2003

From subject gateways to portals: The role of metadata in accessing international research

Susan J. Heron
University of South Florida, sheron@usf.edu

Ardis Hanson
University of South Florida, hanson@usf.edu

Recommended Citation

Heron, S. J & Hanson, A. (2003). From subject gateways to portals: The role of metadata in accessing international research. In N. Callaos (Ed.) Conference proceedings of the SCI 2003: The 7th world multiconference on systemics, cybernetics, and informatics (pp. 529-533). Orlando, Florida: International Institute of Informatics and Systemics.

From Subject Gateways to Portals: the Role of Metadata in Accessing International Research

Susan Heron
Library System, University of South Florida
Tampa, Florida, United States of America

Ardis Hanson
The Louis de la Parte Florida Mental Health Institute, University of South Florida
Tampa, Florida, United States of America

OVERVIEW

Traditionally, academicians and researchers have turned to librarians to assist in the organisation and, ultimately, the retrieval of information. Since the beginning of scholarly activity libraries have organized information based upon the scholar's area of study. Subject collections comprised of materials in various formats have drawn scholars to the library. Today libraries organize myriad online resources to assist selected user communities, again based upon areas of study through the development of subject gateways or portals. However, a critical issue remains how best to broker access to heterogeneous information and learning resources.

ORGANISING KNOWLEDGE

Knowledge is organised according to some sort of data structure or framework. According to the American Heritage® Dictionary of the English Language [1], a framework is "a set of assumptions, concepts, values, and practices that constitutes a way of viewing reality." From this conceptual beginning, data structures are developed to provide access to content. Two common terms that will be used extensively in this paper are metadata and MARC. "Metadata is data that describes the content, data definition and structural representation, extent, …quality, availability, status, and administration of a … dataset" [2]. MARC (MAchine Readable Cataloging) is a communications protocol developed in the 1960s by the Library of Congress for representing bibliographic records in a computer stable form [3]. Both of these formats create a 'meta-vocabulary" for descriptive elements and/or content of an item.

There are two major threads in the conversation on "meta-vocabulary". The first thread deals with a standard bibliographic description. In librarianship, the International Standard Bibliographic Description (ISBD) is a bibliographic format developed to easily translate data across physical borders and machine environments. Endorsed by the International Federation of Library Associations (IFLA), the ISBD uses standard set of punctuation (periods, commas, semicolons, colons, dashes, spaces, and slashes) to delineate the fields (primary elements) of a cataloging record, in any language or character set. ISBD transcends international boundaries, allows records to be incorporated into catalogs of other countries, and finally, allows records to be converted into machine-readable form with a minimum of effort.

The second thread deals with the creation of a standard record format that accepts data into pre-defined fields, is governed by a set of rules defining what information goes where, and last, but not least, is extensible. Since this view of bibliographic data requires standards for data entry and configuration, the criteria used allow for easy retrieval and precision and relevance in one's recall. Why is this significant? The importance of frameworks in knowledge organisation and management can best be explained by the concept of 'cognitive miserliness', the tendency of the human mind to expend the least effort in acquiring information. Coined by social psychologists Fiske and Taylor [4], the term 'cognitive miser' describes an individual's interest in conserving energy and reducing cognitive load, i.e., sifting through the mass of information that bombards us everyday, ignoring anything unimportant to us, and retaining the information that is important.

The use of metadata, MARC, or any other knowledge organisational tool is based upon some form of Cutter's principles of organisation. Cutter's *Objects* were to 1) enable a person to find a book of which either the author, title, or subject is known; 2) show what the library has by a given author, on a given subject, or in a given kind of literature, and 3) assist in the choice of a book, as to its edition (bibliographically) or to its character (literary or topical). His *Means*, or method of doing so, provides numerous access points, including author-entry with necessary references; title-entry or title-reference; subject-entry, cross-references, and classed subject-table; form-entry; edition and imprint, with notes when necessary [5].

