes. For example, the Unified Medical Language System (UMLS) from the National Library of Medicine has defined more than 40 relationships, many of which are associative. Preferred terms for indexing and retrieval are identified. Entry terms (or nonpreferred terms) point to the preferred terms to be used for each concept.

There are standards for the development of monolingual thesauri (NISO 1998; ISO 1986) and multilingual thesauri (ISO 1985). In these standards, the definition of a thesaurus is fairly narrow. Standard relationships are assumed, as is the identification of preferred terms, and there are rules for creating relationships among terms. The definition of a thesaurus in these standards is often at variance with schemes that are traditionally called thesauri. Many thesauri do not follow all the rules of the standard but are still generally thought

of as thesauri. Another type of thesaurus, such as the *Roget's Thesaurus* (with the addition of classification categories), represents only equivalence.

Many thesauri are large; they may include more than 50,000 terms. Most were developed for a specific discipline or a specific product or family of products. Examples include the Food and Agricultural Organization's *Aquatic Sciences and Fisheries Thesaurus* and the *National Aeronautic and Space Administration (NASA) Thesaurus* for aeronautics and aerospace-related topics.

Semantic Networks. With the advent of natural language processing, there have been significant developments in semantic networks. These KOSs structure concepts and terms not as hierarchies but as a network or a web. Concepts are thought of as nodes, and relationships branch out from them. The relationships generally go beyond the standard BT, NT, and RT. They may include specific whole-part, cause-effect, or parent-child relationships. The most noted semantic network is Princeton University's WordNet, which is now used in a variety of search engines.

Ontologies. Ontology is the newest label to be attached to some knowledge organization systems. The knowledge-management community is developing ontologies as specific concept models. They can represent complex relationships among objects, and include the rules and axioms missing from semantic networks. Ontologies that describe knowledge in a specific area are often connected with systems for data mining and knowledge management**.**

All of these examples of knowledge organization systems, which vary in complexity, structure, and function, can provide organization and increased access to digital libraries.

The Origin and Use of Knowledge Organization Systems

In the physical library, classification schemes such as Library of Congress (LC), Dewey Decimal System, and the Universal Decimal Classification reflect, among other things, the need to store a single item at a single location on a shelf. To provide multiple access points beyond the limits of a single physical location, subject headings are applied. Libraries use subject heading schemes such as LCSH, Sears, or other specialized schemes developed for specific content or specific collections. At the level of specific content, libraries have used authority files to control variant forms of personal, organizational, and geographic names.

However, KOSs can be found in settings other than libraries. An awareness of the KOSs available from alternative sources is valuable when considering the development of digital libraries for a specific audience.

Abstracting and Indexing Services

Abstracting and indexing (A&I) services developed as an outgrowth of traditional bibliographies and the explosion of journal literature. In the sciences, the development of A&I services was spurred by the post-World War I concerns about inadequate access to scientific information. In the 1950s, investment in A&I services was fueled by the Cold War and Sputnik. Abstracting and indexing services in the humanities, such as the Bibliography of the History of Art or the Modern Languages Association (MLA) Bibliography, generally took a different growth path than did their scientific and technical counterparts, but they also quickly became important resources for scholarship in the online environment. The scope of A&I services varies from broad discipline-oriented services (e.g., chemistry, architecture, biology, and physics) to narrowly defined aspects of the literature (e.g., peaceful uses of nuclear energy) and subdisciplines (e.g., aquatic sciences).

Special KOSs, such as thesauri and subject categories, were developed to support A&I services and their specific products and audiences. These organizations applied increasingly complex schemes to provide subject access to the literature in a variety of subjects. By the 1960s, A&I services were moving from the provision of print-only products to print and online services through large online vendors such as Dialog. Later, the products were distributed on CD-ROM and now, increasingly, on the Web. In many cases, the KOSs migrated from print to electronic media following the products they supported. While increased computing power, more sophisticated search engines, and more independent end-user searching have led to changes in some KOSs, most have retained their importance, even in the Web environment.

For many years, the KOSs related to A&I services were applied only by catalogers and indexers trained in using the KOS indexing for a particular product or products. The primary users of KOSs were librarians and other professional searchers. However, the proliferation of electronic data, the explosion of electronic publishing, and increasing concerns about the difficulty of locating information have led to a renewed interest in these KOSs for use not only by professionals but also by end users.

Publishers

As publishers have migrated to electronic composition systems, they have become increasingly involved in the production of A&I products. Large journal publishers such as Academic Press and Elsevier have developed their own systems to provide bibliographic records linked to the full text of documents. As the content of online journals has grown, it has become necessary to move from systems that provide browsing by table of contents and journal issue to systems that support searching by both free text and by KOS. Electronic journals have resulted in additional KOSs, particularly classification and categorization schemes. For example, Elsevier's Web site has a subject

categorization scheme to provide access to individual Web sites of its more than 2,000 titles.

Trade, Professional, and Governmental Organizations

A variety of authority files and classification schemes are used to support business and commerce. They range from the Standard Industrial Classification (SIC) code and the North American Industrial Classification System (NAICS), used in procurement and government statistics, to disease codes used to communicate patient illnesses and treatments among physicians, hospitals, and insurance companies. As more organizations develop Web sites, additional KOSs are being developed to support them.

Internal Projects

Organizations are among the most prolific creators and users of KOSs. Developers of corporate intranets and knowledge management systems have discovered hundreds of specific classification schemes, glossaries, categorization schemes, and other vocabularies in use within organizations. Many of these are geared toward specific tasks and are, therefore, very narrow both in subject scope and target audience. However, for these audiences, they can also be rich sources of information.

For example, the Department of Energy (DOE) Environmental Management Science Program (EMSP) and the Office of Scientific and Technical Information are developing a digital library to support EMSP program managers. Program managers and researchers have developed “needs categories” and “science categories” to organize the Environmental Science Network (ESN). The categories are used primarily to support the process of grant submission and award; however, the ESN also uses them to provide access to related material from within DOE and from other distributed databases from the EPA, the Department of Defense, and NASA. Vocabulary is currently being organized around these categories for use with a Web mining tool that will provide highly relevant Web resources for project managers in specific areas.

Summary

Knowledge organization systems include a variety of schemes that organize, manage, and retrieve information. They range from authority files to classification schemes, thesauri, and ontologies. Libraries and other information management organizations have developed KOSs to organize and retrieve information. In addition to their primary function, which is to provide access to materials for a specific community or audience, KOSs can perform functions that further enhance the digital library.

Linking Digital Library Resources to Related Resources

2

This section emphasizes the ability of knowledge organization systems to link digital library resources to other related resources. The basis for this linking is the identification of information within a digital resource that can be extracted and used to search and locate information within a KOS. The KOS may then be used to expand codes to more explanatory full text, to provide more descriptive records, or to link entity names to resources of physical specimens.

Expanding Codes to Full Text

Practitioners of a discipline use coding schemes to facilitate communication within that discipline. It is often helpful to connect these coding schemes to the full names for which the code stands. The examples provided here include links between databank registration codes and the biological sequence data, and between industrial codes and the full name that the code represents.

Linking Sequence Numbers to Biosequence Databanks

The lengthy biochemical and genetic sequences that molecular biologists, biotechnologists, and geneticists identify each day are kept in databanks. Several databanks have been developed, for example, to cover protein sequences, nucleotides, and cell lines. One of the largest databanks contains information on the mapping of the human genome. As molecular biologists began to discover these sequences, they reported them in scientific journals. Difficulties in composing, proofreading, and printing the text soon arose. Through an ad hoc standards process, major biomedical publishers agreed to require the inclusion of codes or databank numbers for these sequences in articles when they are published. In addition, the sequence itself must be registered in a databank before the paper can be published.

