Hill, L. (2002). Integration of Knowledge Organization Systems into Digital Library Architectures: Position Paper for 13th ASIST SIGICR Workshop, Reconceptualizing Classification Research. 13th ASIS SIG/CR Classification Research Workshop, 46-52. doi:10.7152/acro.v13i1.13835

Integration of Knowledge Organization Systems into Digital Library Architectures: Position Paper for 13th ASIS&T SIG/CR Workshop, "Reconceptualizing Classification Research"

Linda Hill, Olha Buchel, Greg Janée
Alexandria Digital Library Project
University of California, Santa Barbara
lhill@alexandria.ucsb.edu, obuchel@alexandria.ucsb.edu, gjanee@alexandria.ucsb.edu

Marcia Lei Zeng
School of Library and Information Sciences
Kent State University, Kent, Ohio
mzeng@kent.edu

1. INTRODUCTION

Digital library (DL) research and development has concentrated primarily on *collections* and on the *services* to build and access them (Arms, 2000). To some extent, there has also been a focus on *users and uses* and on how well digital library constructions satisfy them (Borgman, 2000). A class of DL components has been missing from this development. This class we call knowledge organization system (KOS) resources and by this we mean the set of familiar and evolving systems that organize and define the terminology and notations we use to represent and organize concepts and real world objects. Just as DL *collections* and *services* can be modeled in general frameworks that support the building of library architectures, it is also possible to integrate the variety of KOSs into the DL context and to extend DL architectures to include them in the development and use of collections and services. Treating KOS resources as tightly integrated components of DL architectures raises new research and development issues for the DL and the classification research communities.

2. KNOWLEDGE ORGANIZATION SYSTEMS IN DIGITAL LIBRARIES

The set of KOSs includes the following (modified from Networked Knowledge Organization Systems/Services Group & Hodge, 2000).

- Classification and Categorization
  - Categorization Schemes: loosely, any grouping scheme
  - Classification Schemes: hierarchical and faceted arrangements of numeric or alphabetic notation to represent broad topics.
  - Subject Headings: schemes that provide a set of controlled terms to represent the subjects of items in a collection and sets of rules for combining terms into compound headings.
  - Taxonomies: divisions of items into ordered groups or categories based on particular characteristics

ISSN: 2324-9773

- Metadata-like Models
  - Directories: lists of names and their associated contact information
  - Gazetteers: geospatial dictionaries of named and typed places, where relationships between places are represented inherently through geospatial representations as well as through explicitly stated relationships such as "IsPartOf" (Hill, 2000); the scheme is extendable to the representation of events (e.g., hurricanes) and named time periods where the geospatial representations become time ranges.
- Relationship Models
  - Ontologies (Concept Spaces): specific concept models representing complex relationships between objects, including the rules and axioms missing from semantic networks.
  - Semantic Networks: sets of terms representing concepts, modeled as the nodes in a network of variable relationship types.
  - Thesauri: sets of terms representing concepts and the hierarchical, equivalence, and associative relationships among them. Thesaurus structures of this type are based on NISO (National Information Standards Organization (U.S.), 1994) and ISO (International Organization for Standardization (ISO), 1986) standards. Another type of thesaurus (e.g., Roget's Thesaurus) represents only the equivalence (synonymy) of terms, with the addition of classification categories.
- Term Lists
  - Authority Files: lists of terms that are used to control the variant names for an entity or the domain value for a particular field.
  - Dictionaries: alphabetical lists of terms and their definitions that provide variant senses for each term, where applicable.
  - Glossaries: alphabetical lists of terms, usually with definitions.

In a DL environment, this KOS class can be distinguished by its common elements and by a set of common functionalities. In general, all of these schemes feature *labels* (including terms and notations), *their meanings*, and *their relationships*. The functions they support are:

- description: controlled set of labels for describing an object;
- definition: meanings associated with labels;
- translation: mappings between equivalent representations; and
- navigation: links within an organized structure of representation.

In addition, all of the members of the KOS class represent a *point of view*. They model a domain of knowledge and they are often designed for a special purpose. Therefore, each has an overall structure, scope and purpose, the understanding of which is necessary for the interpretation of the contents.