Cutter's Means were created based on the principles of how individuals search for information, using his own observations. His experiences in 1904 are echoed in a 1998 report on the functional requirements for the bibliographic record (FRBR) entity relationship model for works, expressions, manifestations, and items. However, on major difference between Cutter's experience and the libraries of the 21st century is that today's literature clearly demonstrates that library patrons see libraries more as remote resources, rather than as walk-in facilities.

The IFLA Study Group on the Functional Requirements for Bibliographic Records [6] reviews the applications users may make use of bibliographic records, for example: "to determine what information resources exist, perhaps on a particular subject or by a particular person, within a given "universe" (e.g., within the totality of available information resources, within the published output of a particular country, within the holdings of a particular library or group of libraries, etc.); to verify the existence and/or availability of a particular document for

purposes of acquiring, borrowing or lending; to identify a source or sources from which a document can be obtained and the terms under which it is available; to determine whether a record already exists for an item being added to a collection or whether a new record needs to be created; to track an item as it moves through a process such as binding or conservation treatment; to determine whether an item can be circulated or sent out on interlibrary loan; to select a document or group of documents that will serve the information needs of the user; or to determine the physical requirements for use of an item as they relate either to the abilities of the user or to special requirements for playback equipment, computing capabilities, etc. (p. 8).

Based upon this analysis of user needs, the IFLA Study Group determined that the generic information tasks users perform are:

1. "using the data to find materials that correspond to the user's stated search criteria (e.g., in the context of a search for all documents on a given subject, or a search for a recording issued under a particular title);
2. using the data retrieved to identify an entity (e.g., to confirm that the document, described in a record corresponds to the document sought by the user, or to distinguish between two texts or recordings that have the same title);
3. using the data to select an entity that is appropriate to the user's needs (e.g., to select a text in a language the user understands, or to choose a version of a computer program that is compatible with the hardware and operating system available to the user);
4. using the data in order to acquire or obtain access to the entity described (e.g., to place a purchase order for a publication, to submit a request for the loan of a copy of a book in a library's collection, or to access online an electronic document stored on a remote computer)." (p. 9).

These principles are still the foundation of best cataloging practice, including the notion of specificity, the consideration of the user as the principal basis for subject-heading decisions, the practice of standardizing terminology, the use of cross-references to show preferred terms and hierarchical relationships, and solving the problem of the order of elements. They organize the information in such a way that allows the user to eliminate irrelevancies or false cognates and to focus on specifics, thereby reducing cognitive overload.

The library approach -- to use experts to impose order on the available materials to facilitate precise retrieval -- is not the approach taken by most web search engines. The latter rely on keyword access augmented by a voluntary contribution from site authors. The result is predictable: web searches retrieve thousands of possible "hits", many irrelevant, which the user has the task of winnowing. For researchers this is more than simply frustrating: they want an exhaustive list of relevant material or a short list of only the most focused items.

To bring the terminology of the 19th century into the 21st century, replace Cutter's "book" with "resources", prefix it with any number of adjectives (e.g., print, digital), and filter it through the lens of the user of today's network user.

NETWORK RESOURCES

Before the Internet, an individual might look to a common reference tool, such as the *Yellow Pages*, an encyclopedia, or the local library catalog to find an answer. Increasingly, however, individuals looked to search engines and other services to bring specific resources and services worldwide into one shared space. The access protocol (ftp, Gopher, HTTP) defined these resource "spaces" and each resource space developed associated search services (Archie for searching ftpspace, Veronica for searching Gopherspace, Mosaic for HTTP).

Today, these "resource spaces" are defined as gateways and portals. Koch [7] provides a useful typology of gateway services, which delineates some of the finer distinctions between different types of services, including: quality criteria and quality control; the extent of metadata provided, and by whom; the intended scope (subject, geographical, language). Wegner [8] describes the importance of identifying core content in subject areas and creating an environment of comprehensive facilities for accessing research papers: literature information databases, projects for "one-stop shopping" sites, freely available digital content of classical publications, and access to electronic grey literature.