Some of the most frequently referenced databanks are listed on the Web site of the National Center for Biotechnology Information. They include GenBank and the Research Collaboratory for Structural Bioinformatics Protein Data Bank. Each sequence number is different, but all begin with a persistent code identifying the databank.

How can the link be made between the literature and the databank? Through a search profile, a text analysis program, or keyword indexing, the text can be analyzed and the sequence databank numbers identified. An active link can be embedded. The active link consists of a search strategy (possibly written as a CGI script) to locate

that sequence number in the databank where the actual sequence is stored. When the user clicks on the active link, the script is generated and launched from the user's browser. The Web-enabled database is searched, and the sequence record is returned to the user. Depending on the services provided by the databank site, the user can analyze the sequence using a number of tools provided by the databank or download the sequence for local manipulation.

This type of connection exists between the National Library of Medicine's (NLM) search service, PubMed, and GenBank at the National Center for Biotechnology Information. If a search in PubMed yields records that have GenBank numbers, the user can automatically search and display the sequence records from GenBank.

Linking Individual Industrial Codes to the Full Scheme

In business, classification schemes serve to communicate important facts about a company or product. These codes are generally controlled by a government, professional, trade, or international standards organization. They often serve as shorthand for users interested in material in a particular area of industry or a specific business sector.

Perhaps the most familiar scheme is the SIC code, which was last updated in 1987. The SIC codes have been used by the U.S. government, economists, financial markets, regulators, and procurement offices to identify manufacturing, agriculture, and service sectors of the economy. In 1997, a new scheme was approved for use within the United States. The North American Industrial Classification System was developed with Canada and Mexico as a means of providing an agreed-upon scheme for the collection, reporting, and analysis of information about the economy by sector, both within and across borders. Information about NAICS is available from the Web site of the U.S. Census Bureau (see references for address).

The digital library can provide related information by using the authority files for the coding schemes as a linked authority file. If a company or economic sector mentioned in the digital library's collection can be linked to an SIC or NAICS code, the code can be searched against the official tables of definitions maintained by the U.S. Census Bureau. These files provide definitions of the codes and place each code in the classification scheme with other economic sectors.

The digital library's content can be further enhanced by making a link between the SIC and NAICS codes. If the digital library resource has the SIC code, it can be extracted and searched against the Census Bureau's *1997 NAICS and 1987 SIC Correspondence Tables.* The table returns the corresponding code from the alternate scheme.

Linking to Descriptive Records

Linking the name of an entity, such as a personal name, organization, or location, to additional information about that entity was one of the first uses of hyperlinking. Knowledge organization systems such as dictionaries, glossaries, and classification schemes can be used to

link the entities in one resource to richer descriptions of that entity in another resource. This is particularly helpful for users who are new to a topic and in cases where the additional information can make the user's task more efficient.

The examples that follow are from three disciplines. The first example links organism names to records that not only describe the species more fully but also put it in the context of the overall classification scheme for living organisms. The second example links chemical names to descriptive records and molecular structures. In the third example, proper names are linked to the biographies for the person.

Linking Organism Names to Taxonomic Records

Genus-species names are the Latin names for organisms—e.g., plants, animals, and microorganisms. Taxonomists, who study and classify living organisms, create records for each of these organisms. Generally, these records are linked relationally to the other organisms in a hierarchy. Beyond the organism name and the information that it and its placement in the hierarchy convey, taxonomic records use other elements to describe the organism. These may include distribution patterns, the authority for naming and classification, and the date the organism was identified. Scientists base the information on specimens that are retained because they serve as the physical evidence of the description. Natural history museums, private collections, and individual scientists number, or code, the specimens in their collections. Sometimes specimens are supported by photographs or line drawings, which may be digitized.

By using a taxonomic authority file as an intermediate authority file, one can link a text or an image file containing a name or picture of an organism to additional related information. By automatically processing the text or embedding a link from the organism name in the text or from the image to the taxonomic authority record, one can extend the knowledge conveyed by the text. The text can include the descriptive and historical information in the taxonomic record and, ultimately, link to a photograph, a drawing, or appropriate video or audio segments.

Because of the ambiguity in organism names, many examples of this type are now created manually. However, depending on the extent of the files involved, the ambiguity of the Latin and common names for organisms can be overcome. An example of a taxonomic intermediate file is the Integrated Taxonomic Information System (ITIS). ITIS is a partnership of U.S., Canadian, and Mexican government agencies, private organizations, and taxonomic specialists cooperating to develop an online, scientifically credible list of biological names of North American plants and animals. It is used by many U.S. government agencies for consistent naming of plants and animals for regulatory and monitoring purposes. To link textual material in a digital library to the ITIS record, the organism name can be identified manually or automatically in the text and submitted as a query to the ITIS database. When a match is found, ITIS presents the

ITIS record, which provides essential information about the organism. The information includes synonymous names, including some common names, and an indication of the placement of the organism in the larger taxonomic classification scheme.

Linking Chemical Names to Molecular Structures

The unique identification for a chemical substance is not its name but its molecular structure. However, chemical names are commonly used in research documents, project plans, catalogs, and directories, all of which may be resources in a digital library. There are competing systems of nomenclature (i.e., that of the Chemical Abstracts Service [CAS] and of the International Union of Pure and Applied Chemistry) as well as common and commercial synonyms.

The ambiguity is resolved by providing links between the chemical names in the text and the molecular structure. This is done through a chemical registry number or code that is connected to a particular chemical name (using certain nomenclature standards) and an authority record that provides additional information about the chemical. This information includes the chemical's synonyms and some of its chemical and physical properties. Most important in today's research environment is the link from this authority file to a chemical structure file. Structure files, used with the appropriate software, graphically depict the molecular structure. This sophisticated software allows for three-dimensional visualization, rotation, and substitution of the chemical bonds.

An example of the use of the chemical registry number to link chemical names with molecular structures can be seen in the work of BIOSIS, the world's largest not-for-profit producer of biological and biomedical databases. In 1993, BIOSIS began processing its bibliographic citations (titles and keywords) to automatically identify chemical names (Hodge, Nelson, and Vleduts-Stokolov 1989). BIOSIS assigns CAS Registry Numbers (RNs) to the chemical names identified in this process. In the STN International online system, hosted in the United States by CAS, a user of BIOSIS can select one or more of the records resulting from a search and extract the RN. The extracted RN can be applied against the CAS Registry File, which contains more than 21 million substances, including organics, inorganics, biosequences, metals, and alloys. The registry file record for the chemical name, including the link to the synonyms for the chemical name and the structure file itself, can then be accessed. With special tools developed by CAS, the structure can be viewed and manipulated. It can be imported into modeling tools that allow the chemist to manipulate the structure and thereby envision new chemicals. Alternatively, the user can start with any database that contains CAS RNs and extract the resulting RNs to perform a search for complementary bibliographic records in the BIOSIS database.