In contrast, DL *collections* can be distinguished as "groups of objects" represented by various forms of item-level metadata, ranging from *ad hoc* collections to formally and institutionally managed collections (Hill, Dolin, Frew, Janée, & Larsgaard, 1999). In DLs, collections, in general, support the functions of:

- selected content: selected subset of available objects;
- organization: application of consistent ordering principles;
- documentation: contextual, inherent, and administrative metadata; and

- archiving: long-term stewardship.

Collection development is driven by scope and purpose, but, in contrast to KOSs, collections are fundamentally open-ended with the intention of continuing the process of collection. Understanding the scope and purpose of a collection provides guidance on what to expect to find there. It is useful to consider that a KOS expands by *infill*, while a collection expands by *acquisition*. That is, KOSs add entries within their sets of records while collections add additional resources.

DL *services* interact with collections, KOSs, and users (both machines and human). They tend to be modular and designed for special purposes. They are designed to work through networks and to be compatible with networking standards for machine-to-machine communication. In general, DL *services* support the functions of:

- acquisition and cataloging: collection building, metadata creation, and maintenance;
- search and retrieval: distributed query and response, query enhancement, access methods; and
- analysis and evaluation (including visualization).

3. ALEXANDRIA DIGITAL LIBRARY PROJECT

The Alexandria Digital Library (ADL) Project at the University of California, Santa Barbara focuses on the design and implementation of distributed georeferenced digital libraries and has been involved with the design and building of collections, services, and KOSs since the beginning of NSF DL funding in 1994 ("ADL Homepage", 2002). Included in this, in addition to actual collection building, has been

- collection-level metadata structure (Hill et al., 1999)
- metadata for representing computer models (Hill, Crosier, Smith, & Goodchild, 2001)
- gazetteer design and implementation ("ADL Gazetteer Development Page")
    - Gazetteer Content Standard (see "ADL Gazetteer Development Page")
    - Gazetteer Service Protocol (Janée & Hill)
    - Textual-Geospatial Integration (Frew & Smith, 2001)
- thesaurus development
    - ADL Feature Type Thesaurus (Hill, 2002)
    - Object Type Thesaurus ("ADL Object Type Thesaurus")
    - ADL Thesaurus Service Protocol (Janée, Ikeda, & Hill, 2002)
- search bucket architecture for unified search across dissimilar collections (Janée & Frew, 2002)
- concept space design and development in support of science education (Smith, Zeng, & ADEPT Knowledge Team, 2002)
- visualization of geospatial objects and of concept spaces (Ancona & Smith, 2002)

Gazetteers have always been central to ADL. At one time, the gazetteer was considered to be a collection and treated just like other collections of aerial photographs, remote sensing images, maps, and other georeferenced documents. But the function of the gazetteer is special in the ADL architecture. It answers questions like "Where is Bakersfield?" It provides a translation function in the processing of a query such as "What remote sensing images does the library have

ISSN: 2324-9773

covering Bakersfield?" where the placename "Bakersfield" needs to be translated into longitude and latitude coordinates in order to locate remote sensing images covering the area. Gazetteer data can be superimposed on geospatial images and maps to identify and label the features and provide the context necessary to evaluate geospatial data. Gazetteers support the geoparsing of text documents, wherein the geographic areas that documents are about can be represented with coordinates, thus making them objects suitable for a geospatial digital library. So, the ADL Gazetteer became a class of its own in the ADL architecture.

The ADL Gazetteer Service Protocol was developed to provide general programmatic access to gazetteers of various structures. It supports the searching of gazetteers by the principal attributes of geographic places (names, footprints, types, and relationships) and the return of reports in a *standard* structure. All that is required for its use is the implementation of a gazetteer server that can accept the specified XML queries and return the specified standard reports.

Thesauri and authority files are used in several ways in the ADL architecture. Gazetteer entries are grouped by (classed with) terms from the ADL Feature Type Thesaurus. ADL search buckets specify that hierarchical sets of terms be used for *object types* and *format* descriptive elements for collections. The intent is that these terminologies be used as descriptive content in object metadata and gazetteer entries and that the KOS structures assist users in finding appropriate information and navigating within the collections. To support these uses, the ADL Thesaurus Service Protocol was developed as a general protocol to programmatically access online thesauri. Like the Gazetteer Service Protocol, all that is required for its use is the development of a thesaurus server that can accept the specified XML queries and return the specified standard reports.