The term "subject gateway", or "subject-based information gateways", emerged during the early to mid 1990s, particularly in the research, educational, or cultural domains. It described "a network resource discovery service which provides database(s) of Internet resource descriptions with a specific subject focus and created according to specific selection and quality criteria" [9]. Wells, Calcari, and Koplow [10] provide an excellent review of twelve early projects (eight U.S. and four British) that organized information using a library information management model on the Internet.

An example of the evolution of subject gateways into portals is EEVL (*Enhanced and Evaluated Virtual Library)*, an award-winning free service, which provides quick and reliable access to the best engineering, mathematics, and computing information available on the Internet. Materials are selected, cataloged, classified and subject-indexed by experts to ensure that only current, high-quality and useful resources are included, such as those from e-journals, databases, training materials, professional societies, university and college departments, research projects, bibliographic databases, software, information services and recruitment agencies. Newer services include access to complementary databases, Web indexes of sites included in the gateway, news, and current awareness.

Portals are the newest resource. According to the literature [11-18], effective portals accomplish any or all of the following three goals:

1. establish procedures for creating web portals that link the expertise of interdisciplinary researchers,
2. establish a procedure for digital libraries to exchange and share documents, queries, and services among digital collections as well as within a single digital collection, and
3. address the different levels of interoperability.

Interoperability ranges from defining document and query types, managing documents and items contained or described in the portal, establishing intellectual property rights, and providing as comprehensive and as international collection pertinent to the subject area.

We postulate that portals are most successful when they use a library information management model. Portals should guide the user to his or her right answer through the use of effective metadata, guided queries, and human-factor architecture that will provide targeted, online content with increasing dependability and convenience to users of every skill level through standards, dynamic content linking tools, semantic web engines, and standardized, customizable user interfaces.

SEARCHING ONLINE

Butler [11] reported that seventy percent of Web users typically type in only one keyword or search term. The implications from this study are that Web users seldom bother with more than one keyword, are unable to think of an appropriate second keyword, or have not yet mastered the art of the Boolean operator. Bergman [13] states that a quality search result is "not a long list of hits, but the right list." Further, he states "effective searches should both identify the relevant information desired and present it in order of potential relevance -- quality. Sometimes what is most important is comprehensive discovery -- everything referring to a commercial product, for instance. Other times the user requires the most authoritative answer, for example, the complete description of a chemical compound. The searches may be the same for the two sets of requirements, but the answers will have to be different."

Another area that will be equally important to consider is the capability to use seamless languages by the reference provider and the library patron [19]. Search languages will need to ensure consistency, accuracy, precision, and negotiation power between the remote parties as well as to accommodate whatever communication languages will be needed for disadvantaged users if the Library of Congress' CDRS becomes the standard for 24/7 international e-reference [20-21]. This becomes even more important as librarians and other information professionals across national boundaries will be relying upon their library-based bibliographic systems as well as commercial and general Internet reference tools to provide reference and research assistance to their patrons.

For example, simply allowing the user to search through all of the words in a site, while inexpensive and easy, is ultimately crude and inadequate for the goal of optimal retrieval. An understanding of the context of the information should be built into a site, expressing its relationship to the field of knowledge is essential, especially in a multicultural setting. Also, many electronic resources misuse the term "index", when they really mean "concordance". A concordance is a list of all occurrences of a word in a site or resource (minus stop words). An index, on the other hand, lists concepts and handles the problem of synonymous terms by collocating under a preferred term, much as the expansion of Cutter's *Means* provide the framework.

By creating these indexes based on the conceptual components of a resource, vendors and developers are essentially creating taxonomies. According to Gilchrist [22] "A taxonomy aspires to be: a correlation of the different functional languages used by the enterprise ... to support a mechanism for navigating, and gaining access to, the intellectual enterprise ... by providing such tools as portal navigation aids, authority for tagging documents and other information objects, support for search engines, and knowledge maps ... and possibly ... a knowledge base in its own right."