Linking chemical names to structures using RNs on a large scale is neither inexpensive nor easy. There are two approaches to identifying chemical names in text. Some journal articles include the CAS RN for the major chemicals discussed. In this case, an analysis of the

text for the terms "RN," "CAS RN," and variations preceding numerics can identify RNs that can be used as a link. Alternatively, a program to identify chemical names in text, similar to that developed by BIOSIS, could be devised. Developing the identification program, as well as searching chemical databases, is costly; however, if the digital library has license agreements for chemistry databases, this type of linkage may be possible. In addition, many organizations have small chemical files of their own that may include RNs and other information of particular relevance to the organization's research. It may be possible to link to these local databases using methods that are more direct.

Linking Personal Names to Biographical Information

A common type of authority file is the personal name authority, which controls variants of personal names. For example, the Library of Congress Name Authority File (LCNAF) is used to control variant personal names for authors, editors, artists, and others. The Union List of Artist Names (ULAN), developed by the Getty Vocabulary Program, is another example. Name authorities serve as tools for catalogers and indexers. They ensure that the proper form of the name, rather than an unapproved variant, is used and bring together all works by or about the person.

A name authority file can also be used to link a bibliographic record or document containing the person's name to a variety of other related materials. If the digital library's resource has a standardized form of the name, it can be identified and searched against the authority file to locate variants. The standardized and variant forms can be joined in a search against a variety of other resources that can provide related information.

For example, in the case of a digital library of images of artists' works or biographical or critical text, a name authority file such as the ULAN or the LCNAF can act as an intermediate file to provide additional information. The file, which contains integrated variant names, can be searched by the name appearing in the digital library collection. When the record is found, the information about the artist can be displayed, providing a wide range of contextual material for the user. Citations to significant biographical or critical works about the artist, some of which may also be available on the Web, may also be provided in the name authority file.

The variant names from a name authority can also be used to locate and provide automatic links from the personal name in the text to a biography, without requiring that the name be presented in the same fashion in the two resources. One such resource that could be linked to for biographical information is Gale's Biography Resource, which contains more than 142,000 biographies and related citations from more than 1,000 periodicals.

However, to produce this kind of link, there must be a mechanism for locating personal names in text. Several programs can do this type of text analysis; among those that have been developed commercially are NameFinder from the Carnegie Group and the In-

telligent Agent from IBM. In addition, variant names can be extracted from the name authority itself, grouped, and run as a search against the text to locate name occurrences.

Linking Entity Names to Physical Specimens

In some cases, it is possible to go another step and connect entity names in the digital library resources to physical specimens. The curation of physical specimens or artifacts is critical to the advancement of many disciplines. Exhibition catalogs describe the art objects in a particular exhibition. Museum catalogs provide inventories of the art, natural history, or cultural objects held by a particular museum. These catalogs, increasingly available as computerized databases, are knowledge organization systems that not only provide descriptive records but also point to the location of the object in a museum, an archive, or another collection.

For example, in biology, a physical specimen is particularly important when it is the result of the discovery and description of a new organism or of the reclassification of a known organism. A type specimen is the example collected from the field by a taxonomist to serve as the prime example for the description of the organism and the validation of its taxonomic classification and naming. These specimens are held by natural history collections, and their deposit is required by the rules of various taxonomic societies.

As part of the curatorial activity, the collections assign identification codes. While the primary use of identification codes has been to organize the physical collections, numerous projects are under way in the natural history community to digitize photographs of specimens and create database records for the specimens, including their identifiers, and thereby make them more readily accessible. The degree of digitization varies from specialty to specialty. For example, in botany, virtually all significant research herbaria are digitally cataloging their type collections instead of maintaining paper records. Many are also making digital photographs of the type specimens available over the Web.

The publication of identification codes in the journal literature is also changing. Historically, identification codes have been presented in the “Materials Used” sections of journal articles. The level of specificity of the identification code has varied, depending on the biological discipline. For example, botanical journals tend to list only the institution and the catalog, while vertebrate journals provide the code to the specimen level. The current trend is to require lists of specimens that are more detailed. As the lists become longer and the printing costs increase, journal publishers are beginning to request links to independent Web sites maintained by the researchers or their organizations that carry all the specimens used in the study and provide some level of identification.

If the digital library collection contains resources that include the identification codes, these codes can be extracted and matched

against the Web-based catalogs or databases. This link can provide users with location and contact information to allow them to access the physical object mentioned in the digital library resource.

Curators or registrars of artistic, archaeological, and cultural history collections also assign inventory or accession numbers to items in their collections. Identification numbers may also be found in scholarly catalogues raisonnés. Links similar to those described for natural history can be made between text related to works of art and the physical work in a particular collection. An article about a work of art can be linked to additional information about the physical specimen by linking the identification number in the text with an online catalog containing the number and additional information about the work.

As museums digitize their collections to establish a presence on the Web or to reduce the handling of the physical objects, KOSs that can link the digital library resources to the physical object are being developed. If there is a museum with a collection that complements that of the digital library, it is worthwhile to discuss ways in which the digital library and digital museum collections may "co-evolve."

Summary

Digital libraries can use KOSs to link digital resources to other digital resources or, indirectly, to physical objects. A simple example is the expansion of codes and acronyms. Descriptive records may also be provided either directly from the KOS or indirectly by using the KOS to capture a search key that can be used to access another resource. This concept may be taken a step further by using a KOS, such as a museum or exhibition catalog, to provide information about the location of the physical object.

Making Resources Accessible to Other Communities

3

Someone recently compared the Web with a large room filled with books that were scattered all over the floor. The Web is the world's largest mass of bits and bytes. It is a meeting place that brings together disparate communities. The "Internet Commons," as this meeting place has been called, requires connections between and among disparate communities in order for an "economy" to develop (Weibel 1999). This economy will provide the framework within which both commercial and noncommercial transactions can occur. KOSs are one means of connecting these disparate communities. Knowledge organization systems can be used to (1) provide alternate subject access, (2) add modes of understanding to digital library resources, (3) support multilingual access, and (4) supply terms for expansion of free-text searches in domains that are relatively unknown to the user.

Providing Alternate Subject Access

Alternate subject access refers to the provision of one or more additional subject orientations that make the resources of the digital library accessible to different audiences. This approach is particularly valuable when the digital library resources appeal to groups that do not share a common terminology. It can be a system of subject headings, a classification scheme, or any other subject-oriented system. Alternate subject access can be provided by

- indexing or classifying the resources using multiple schemes,
- retaining original schemes from organizations that contribute to the digital library, or
- mapping between the primary scheme and an alternate scheme.

Indexing the Material with Multiple Schemes

The most direct method for providing alternate subject access to a collection is by classifying or indexing the resources with multiple schemes, but it may also be the most costly. This approach requires redundant cataloging or catalogers who are knowledgeable in both schemes. It may also require modifications to the cataloging tools and procedures. However, if the cataloging is at a high level (resources versus individual documents), or if the schemes are not difficult or detailed, it may be a reasonable approach.

Retaining Alternate Indexing from Contributors

If the digital library is being built through contributions from a variety of sources, the originating organization may have applied an alternate scheme that could be used. For example, the NASA database on aeronautics and astronautics receives relevant bibliographic records from other U.S. agencies, such as the Department of Defense and the Department of Energy. The controlled vocabulary terms assigned by the contributing organization are processed through a machine-aided indexing process to create candidate indexing terms from the *NASA Thesaurus* for review by NASA's indexers. However, the final records contain both the *NASA Thesaurus* terms and the controlled vocabulary terms from the contributing organization, with the alternate indexing terms retained in a separate data element in the bibliographic record. The terms collected from other organizations can be viewed as an alternate access point, so that at least part of the collection is accessible through another discipline's terminology.