A current research focus, as part of the ADEPT project (Smith et al., 2002), is to build a concept space model to represent the concepts in a field of science and their relationships as a primary approach to the teaching of science for undergraduate education. The concept space model extends the thesaurus model by characterizing sets of concepts to represent domains of knowledge more completely. Such a model encourages the representation of *relationships* as a separate set of components – a KOS in itself.

This same realization that relationships are separate definable components of KOSs became evident in the development of the Gazetteer Content Standard. Gazetteer entries can be related to one another by explicit statements of relationship (as well as through geospatial relationships), such as that one entity (e.g., a county) is *PartOf* another entity (e.g., a state). The set of relationships between gazetteer entities can be modeled as a thesaurus with, for example, more specific types of *PartOf* relationships.

With relationships modeled as separate components, the commonality of gazetteers, thesauri, and concept spaces becomes clear. Each represents *concepts* (nebulous entities) with labels, some of which are designated as *preferred* for convenience of reference. The terms are *defined* through associated attributes and through *relationships* with other terms in the system and, optionally, through referral to external resources.

An instructive consideration is that gazetteers occur as thesauri, the most notable example of which is the Getty's Thesaurus of Geographic Names (Getty Information Institute, 1997), and as metadata-like individual records. For thesaurus-modeled gazetteers, the explicit representation of relationships between geographic entities is embedded in the thesaurus structure (usually an administrative partitive hierarchy). The ADL Gazetteer Content Standard, on the other hand, and the gazetteers of the U.S. federal government, are modeled on the basis of individual records for each place with the relationships between places represented as attributes of those records. The thesaurus model can be easily converted to the metadata-like model by considering the relationships as attributes of entries rather than as the structural design of the thesaurus model itself.

4. IMPLICATIONS FOR DL AND CLASSIFICATION RESEARCH AGENDAS

The agendas for DL and classification research should consider the implications of this recasting of KOSs as tightly integrated components of the DL environment. Given a common core of characteristics and functionality among a variety of KOS types, it should be possible to

- Within KOS interoperability and integration into DL services
    - develop a taxonomy of KOS, specified to the point that the content and functions of KOS types can be anticipated and linked to associated DL service protocols
    - develop registries of KOS and the KOS-level metadata to represent them
    - develop XML/RDF standard representations for KOS content that can be customized for different types of KOS
    - identify divergent practices among KOS content guidelines and structures that complicate interoperability among them
    - develop a core set of relationship types that have the same meanings across all KOS
    - explore KOS integration into DL architectures and services
- Within DL services
    - develop a general KOS service protocol from which protocols for specific types of KOS can be derived
    - develop a robust linking model in which DL entities (collections, objects, and services) can refer to KOS entities (concepts, labels, and relationships) in ways that support referential integrity, versioning, and synonym mapping
    - develop visualization tools that fully use and display the rich semantics embedded in KOS

5. CONCLUSION

Collections, KOSs, and services need to work together in DL architectures. KOSs play a part in collection building, discovery and searching, navigation, evaluation, and visualization. A formal and consistent set of definitions for KOS types, methods for identifying, locating, and referring to individual KOS resources, and protocols for their use will integrate these valuable resources into the overall DL environment. The KOS resources preferred by different communities will be accessible outside of that community for the increasing necessity of cross-domain access to information. The existence of free-standing and accessible KOS resources will counter the tendency to build such systems into particular metadata standards and service protocols.

ACKNOWLEGEMENTS

This position paper has benefited from the comments of the ADL implementation and knowledge organization teams, in particular Michael Freeston, James Frew, and Terence Smith.

REFERENCES

Alexandria Digital Library Project. *ADL Gazetteer Development Page*. Retrieved 7/8/2002 from www.alexandria.ucsb.edu/gazetteer.

Alexandria Digital Library Project. *ADL Homepage*. Retrieved 7/8/2002 from http://www.alexandria.ucsb.edu.