A recent corporate survey asked corporations to determine how they were handling the retrieval of electronic information. The questions included the value of taxonomies, what processes were used to build those taxonomies, the optimal mix of machine/human interaction in generating taxonomies [22]. However, three pertinent questions that are of particular interest to this discussion are 1) how roles differ between producers and users of taxonomies, 2) how taxonomies should be developed to represent more than just documents (people, artifacts, etc.), and 3) how to deal with multiple cultures and languages.
These questions are not unique to the business community. However, we feel that workable solutions can be found within the theoretical and applied aspects of cataloging and classification within the library world.

HANDLING GLOBALLY-BASED INFORMATION: A LIBRARY PERSPECTIVE

With the increased ease of access to networked resources, the roles of national bibliographic databases as combined metadata repositories and knowledge management systems will also play a part in the globalisation of information. How can we expand current frameworks to handle emerging information resources to allow efficient information (and cognitive) processing?

As our environment becomes more complex and more international, the need to handle information in an appropriate, efficient, and verifiable manner has grown. The International Federation of Libraries [6] wonders how catalogers will guarantee the quality and relevance of bibliographic access within the exploding world of online materials. If so, what kind of bibliographic records will be required to meet the different uses and user needs? Finally, how should these bibliographic data be organized and structured for intellectual and physical access to the documents? Let us address record requirements from the perspective of the user framed within MARC and AACR2r.

Display Issues

What are optimal (good enough) display elements and relationships between the different entity groups? Questions as to how well the display elements are on a page or how fully the MARC record might convey the "substance" of an item need additional consideration as librarians "push" OPAC pages to users who may or may not be conversant with the existing screen display. Quality assurance issues, such as authenticity, provenance, permanency, reliability, and validity, take on new meaning as librarians interact with remote patrons who expect a level of integrity in the material they are receiving [23].

Display of complex bibliographic information is increasingly vital as we look at fullness of records, related and associational links, and contextually related materials. These new complex records provide a level of analysis with co-citation studies, publishing clusters, active bibliography, and if the document is on an external website, similar or related documents on the same site (See figure at end of this paper). However, the more complex the record becomes and the larger the database, subject access and forms of name become more critical to collocate contextually based and synonymous information.

Subject Access, Naming Conventions, and Keywords

With the stated aim of creating an inventory of globalization resources available electronically, there will be groupings of

identified materials based upon some sort of subject access. If the assumption is that the users of this database will be searching for works by a particular author or group of authors, titles of works, and or subject areas, there will be a need to establish naming conventions for these access points. There are a number of the studies across the library science and information science literature that attest to the need for naming conventions to enhance precision and relevance in one's retrieval when searching [24-26]. In fact, the literature attests to user frustration when trying to find a relevant something and then having to sift through hits that are contextually irrelevant although their term(s) might be somewhere in the record. Thesauri that can create the hierarchical and bibliographical relationships among content and context of items are critical. Catalogers attempt to create listings of various depths and degrees of detail to record the existence of research materials. Researchers then search for answers to their questions and to make the best possible use of recorded knowledge. As Smiraglia [27] states "That is, they [researchers] seek to exploit what is already known, so as to create new knowledge."

Some of the most important aspects of these databases are the enhancements added to the base record, such as the extensive "keywords" added by database vendors. These are not truly "keywords" as traditionally defined; they are part of a controlled vocabulary that is assigned by an individual who reviews the context and content of each item and adds these words to enhance precision and relevance in one's search. So many people assume that "keyword" searching (which is really "natural language query") will retrieve all relevant or pertinent topics within a database.

There are many reasons why this is a false assumption. First, my "keyword" may not be yours . Second, if my "keyword" isn't in a relevant document, the document will not be retrieved. Third, if the concept for which I am searching is in the database does not explode or map terms to analogous (related) terms, my retrieval will be degraded. In actuality, few databases map to online thesauri. In addition, most "home-grown" online databases use the "keyword" as their base.