Mapping Multiple Schemes

The third method for providing alternate subject access is the most indirect, that of mapping one or more schemes. Several examples of this approach can be found among A&I services. Both BIOSIS, the world's largest private sector A&I service in the life sciences, and the NLM apply MeSH to BIOSIS documents. The records that BIOSIS contributes to NLM's TOXLINE database are processed automatically to have appropriate MeSH terms added. This is based on a mapping of the natural language terms that occur in the toxicology literature and BIOSIS' normalized natural language keyword indexing with the MeSH terminology. In the new BIOSIS relational indexing structure, BIOSIS builds and maintains authority files that connect natural language disease names to the MeSH-controlled disease terms. When the BIOSIS indexer assigns the free text keyword for the disease name, the appropriate MeSH term is also added to the record as an alternate access point (BIOSIS 1999). The assignment is based on the development over time of a mapping between the terminology used by BIOSIS and the MeSH-controlled terms.

In addition to providing alternate access points to BIOSIS products, the inclusion of the MeSH terms makes it possible to perform cross database searching on the indexing field with MEDLINE and other databases that include MeSH terms. From 1999 forward, users can search BIOSIS databases using MeSH disease terms. The disease terms can be extracted from the MeSH authority file or from a MEDLINE record and then used in a search against the BIOSIS files, or vice versa. This helps users find relevant records that are unique to either BIOSIS or MEDLINE. The inclusion of terms from an alternate KOS, such as MeSH, therefore supports the use of BIOSIS by medical librarians and practitioners who are familiar with MeSH terminology.

A more extensive example of mapping variant schemes is the metathesaurus developed by the NLM's Unified Medical Language System (UMLS). This system has linked more than 40 separate KOSs from various medical specialties. They range from MeSH to coding

and classification schemes used by insurance companies and physicians to describe treatments and diseases on patient records. The UMLS is licensed by many other organizations for inclusion in applications that can bridge various health care communities.

How can digital libraries use alternate indexing? While many digital libraries do not have the A&I resources of large database producers such as NLM and BIOSIS, the concept of applying alternate indexing can be scaled to fit. While the systems described deal with item-level bibliographic records, alternate indexing can be applied at several levels. Alternate subject access can be applied only at the resource level, for the database, electronic book, electronic journal, or image collection, so that other communities can identify resources of interest that must then by searched or browsed individually. This concept is conducive to use with portals that provide access to the same resources with different views for different audiences. Alternatively, if the digital library has bibliographic records or metadata records at a very detailed level, it may be possible to develop switching programs that will translate concepts from the original organization of the digital library or resource to that of the alternate scheme.

Adding New Modes of Understanding to the Digital Library

People perceive the world through many modes, including textual and graphical. Some people comprehend information more easily in one mode than another. Most people benefit from a variety of modes that reinforce one another or that can be used when appropriate to the context. Many digital library projects remain text-based; however, this text-only dimension is changing as digital libraries become oriented more to multimedia and as other modes of information presentation become viable on the Internet.

KOSs can be used to bring new dimensions to an information resource or a collection in a digital library. In the digital library environment, these dimensions can be viewed as layers that can be added on top of one or more objects. Various tools and services can be developed that are geared to a particular mode. For example, the results of a text search can be presented in graphical or visual form, based on the number of occurrences of a term or concept or on the occurrences of documents from a particular country, journal title, or author.

A more complex dimension that can be added is the geospatial dimension, which emphasizes access by place. A “geolibrary” is defined as a digital library consisting of “geoinformation,” or material that can be accessed by place (National Research Council 1999). This so-called georeferencing can be either direct (by a geospatial footprint, a series of latitudes and longitudes for the location) or indirect (by a textual place name). Georeferencing of textual objects is facilitated by a gazetteer, which brings together the place name and the spatial footprint for its location.[2] Many gazetteers also include feature types for each footprint. The vocabulary used for the feature

types varies among gazetteers, but may include terms such as "airport," "harbor," and "railroad station."

Although many organizations, including federal and state agencies, are currently required to provide geospatial referencing as part of the National Spatial Data Infrastructure Program, the geospatial referencing is not readily available for older works. How can the data sets of today be integrated with the textual information of yesterday? The answer is by adding geospatial referencing to the text resource. Geospatial referencing requires that the text name for a place have an associated spatial footprint. This can be achieved by using a georeferenced, digital gazetteer that provides geospatial footprints for place names.

Through this type of knowledge organization system, place names in a library catalog or bibliographic database can have footprints assigned (Blair 1999; Tahirkheli 1999). If one or more of the library's resources have latitude or longitude coordinates in the catalog record or in the full text but no place name, the coordinates can be extracted and submitted to the gazetteer service. The service will return the place name for the footprint. Alternatively, the resource may have a textual place name. This place name can be extracted and searched against the gazetteer, and the footprint can be provided to a mapping application. The latter search may result in more than one footprint, since place names may be ambiguous. Therefore, it is important that the user interface be designed to allow the user to distinguish the locations. Once the footprint has been determined, a user can access the text resource through a geographic mapping tool. Alternatively, a user of the text resource can find a set of results and have the place names displayed as footprints on a map.

In disciplines such as ecology, environmental science, and even public health and epidemiology, it would be beneficial to build a digital library with access to such a digital gazetteer service. Users could then access the system through the text mode or the geographic

---

2. A recent National Science Foundation-sponsored workshop, "Digital Gazetteer Information Exchange," addressed the issues of digital gazetteers. One of the critical issues is that there is no standard for the interchange of information, either to provide gazetteer information physically to another gazetteer or to interoperate with one or more distributed gazetteers through the Internet. The workshop participants emphasized the need for such protocols and for enhancements to current gazetteers. (Many gazetteers do not include coordinates or are incomplete in this regard.) The goal is to develop a digital gazetteer service that can be accessed by any application.

Such a service is central to the vision of a geolibrary. A report on distributed geolibraries from the National Research Council (1999) envisions the geolibrary as a physical globe. One would walk into such a geolibrary and be confronted not by a card catalog or an OPAC terminal but by a large physical globe. The user would indicate his or her area of interest by pointing to a place on the globe. The librarian would use the geospatial location information to retrieve and present materials related to that place. By comparing feature types, the user could ask for other place names and locations that were similar to the original.

Significant work into digital gazetteer services and geospatial libraries has been conducted by the Alexandria Digital Library (ADL) Project at the University of California at Santa Barbara, with support from the National Science Foundation's Digital Library Initiative-1 (Hill and Zheng 1999). An ADL Gazetteer was created by merging place name authority files from the National Image Mapping Agency and the U.S. Board on Geographic Names of the U.S. Geological Survey. The project also added controlled feature types to the gazetteer. With the aid of a visualization tool, the information can be provided on a map and accessed using other geographic visualization tools.

mode, depending on their comfort level and the type of information needed. Presenting the results on a map allows users to make new associations and analyze the results more easily. Through a geospatial KOS, they can see connections between disparate data, because the data are presented in an alternate mode.

Providing Multilingual Access

A third way that KOSs can support the use of digital libraries by disparate communities is to provide multilingual access. A variety of sources, including multilingual dictionaries and multilingual thesauri, can support this type of access.

One of the most extensive multilingual thesaurus efforts is the Generalized Multilingual Environmental Thesaurus (GEMET) from the European Environment Agency (EEA), produced by Italy's research council, the Consiglio Nazionale delle Ricerche (CNR). The GEMET is available in 12 languages, and plans for a global environmental thesaurus in many more languages were recently announced. GEMET is available by agreement with the EEA.