Alexandria Digital Library Project. *ADL Object Type Thesaurus*. Retrieved 7/8/2002 from http://www.sdc.ucsb.edu/~mary/objtype.tes/index.htm.

Alexandria Digital Earth Prototype. *Project Description*. Retrieved 9/7/2002 from http://www.alexandria.ucsb.edu/adept/adept.html.

Ancona, D., & Smith, T. R. (2002). *Visual Explorations for the Alexandria Digital Earth Prototype*. Second International Workshop on Visual Interfaces to Digital Libraries, at the ACM+IEEE Joint Conference on Digital Libraries, July 18th, 2002, Portland, Oregon, USA. Retrieved from http://www.ohsu.edu/jcdl/main.cgi?opt=sked-ws.

Arms, W. Y. (2000). *Digital libraries*. Cambridge, Mass.: MIT Press.

Borgman, C. L. (2000). *From Gutenberg to the global information infrastructure : access to information in the networked world*. Cambridge, Mass.: MIT Press.

Frew, J., & Smith, T. R. (2001). *Textual-Geospatial Integration Services for the National SMETE Digital Library*. University of California, Santa Barbara. Retrieved 7/8/2002 from http://www.ehr.nsf.gov/gpra/gpra_award_info.asp?AWD_ID=0121578 .

Getty Information Institute. (1997). *Thesaurus of Geographic Names*. Retrieved from http://www.ahip.getty.edu/tgn_browser/.

Hill, L. L. (2000). Core Elements of Digital Gazetteers: Placenames, Categories, and Footprints. In Borbinha, J. & Baker, T. (Eds.), *Research and Advanced Technology for Digital Libraries : Proceedings of the 4th European Conference, ECDL 2000 Lisbon, Portugal, September 18-20, 2000.* (pp. 280-290). Berlin: Springer.

Hill, L. L. (2002). *Feature Type Thesaurus*. Alexandria Digital Library Project. Retrieved 9/6/2002 from http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/FTT_metadata.htm .

Hill, L. L., Crosier, S. J., Smith, T. R., & Goodchild, M. F. (2001). A content standard for computational models. *D-Lib Magazine, 7*(6) (June).

Hill, L. L., Dolin, R., Frew, J., Janée, G., & Larsgaard, M. (1999). Collection Metadata Solutions for Digital Library Applications. *Journal of the American Society for Information Science (JASIS). Special Topic Issue on Metadata, 50*(13), 1169-1181.

Hodge, G. (2000). *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*: Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/reports/pub91/contents.html.

ISSN: 2324-9773

International Organization for Standardization (ISO). (1986). *Documentation - Guidelines for the establishment and development of monolingual thesauri* (ISO 2788:1986). Retrieved from http://www.nlc-bnc.ca/iso/tc46sc9/standard/2788e.htm.

Janée, G., & Frew, J. (2002). *The ADEPT Digital Library Architecture*. ACM+IEEE Joint Conference on Digital Libraries, July 18th, 2002, Portland, Oregon, USA. Retrieved from http://alexandria.sdc.ucsb.edu/~gjanee/archive/2002/jcdl-adept.doc.

Janée, G., & Hill, L. L. *Gazetteer Service Protocol*. Alexandria Digital Library Project. Retrieved 7/8/2002 from www.alexandria.ucsb.edu/gazetteer/protocol .

Janée, G., Ikeda, S., & Hill, L. L. (2002). *The ADL Thesaurus Protocol*. Alexandria Digital Library Project. Retrieved 7/8/2002 from http://www.alexandria.ucsb.edu/thesaurus/protocol/.

National Information Standards Organization (U.S.). (1994). *Guidelines for the Construction, Format, and Management of Monolingual Thesauri* (Z39.19-1993). Bethesda, MD: NISO Press.

Networked Knowledge Organization Systems/Services Group. *Webpage*. Retrieved 7/8/2002 from http://nkos.slis.kent.edu/.

Smith, T. R., Zeng, M. L., & ADEPT Knowledge Team. (2002). *Structured models of scientific concepts for organizing, accessing, and using learning materials*. Paper presented at the Seventh International ISKO Conference. 10-13 July 2002, Granada, Spain.

ISSN: 2324-9773