For example, the Foundation Center database does not use a controlled vocabulary or field delimiting to distinguish between potential grantees and those individuals, agencies, and organizations to which they do not award grants. When one searches their databases, one searches on all words within a record, which certainly explains why one retrieves grant opportunities for which one is not eligible.

Further, when looking at the differences in discipline-based terminology, establishing a controlled vocabulary for a multi-disciplinary database is not for amateurs. Creating a concordance is relatively easy compared to creating and maintaining a hierarchical, expansive thesaurus for a database that maps across terms and creates those narrower, broader, and related terms that are critical to ensure that what one is searching for is really what one wants.

Language Access

If this to be a globally defined collection, what is the primary language for searching? If English is assumed to be the primary language, then titles and subject areas will need to be enhanced with translations, particularly for transliteration for non-Latinate languages. Further, if this is to be a multi-lingual database, defining the search parameters becomes fairly critical. There will also be a need to create variant or translated titles/abstracts, etc. for non-English materials to provide access to those items and possibly a translation engine to create some sort of translation of the item (if in HTML, Word, etc.) and vice-versa.

Differing formats and hardware/software necessary to view content would require notes to inform the user that to view this data one would need X software/plug-in application(s), Y amount of space on their drive to install and run said plug-in, etc. An example would be plug-ins to display kanji or other Asian syllabary or ArcView to visually display geographic or spatial data.

User Behaviors

Developers of the back-end of the database will need to review anticipated user behaviors for the database. For complex multi-disciplinary and multi-lingual databases, fields and limiting factors need to be defined. These would include language, subject field?, format of data, software and hardware needs to access that data, etc. In addition, how do people search? Librarians have decades of experience with a wide variety of user populations searching for information.

Conclusions

This paper discusses that the use of metadata formatted in a uniform way, using thesauri and authority files, aids users in efficient retrieval. Authors' names represented in a variety of ways on his/her publications, variant spelling (e.g. colour or centre), an updated term (NEGRO to AFRICAN AMERICAN), evolving geographic entities (SOVIET UNION now RUSSIA), a foreign language phrases, and contextually ambiguous terms (false cognates) can lead to missed sites when keyword indexing is the only access. A thesaurus of subject terms enhances searching precision, eliminating false leads, which keyword-only searching promulgates, while including relevant materials which might have been missed because their titles lacked common keywords. There is a critical need for new approaches to the problem of information overload, such as may be offered by taxonomies. Considering the scale and variety of information now being provided, there is a growing demand for a wider range of search aids. According to Gilchrist [22], there was a feeling among survey respondents that an over-reliance on software solutions was dangerous. As a consequence, the survey respondents were prepared to invest significant human resources in building and maintaining classifications, thesauri and taxonomies.

Finally, the latest research on the use and construction of bibliographic records is weaving together a continued emphasis on relevance and precision in retrieval, needs of the unmediated search behavior, and international use.

References

[1] **American Heritage® Dictionary of the English Language** (2000, 4th edition).

[2] J. Smits, "Metadata: An introduction". **Cataloging & Classification Quarterly**, Vol. 27, No.304, 1999, pp.303-19.

[3] S. J. Heron & C. L. Gordon, "Cataloguing and Metadata Issues for Electronic Resources". In A.Hanson & B.L.Levin (eds.)**The Building of a Virtual Library.** Hershey, PA: IDEA Group Publishing, 2002, pp. 78-94.

[4] **S. T. Fiske & S. E. Taylor,** Social cognition **(2nd ed.). New York : McGraw-Hill, 1991.**

[5] **C. A. Cutter,** Rules for a Dictionary Catalog**, 4th edition. Washington, DC: Government Printing Office, 1904.**

[6] IFLA Study Group on the Functional Requirements for Bibliographic Records **Functional Requirements for Bibliographic Records: Final Report**: IFLA Study Group on the Functional Requirements for Bibliographic Records. Deutsche Bibliothek: Frankfurt am Main, 1998. http://www.ifla.org/VII/s13/frbr/frbr.pdf.