The European Topic Centre on Catalogue of Data Sources in Germany is developing a system that will link data sources and metadata information in a virtual library. GEMET will be used to convert a search in one language into searches for the same concepts in other languages. Users will retrieve documents not only in their native language but also in other languages. This will allow data systems from throughout the EEA and beyond to be accessed as a virtual library collection with both controlled vocabulary and free-text term searching in multiple languages.

Expanding Free-Text Search Terms

Free-text searching is the main method of searching on the Web. Only a small percentage of Web resources have metadata, and an even smaller percentage have controlled vocabulary assigned. However, variations in natural language make free-text searching problematic. Even a knowledgeable user may not know all the terminology (synonyms or related terms) that can be used in the literature to express a concept. The problem is exacerbated when the user is unfamiliar with the topic or is interested in an interdisciplinary area. How can the user expand his or her search to overcome these terminology differences? One possibility is to use KOSs as aids to the selection of free-text keywords.

The Getty Vocabulary Project emphasizes support for searching as a significant application of its vocabularies. Harpring (1999) reports that the vocabularies are increasingly being used in search engines to look for different terms that refer to the same concept. The Getty vocabularies (the Art and Architecture Thesaurus, the Union List of Artists Names, and the Thesaurus of Geographic Names) are particularly rich in equivalence relationships. "When these equivalence relationships are exploited in search engines, there are typically

two possible scenarios: the user may be allowed to first query the vocabulary database, locating appropriate terms, and then applying those chosen terms in a query across target databases; or there may be little or no user interaction with the vocabulary, when the vocabularies are used behind the scenes [to expand the search] . . . " (Harpring 1999). Getty developed a prototype called *a.k.a.* to experiment with the use of equivalence terms to broaden or narrow searches across databases on the Web.

In addition to expanding routine search queries, KOSs can be used in Web mining tools. Northern Light has developed a Web mining tool that reportedly returns a high degree of relevant hits. The KOS that supports the Northern Light site was built by ingesting large existing vocabularies and thesauri. The result was then organized under an extensive classification scheme developed by Northern Light. The terms can be used to extend a user's search or to distinguish between multiple meanings of the terms supplied by the user. The results of a search are organized into "folders" based on the classification scheme. These high-level categories, represented by the folders, help distinguish multiple meanings of the same term. For example, an ambiguous word such as "pitcher" might result in two folders being presented to the user. One folder would be titled "Sports" (as in baseball pitcher), the second "Decorative Arts" (as in water pitcher). The user who chooses only the Sports folder will be presented with only those Web resources that use "pitcher" in the baseball sense. The user who selects the folder called "Decorative Arts" will be presented only with those resources that are related to water pitchers.

KOSs can be very powerful in supporting free-text searching within digital libraries and in integrating Web resources into existing digital libraries. However, these systems must be used with caution. KOSs have generally been developed for a specific discipline, task, or function, or for the indexing of a specific collection or database. Therefore, depending on the domain in which the KOS is being used and the complexity of the system, it may or may not suggest relevant free-text terms. Expanding a search with related terms, rather than pure synonyms, may return hits that are only peripherally relevant to the user.

Summary

One of the benefits of the Internet, the Web, and digital libraries is the degree to which resources can be made available to broader audiences. The technology facilitates the connection of disparate knowledge communities at the network level. However, discovery of the resources and true accessibility require that the content and its organization be understood by these disparate communities. By providing alternate subject access, adding modes of understanding, supporting multilingual access, and supplying terms for expanding free-text searching, KOSs can facilitate discovery and understanding by disparate communities, and allow these communities to interact in new ways.

Planning and Implementing Knowledge Organization Systems in Digital Libraries

4

This section provides general guidelines that may be useful for an organization that wants to use knowledge organization systems to organize a digital library. The framework described is applicable for KOSs of any type or subject.

Planning Knowledge Organization Systems

Analyzing User Needs

Of primary importance to any digital library project is an analysis of its users' needs, in terms of content and functionality. Many volumes have already been written about needs assessment, and providing detailed guidance on this subject is beyond the scope of this paper. However, when analyzing how a KOS might be used with a particular digital library, it is essential to thoroughly understand the environment of the user. One must look not only at the needs for organizing the digital library materials but also at possible links between content within and outside the digital library walls. This is particularly important for KOSs that are acting as intermediate authority files, because in such cases the links may not be readily apparent. It is important to consider other views that might be valuable for users and peripheral communities that might benefit from the digital library's content were it accessible to them through a KOS.

Locating Knowledge Organization Systems

Once the user's needs have been analyzed, it is necessary to locate KOSs to meet the need. While an alternate system can be built locally, it is preferable to find an existing KOS for several reasons. First, it is costly and time-consuming to build a KOS. Second, KOSs often benefit by having been built over time. Many of the systems described in this report have been built over decades; some existed in paper before digitization. The value of a KOS comes from its acceptance by the user community; sources built by noted authorities such as learned societies, trade associations, or standards groups will be viewed as more trustworthy than those built internally. Finally, the networked environment has resulted in both an explosion of primary materials, including documents, electronic journals, and Web-based databases, and in an equivalent explosion of KOSs on the Web.

There are several ways to identify KOSs that may be of interest. Many users are already aware of KOSs on the Web within their disci-

pline. Developers may also turn to directories, librarians in the field, and reference sources, or they may perform a general search of the Internet.

Planning the Infrastructure

It is necessary to make decisions about the architecture of the KOS in the context of the digital library setting. The physical location of the KOS is important. Will the system be held externally or internally? There are pros and cons to either approach.

If the system is available on the Web, it is possible to consider linking to the KOS as an external system. This architecture requires a script or some search query to locate the resource. One must then launch a query against the resource to obtain the piece of information that will serve as the key between the two files. This key could be a universal resource locator (URL) or input to another search query. A query may be necessary if the KOS is stored in a database. The script may transfer log-on information (including user ID and password) from the digital library system to the external KOS, in order to provide access to the Web-enabled database. In the case of a more direct link, the access may be by URL.

However, the use of a URL as the link has the same problem with persistence as does direct access via a URL from a browser. The organization may move the KOS, thereby changing the URL that is being used as the key. It is important to determine how often the URLs in the KOS change, whether there is a means of notification of these changes, and whether it is possible to consider an alternative that would be more persistent. Schemes such as the Digital Object Identifier and the Persistent URL have been devised to enable resources to be physically moved among servers without having their names changed. Another alternative is the use of other Uniform Resource Identification (URI) schemes and the Uniform Resource Name (URN), which can be sent from the newer Web browsers. The benefit of linking to a remote resource is that the resource will always be up-to-date. The maintenance of the KOS is in the hands of the owner, not the digital librarian. It may also be more apparent to users that the KOS is not owned by the digital library.

Linking to a remote KOS also has disadvantages. Persistence and unexpected changes in the organization and content of the system may cause problems. The software or telecommunications route between the digital library server and the KOS may be unreliable. In systems requiring fast response time or large amounts of data transfer, and, therefore, high bandwidth (such as full-motion video or detailed graphics), the fact that a connection must be made between the digital library and the external KOS may make the system unacceptable to the user.