[7] T. Koch, Traugott "Quality-controlled subject gateways: definitions, typologies, empirical overview" **Online Information Review**, Vol. 24, No. 1, 2000, pp.24-34.

[8] B. Wegner, "EMIS 2000: The European Mathematical Information Service and its developments," **Online Information Review**, Vol. 25, No. 3, 2001, pp. 165-172.

[9] L. Dempsey, "The subject gateway: experiences and issues based on the emergence of the Resource Discovery Network," **Online Information Review**; Vol. 24 No. 1; 2000, 8-23.pp.

[10] A. T. Wells, S. Calcari, & T. Koplow, Travis. **The Amazing Internet Challenge: How Leading Projects Use Library Skills to Organize the Web**. Chicago, IL: American Library Association, 1999.

[11] D. Butler, "Souped-up search engines," **Nature**, Vol. 405, No. 6783, 2000, pp. 112-15.

[12] C. Frappaolo & H. Reynolds, "Now it's personal [Web portals]," **Intelligent Enterprise**, Vol. 3, No. 17, 2000, pp. 30-38.

[13] M. K. Bergman, "White Paper: The Deep Web: Surfacing Hidden Value," **Journal of Electronic Publishing**, Vol. 7, No. 1, 2001. http://www.press.umich.edu/jep/07-01/bergman.html.

[14] D. R. Kolzow & E. Pinero, "EDRI White Paper: The Internet economy and its impact on local economic development," **Economic Development Review**, Vol. 7, No. 3, 2001, pp. 82-99.

[15] R. Kotorov & E. Hsu. "A model for enterprise portal management," **Journal of Knowledge Management**, Vol. 5, No. 1, 2001, pp. 86-93.

[16] J. Koenemann, H.-G. Lindner, & C. G. Thomas, "Enterprise information portals from search engines to knowledge management," **NFD Information - Wissenschaft und Praxis**, Vol. 51, No. 6, 2001, pp. 325-334.

[17] B. Ainsbury, "Cataloging's comeback classifying and organizing corporate documents," **Online**, Vol. 26, 2002, pp.2-3

[18] L. Andresen, "Immediate access to Danish libraries - with bibliotek.dk," **The Electronic Library**, Vol. 20, No.3, 2002, pp. 187 - p194.

[19] Z. Ercegovac, "Collaborative E-Reference: A Research Agenda," **67th IFLA Council and General Conference**, August 16-25, 2001. http://www.ifla.org/IV/ifla67/papers/058-98e.pdf.

[20] E. G. Abels, "The e-mail reference interview," **RQ**, Vol. 35, No. 3, pp.345-358.

[21] B. Dervin & P. Dewdney, "Neutral questioning: A new approach to the reference interview," **RQ**, Vol. 25, No. Summer, 1986, pp. 506-513.

[22] A. Gilchrist, "Corporate taxonomies: report on a survey of current practice", **Online Information Review**, Vol. 25 No. 2; 2001, pp. 94-103.

[23] A. T. Wells & A. Hanson, " E-reference" In A. Hanson & B. L. Levin (eds.). **The Building of a Virtual Library**. Hershey, PA: IDEA Group Publishing, 203, pp. 95-120.

[24] E. Svenonius, "The intellectual foundation of information organization" In **Digital libraries and Electronic Publishing**, Cambridge, MA: MIT Press, 2000.

[25]A. G. Taylor, "Research and theoretical considerations in authority control," **Cataloging and Classification Quarterly**, Vol. 9, No. 3, 1988, pp. 29-56.

[26] B. Hjørland, & H. Albrechtsen, "An analysis of some trends in classification research," **Knowledge Organization**, Vol. 26, No. 3, 1999, pp. 131-139.

[27] R. Smiraglia, "The progress of theory in knowledge organization," **Library Trends**, Vol. 50, No. 3, 2002, pp. 330-349.

Figure 1. Sample display of complex record