Alternatively, the KOS may be obtained from the owner and loaded locally. In many cases, this requires licensing that may not be required when the KOS is accessed remotely, because a copy of the whole resource is being provided to the digital library. Loading a KOS locally also requires that one consider issues such as mainte-

nance, local system administration, and disk storage. If the KOS uses special software, such as a database management system, loading the KOS locally will require a copy of that software, which may require additional purchase or licensing. Other considerations are the need for firewalls and interface design. On the positive side, the KOS is under more local control. Therefore, it may be possible to improve the response time by not accessing the KOS over the Internet. If the KOS is to be used behind the scenes (that is, the system is not visible to the user), concerns of speed and integration become more important. If additional modifications (including digitization) need to be made to the KOS to integrate it with the digital library, it will also be necessary to load the KOS locally.

If the digital library intends to incorporate numerous secondary KOSs, it is important to consider the degree to which the architecture is scaleable. The National Library of Medicine's UMLS incorporates more than 40 different sources. While its main purpose has been to develop a metathesaurus for moving among these vocabularies, the management of the systems, regardless of the mapping issues, has been a major consideration. Ingest has been a major concern, with the need to develop a system that can handle a variety of input formats—from ASCII text files to highly structured database output. The architecture must also accommodate the character sets of the incoming sources. This is particularly important if a mark-up language has been used to represent special characters and diacritical marks. Systems that have been developed in Unicode, which extends ASCII to accommodate diacritical marks and non-Roman character sets, cannot be handled by systems that deal only with ASCII or extended ASCII sets.

Since many digital library systems are being built as extensions or applications of existing integrated library systems (ILS), it is important to consider how the KOSs will integrate with the library system. Unfortunately, many ILS vendors have not considered links to external files or databases in their system designs. In some cases, the vendor may require that the information be stored in the proprietary format of the ILS. The system may require that the files be on the same directory or server as the accessing ILS. The fields that can be linked to the Web or searched may be limited. Outside communications may require Z39.50 client–server connections. With relatively closed systems, ILSs may be a difficult environment in which to implement alternative and nontraditional KOSs.

Digital libraries that are interested in using KOSs should consider this integration when developing requirements for the procurement of a system to support them. Vendors should be encouraged to support relatively open architectures and to consider the extension of traditional library systems to support broader digital library functionality.

In addition to these immediate concerns, it is important to consider the incorporation of future KOSs. Initial success may spur the desire for integration of additional KOSs or enhanced functionality

for the existing KOS. Success may breed additional requirements and increase the strain on hardware, software, and network architectures.

Maintaining the Knowledge Organization System

For a digital library, an outdated KOS can be more of a hindrance than a benefit. Maintenance, both of content and of the system, should be considered when planning a KOS. This is particularly important if the digital library is to be self-supporting or revenue generating.

Version control of the KOS is extremely important. Reloading a new version from the system provider is one way to accommodate changes; however, this may not be acceptable if the locally held version differs substantially from that held by the system's provider. If there has been significant transformation or processing of the original KOS, it may be difficult, or impossible, to reload the original and recreate the changes that have been made.

A transaction-based approach, whereby only changes are transferred between the KOS provider and the library, is also possible; however, this requires that the system provider have the infrastructure, both machine and human, to produce these transactions. It also requires that the changes to the original KOS be identifiable in order to create change transactions. For example, Stuart Nelson of the NLM's UMLS Project recently reported that many systems can create annual transaction records to inform the UMLS about the changes that have occurred to the original system. However, the changes are often not indicated with enough detail to support automatic change transactions in the UMLS. If a change date, for example, is recorded only at the level of the concept record, it is impossible to tell whether the term has changed (a correction of a typographic error for example) or if the relationship between this concept and another concept has changed. Since the UMLS splits the incoming terminology and its relationships into a variety of files, it is often difficult to tell how the UMLS files must be change based on the changes made during the maintenance of the original KOS (NISO 1999).

Presenting the Knowledge Organization System to the User

In addition to deciding which KOS should be used and what functions it should serve, the digital library will need to determine how to present the KOS to its users. A KOS may be exposed to the user or made relatively transparent.

The KOS can be exposed to the user in different ways. Material can be grouped into KOS-related themes or categories on the digital library's Web site. The KOS may be used at a higher level to identify specific portals for different uses or users. If the content of the digital library includes metadata records, the KOS may be displayed as index terms on the records or in its entirety as a navigation aid to searching.

In other cases, the KOS may be transparent. For example, a thesaurus can be used behind the scenes to extend the user's search to

include synonyms, to connect the digital library's resources to other information and resources, or to filter or rank the information obtained.

Implementing Knowledge Organization Systems

Acquisition and Intellectual Property Issues

It is critical to properly handle the acquisition of knowledge organization systems. The first question is whether the KOS is under copyright. If so, the copyright holder should be contacted concerning the KOS. It is important to ensure that the apparent contact is the official one. Many references have been reprinted or put on the Web without proper acknowledgment of the real owner.

Once the contact has been made, there are several points for discussion:

- If the provider maintains the KOS, how will the digital library find out about any changes that may be made in it? Is there a notification mechanism in place? How frequently must the information be updated to be of benefit to the digital library's users? Will the maintenance be self-evident, or must the agreement include notification requirements? What will the owner do if the maintenance can no longer be performed?

- What will happen if the provider discontinues the product or sells or transfers it to someone else?

- What uses can the digital library make of the KOS under the proposed agreement? As with other licensing, it is advisable to aim for the broadest permissions and the longest term possible. At a minimum, the library should be able to renegotiate the terms of the agreement relatively easily.

- In a networked environment, it is beneficial to develop mechanisms for linking to online versions rather than to maintain a local copy of the resource. This ensures that what is presented is up-to-date, and acknowledges more clearly the ownership of the KOS. However, there are numerous factors to consider. Will the KOS be used on an intranet or behind a firewall, where access to the outside or information coming into the organization might be prohibited? Does the KOS service use "cookies" or require knowledge of the user's Internet provider address? Does it require a user ID and password?

- If the KOS is to be accessed remotely, are there service issues? Is it likely to be accessed with bandwidth, model, and computer speeds that are adequate for outside connections of this type? Is the use of such a critical nature that unreliable service on the part of the KOS or the Internet connection will cause the digital library itself to be viewed as less useful? Does the KOS require a special-

ized search engine or search query formulation? Can the digital library system properly display the results, or would the results be better displayed through the KOS system? Will the resulting information be used in its native form or must it be extracted or transformed? If the KOS is to be loaded locally, in what formats can the content be received?

- If the KOS is not available electronically, can it be digitized? Is the owner interested in a cooperative venture, and are the human and financial resources for such an effort available?

Making the Link

There are two parts to establishing the link between the digital library and the KOS. The first is locating the key anchor information in the digital library's resource. The second involves the look up against the target file. The creation of this link may be more or less automatic, depending on the particular situation. The characterization of this activity is meant to be general and to allow both "on-the-fly" links and embedded links.

Regardless of what function the KOS is going to serve in the digital library, the essential information contained in the digital library resource from which the link is to be made must be identified. The mechanism for doing this depends on the type of object from which the link is being made and on the information that is expected to be identified in the digital library's resource.

The first step is to review any metadata related to the digital library resource. Do the metadata carry the term (such as SIC code, artist's name, place name, geographic coordinates) that is needed to make the link? If this information is included, the level at which the metadata are assigned should be reviewed. If the metadata indicate the subject matter of the specific resource in which the user will be interested, the metadata can be used to make the links. However, in some cases, the terms that appear in the title or description at the resource level (e.g., the book) may not be indicative of the subject at the individual item level (e.g., the chapter). Automatically making a link on the basis of the content description for an entire book may misrepresent the content of a chapter. Whether or not the metadata can be used will depend on the amount and type of information given in the metadata and the level at which the metadata are assigned.

If a text resource in the digital library provides no appropriate metadata, the procedure for identifying the key information may involve text analysis. A program to perform simple string searching or a search engine that can preserve hit locations can be used if the text string has distinguishing characteristics, such as a database acronym, or a specific structure, such as a latitude and longitude coordinate. If the text string has no such cues, text mining or more complex text-analysis tools may be necessary. These tools use a variety of semantic and syntactic algorithms to locate key information. There have been significant advances in commercially available text-mining tools,

such as IBM's Intelligent Agent, which includes specific algorithms for identification of names of places and persons.

The second step of the linking activity is to make the connection to the KOS. The methods for doing this vary, depending on whether the system is being loaded locally or is referenced remotely. If the system is loaded locally, it is possible to perform a significant amount of processing to match the two files, assuming that computer resources of this type are available to the digital library organization. If the system is only available remotely over the Web, the interaction will require knowledge of scripting and various Web-based access techniques. Scripting should be considered in both local and remote approaches, since the more integrated the linking is with the resource, the more maintenance may be required if there are changes in either the resource or the KOS. Regardless of the approach that is taken, making the link requires an analysis of both the information in the original digital library material and the corresponding information in the KOS.

If the KOS is being used as an intermediate file to bridge between the digital library's resource and another resource, it is also important to understand the data and the process whereby the search is performed and information returned from the target resource. If the KOS must return a value to the original digital library resource, the data and process must be evaluated in a bidirectional sense.

Choosing the linking mechanism is equally important. The link may be fixed or "on-the-fly." In the case of a fixed link, a specific URL is embedded at the link point in the digital library material. However, as stated before, problems of persistence are inherent in this approach. Alternatively, a URN can be used. The URN requires the creation of a namespace on the point of the target file, and the search is to this namespace rather than to a specific URL. Persistent locators (PURLs) and digital object identifiers (DOIs) can also solve this problem. These schemes are sufficient if the material is an HTML document.

Content in databases is more difficult to retrieve. The National Library of Medicine now supports the searching of a variety of its databases through its Internet Grateful Med (IGM) URL function. IGM users can create URLs that will actually perform searches against the databases. For example, the following script would perform a search for "pneumonia" in the HealthSTAR file: http://igm-02.nlm.nih.gov/cgi-bin/IGM_robot.pl?datafile=HealthSTAR&search=Subject=pneumonia.

Information on the syntax for creating such a URL is provided on the NLM Web site. While the intent is that the search URL will be bookmarked by an individual user, the same concept can be used for creating an active link at the anchor point for the link. With additional scripting, the creation of the term *pneumonia* can be automatically replaced with an active link that picks up the term where the link has been made.

Summary

The framework for developing an infrastructure to support the use of KOSs in digital libraries requires an analysis of user needs, the identification and location of the appropriate KOSs, and the development of the hardware, software, and network architecture to support its integration and maintenance. The digital librarian must make decisions concerning the degree to which they will be presented to the user, acquisition and intellectual property issues, and maintenance and update procedures. There are several technical ways to make the link between the digital library and the KOS. As knowledge organization systems are increasingly available on the Web, requirements are beginning to be defined to improve the interoperability and general use of these resources through the development of knowledge organization services on the Web.

The Future of Knowledge Organization Systems on the Web

5

As online databases moved to the Web, they began to provide their products, including vocabulary aids, in this environment. Portable document format (PDF) versions of printed vocabulary aids are common, since PDF can be easily produced from a Postscript file and it retains the look of the printed product. With Adobe's tools for indexing and searching, the PDF file can provide some level of support for linking. Many of these aids, however, remain in the form of HTML files only—there is no database structure to easily support the linking and searching. In some cases, the full structure of the KOS is not made available on the Web; the only format for a Web-based thesaurus may be an alphabetical list of terms that does not enable the user to navigate easily the hierarchical structure. As unique ways of using these resources are developed, it is hoped that more KOS providers will be encouraged to provide their systems in formats that are conducive to such networked uses.

Some of the requirements for such electronic KOSs were identified at a workshop entitled "Electronic Thesauri: Planning for a Standard" and sponsored by NISO (1999). While the focus of this meeting was digital thesauri, consideration was also given to other KOSs in digital form. The identified requirements include persistent identification at the concept level, the need for a simple protocol for the distributed querying and response from a KOS, and the development of a standard set of metadata attributes for describing a remote KOS.

To facilitate the search and display of information from a previously unknown KOS, the system must have unique and persistent identifiers for each of the concepts in the system. For example, the California Environmental Resources Evaluation System (of the California Natural Resources Agency) and the U.S. Geological Survey have developed a system for remote querying and response (CERES 1999). It requires that each concept in the thesaurus have a unique identifier. In the case of the previously described ITIS, which is accessed remotely by the CERES system, the ITIS record number is used as the identifier. Other unique identifiers could include the DOI, or a classification notation that has been made unique by appending the scheme name or the URL to the notation.

The second requirement is a protocol for the distributed querying and response of KOSs. This is particularly critical for highly structured systems such as thesauri, semantic networks, and ontologies. Work has been done in this area within the Z39.50 community.

(Z39.50 is the NISO standard for searching distributed bibliographic databases.) A profile has been proposed by the Zthes Working Group to tailor the Z39.50 protocol to operate on thesauri that follow the Z39.19 standard.

A similar effort is under way at the CERES Project. Instead of a Z39.50-based protocol, CERES has developed a structure that is based on the Resource Description Framework (RDF) and the HTTP protocol of standard browsers. The RDF's concept of containers is a natural for managing the hierarchical structure of complex systems such as thesauri. The structure proposed by CERES is likely to be encoded using XML, a mark-up format that lends itself to structured information. This protocol for linking distributed vocabularies will support both searching and cataloging. The user will be presented with remote vocabularies that can be displayed and navigated by a local client.

The third major finding from the NISO workshop was the need for a metadata content standard for the description of KOSs. Such a standard is key to provision of knowledge organization services over the Internet. The metadata identify the Web resource as a KOS and provide important information to allow an application to use it remotely without prior knowledge of its content or structure.

A draft set of attributes for describing KOSs available in a networked environment has been developed by a task group of the Network Knowledge Organization Systems (NKOS) Working Group, an ad hoc group of terminology experts from organizations that are interested in issues related to the use and interoperability of KOSs over the Internet. The draft attributes are based on work originally done by Linda Hill (Alexandria Digital Library at the University of California at Santa Barbara) and Michael Raugh (Interconnect Technologies).

The attributes describe the KOS so that content from the system can be transferred over the Internet and handled by a remote browser or client application. The attributes include the depth of hierarchy, the types of relationships included, the subject (described by free text or by a declared classification scheme), storage format, copyright and rights management, and contact information. To facilitate the transfer of information, the attribute set also includes information on character set and file size. To facilitate the acquisition and licensing of the KOSs, the draft content description includes point of contact information.

During discussions about the metadata content standard, workshop attendees identified three methods for storing the metadata for a KOS. First, the metadata could be stored with the KOS, as metadata elements for that resource. Second, the metadata could be stored in a physically separate knowledge organization registry. The third possibility is a hybrid approach, where a minimal set of metadata elements is contained in a central registry (i.e., sufficient information to identify the resource, where it is located, and how more information can be obtained). The more detailed information would be stored with the KOS itself.

There is significant interest in the use of KOSs to organize and search material on the Internet. It is hoped that this interest will result in knowledge organization services that will make these sources more readily accessible to a variety of software applications and to a variety of users. As services and enabled software proliferate, it will be easier to integrate these KOSs into digital libraries.

Conclusion: Enhancing Digital Libraries with Knowledge Organization Systems

6

Given that the digital library field is still quite new, it seems strange to be talking already about enhancing digital libraries. However, in this fast-moving environment, the initial digital libraries resulting from digitization projects, or even virtual collections, are being enhanced as user expectations and technology capabilities allow. In the midst of this furious activity, it is valuable to analyze users' needs and interests and then to identify KOSs that can be used to enhance the digital library.

Knowledge organization systems refer to a range of traditional and nontraditional systems for the organization of knowledge. The systems have been developed in numerous environments outside the traditional library environment, including those of A&I services, publishers and professional organizations, and corporations. Examples exist in many disciplines and for many target audiences.

Knowledge organization systems can enhance the digital library in a number of ways. They can be used to connect a digital library resource to a related resource. The related information may reside within the KOS itself or the KOS may be used as an intermediary file to retrieve the key needed to access it in another resource. A KOS can make digital library materials accessible to disparate communities. This may be done by providing alternate subject access, by adding access by different modes, by providing multilingual access, and by using the KOS to support free text searching.

A well-planned infrastructure for KOSs is required. This includes the resources, processes, and policies for analyzing user needs; locating KOSs to answer these needs; and acquiring, implementing, and maintaining the KOS.

Traditional and nontraditional KOSs provide an opportunity to extend the boundaries of the digital library. By going beyond the initial organization of the digital library, digital librarians can use the network environment to provide additional value to its users.

References

The Web site addresses listed in this section were valid as of April 1, 2000.

BIOSIS. 1999. Now, Cross-File Searching with CAS Registry Numbers and MeSH Disease Terms. *BIOSIS Evolutions: Corporate News and Product Information from BIOSIS* 6(3):1.

Blair, Nancy. Digital Gazetteers: The Traditional Library Perspective. Paper presented at the National Science Foundation Digital Gazetteer Information Exchange Workshop, Washington, D.C. October 13–14, 1999. Available from http://alexandria.sdc.ucsb.edu/gazetteer/dgie/DGIE_website/session3/blair.htm.

CERES Thesaurus. 1999. California Environmental Resources Evaluation System. Available from www.ceres.ca.gov/thesaurus/.

Harpring, Patricia. 1999. Resistance Is Futile: Inescapable Networked Information Made Accessible Using the Getty Vocabularies. In *Knowledge: Creation, Organization and Use. Proceedings of the 62nd Annual Meeting of the American Society for Information Science* (October 31–November 4, 1999). Washington, D.C.

Hill, Linda, and Qi Zheng. 1999. Indirect Geospatial Referencing through Place Names in the Digital Library: Alexandria Digital Library Experience with Developing and Implementing Gazetteers. In *Knowledge: Creation, Organization and Use. Proceedings of the 62nd Annual Meeting of the American Society for Information Science* (October 31–November 4, 1999). Washington, D.C.

Hodge, Gail, Tom Nelson, and Natasha Vleduts-Stokolov. 1989. Automatic Recognition of Chemical Names in Natural Language Text. Paper presented at the 198th American Chemical Society National Meeting, Dallas, TX, April 7–9, 1989.

International Standards Organization. 1986. ISO 2788. Documentation—Guidelines for the Establishment and Development of Monolingual Thesauri.

International Standards Organization. 1985. ISO 5964. Guidelines for the Development of Multilingual Thesauri.

Lesk, Michael. 1997. *Practical Digital Libraries: Books, Bytes, and Bucks*. San Francisco: Morgan Kaufmann Publishers.

National Research Council, Panel on Distributed Geolibraries. 1999. *Distributed Geolibraries: Spatial Information Resources: Summary of a Workshop*. Washington, D.C.: National Academy Press.

National Information Standards Organization (NISO). 1999. NISO/ASI/ALCTS Workshop on Electronic Thesauri: Planning for a Standard, Washington, D.C. November 4–5, 1999. Available from http://www.niso.org/thesau99.html.

National Information Standards Organization (NISO). 1998. ANSI/NISO Z39.19. Guidelines for the Construction, Format and Management of Monolingual Thesauri.

Tahirkheli, Sharon. 1999. Place Names in an Earth Science Literature Index. Paper presented at the National Science Foundation Digital Gazetteer Information Exchange Workshop, Washington, D.C. October 13–14, 1999. Available from http://alexandria.sdc.ucsb.edu/gazetteer/dgie/DGIE_website/session3/Tahirkheli.htm.

Weibel, Stuart. 1999. Dublin Core and the Metadata Landscape: Conventions for Semantics, Syntax, and Structure in the Internet Commons. Paper presented at the 41st National Federation of Abstracting and Information Services (NFAIS) Annual Conference, Philadelphia, PA, February 21–24, 1999.

Web sites noted in this report

Chemical Abstracts Service (CAS) Registry File: http://info.cas.org/casdb.html.

Digital Gazetteer Information Exchange: http://alexandria.ucsb.edu/ gazetteer/dgie/DGIE_website/DGIE_homepage.htm.

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

Elsevier: http://www.sciencedirect.com.

European Topic Centre on Catalogue of Data Sources: http://www.mu.niedersachsen.de/cds/etc-cds_neu/information.html.

Gale Biography Resource Center: http://www.galegroup.com/pdf/facts/brc.pdf.

Generalized Multilingual Environmental Thesaurus (GEMET): http://www.mu.niedersachsen.de/cds/etc-cds_neu/software.html#GEMET.

Integrated Taxonomic Information System (ITIS): http://www.itis.usda.gov/plantproj/itis/index.html.

National Center for Biotechnology Information GenBank: http://www.ncbi.nlm.nih.gov/Genbank/index.html.

National Library of Medicine, Internet Grateful Med: http://igm.nlm.nih.gov/splash/IGM_url.html.

National Library of Medicine, Medical Subject Headings (MeSH): http://www.nlm.nih.gov/mesh/meshhome.html.

National Library of Medicine, PubMed: http://www.ncbi.nlm.nih.gov/PubMed/.

National Library of Medicine TOXLINE: http://www.nlm.nih.gov/pubs/factsheets/toxlinfs.html.

National Library of Medicine, Unified Medical Language System (UMLS): http://www.nlm.nih.gov/research/umls/.

Networked Knowledge Organization Systems/Services (NKOS): http://www.alexandria.ucsb.edu/~lhill/nkos.

Networked Knowledge Organization Systems (NKOS) Thesaurus Registry Working Group: http://www.alexandria.ucsb.edu/~lhill/nkos/Thesaurus_Registry.html.

Persistent URL: http://www.purl.org.

Research Collaboratory for Structural Bioinformatics Protein Data Bank: http://www.nist.gov/srd/nist80.htm.

Resource Description Framework (RDF): http://ceres.ca.gov/thesaurus/.

Union List of Artist Names (ULAN): http://shiva.pub.getty.edu/ulan_browser/ulan_intro.html.

U.S. Census Bureau Web, North American Industrial Classification System (NAICS): http://www.census.gov/epcd/www/naics.html.

Zthes Working Group: http://lcweb.loc.gov/z3950/agency/profiles/zthes-03.html.