THE UNIVERSITY OF RHODE ISLAND

University of Rhode Island
DigitalCommons@URI

Technical Services Department Faculty Publications | Technical Services

2010

2010 New England Technical Services Librarians Spring Conference: Crosswalks to the Future: Library Metadata on the Move

Andrée J. Rathemacher
University of Rhode Island Library, andree@uri.edu

Martha Rice Sanders
HELIN Consortium, martha@helininc.org

See next page for additional authors

Follow this and additional works at: http://digitalcommons.uri.edu/lib_ts_pubs

 Part of the Library and Information Science Commons

Citation/Publisher Attribution

Rathemacher, Andrée J.; Sanders, Martha Rice; and Cerbo, Michael A. II, "2010 New England Technical Services Librarians Spring Conference: Crosswalks to the Future: Library Metadata on the Move" (2010). *Technical Services Department Faculty Publications.* Paper 31.
http://digitalcommons.uri.edu/lib_ts_pubs/31

This Article is brought to you for free and open access by the Technical Services at DigitalCommons@URI. It has been accepted for inclusion in Technical Services Department Faculty Publications by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons@etal.uri.edu.

Authors
Andrée J. Rathemacher, Martha Rice Sanders, and Michael A. Cerbo II

This article is available at DigitalCommons@URI: http://digitalcommons.uri.edu/lib_ts_pubs/31

2010 New England Technical Services Librarians Spring Conference: Crosswalks to the Future: Library Metadata on the Move

Andrée J. Rathemacher, Martha Rice Sanders and Michael A. Cerbo II

This report discusses the program of the 2010 New England Technical Services Librarians (NETSL) annual spring conference, held on Thursday, April 15 at the College of the Holy Cross in Worcester, Massachusetts, entitled "Crosswalks to the Future: Library Metadata on the Move." NETSL is a section of the New England Library Association and a regional group of the American Library Association.

In the opening presentation, Barbara Tillett, an internationally known expert on bibliographic and authority standards, spoke on "Building Blocks for the Future : Making Controlled Vocabularies Available for the Semantic Web." Tillett is chief of the Policy and Standards Division of the Library of Congress and its representative on the Joint Steering Committee for Development of RDA (Resource Description and Access). She previously led IFLA in its work toward a "Statement of International Cataloguing Principles," helped to develop the FRBR conceptual model, as well as FRAD (Functional Requirements for Authority Data), its extension for authorities and has spearheaded the effort to develop the VIAF (the Virtual International Authority File).

As an example of building blocks, Tillett began with a general review of linked data using DBpedia, a community effort to extract structured information from Wikipedia and to make this information available on the Web, highlighting the participation of the National Library of Sweden in this initiative. She discussed how important it is that libraries participate in such experiments.

With this importance in mind, she explained the objectives of VIAF, which are to facilitate sharing of authority data, reduce cataloging cots, simplify authority control internationally, and provide authority data in multiple forms, languages and scripts so that many communities may help to maintain and enhance authority data without the barriers of language or script. In the 1970s, IFLA called for a unified authority file with a single heading for each person and corporate body but this ignored the language needs of its diverse users. The VIAF, hosted at OCLC, began by virtually combining the national authority files of the Library of Congress, the Deutsche Nationalbibliothek, and Bibliothèque nationale de France into a single name authority service. It now matches names across twenty authority files from sixteen institutions and contains 13,000,000 name records representing 10,000,000 personas in 4,500,000 clusters. The database in Unicode is available as linked data with URIs and supports both UNIMARC and MARC21. Work has begun on adding geographic names to the database.

Tillett discussed an associated project whereby large groups of bibliographic records are automatically mined to derive and enhance new authority records. Using multiple bibliographic fields from each record, including the author, added author, title, publication information, etc.,

an authority record is generated that may include attributes, such as the author's broad subject areas, decades of activity, and frequent co-authors. These derived authority records are contributed to the VIAF.

Because VIAF uses linked data, the authority data may be enhanced with such additions as scrollable cover art, alternate forms of each author's name in multiple scripts, maps showing countries of publication, and a timeline of the author's publication history. In addition, the user may find information about the author's nationality, as well as other personal information. She recommended that catalogers use VIAF as a reference tool to resolve conflicts, with questionable data, or forms of name. Next steps include improving searching, adding more linked data, extending participation beyond libraries, and including more name types such as corporate, family, geographic, and uniform titles. Topical terms will never be included.

Tillett next discussed the Simple Knowledge Organization System (SKOS) and how the Library of Congress uses it. SKOS "provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, such as subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabulary." SKOS is based on the Resource Description Framework (RDF), which enables software applications to exchange data and helps to create the semantic Web. Using SKOS, LC created its Authorities and Vocabularies service to enable both humans and machines to programmatically access authority data at the Library of Congress via URIs. This service began by giving access to LCSH (subject, genre/form, children's subject headings, subdivision records, and validation records) and now

links to RAMEAU headings (LCSH in French). LC plans to add the Thesaurus for Graphic Materials and MARC geographic, language, and relator codes. Freely available on the Web and intended for machine use, computer servers can download entire controlled vocabularies, as well as the values within them. In addition, humans may search and view individual headings. Eventually, because of the capabilities inherent in this database, LC plans to generate authority records which will negate the need for "pattern headings" and in the future which may make "free floating" subdivisions unnecessary.

Tillett ended her presentation with information about RDA including a review of the timetable for testing it. RDA relies on RDA-controlled vocabularies/registries. The registries list elements used in building bibliographic records, such as place of publication, dates, media and carrier types, and so on. Each element is assigned a URI that may be used in the bibliographic record instead of the text string it represents. Additionally, the words for each element are listed in multiple languages for facilitating easy translation of the bibliographic data. In conclusion, libraries are joining Google, Amazon, Yahoo and others in "the cloud" to facilitate the sharing and enhancement of our data and ensure its use in the future.

Managing Objects and Data: From Call Numbers to Namespaces

In one of the three morning breakout sessions, Mark J. Caprio (digital services and cataloging librarian, Providence College) and Martha Rice Sanders (knowledge management librarian, HELIN Consortium) explored the assertion that although the work of traditional catalogers has changed as formats have multiplied and collections have become distributed, the fundamental intellectual content of the work of bibliographic description remains the same. Although digital technologies have presented new challenges, tools, and vocabularies for the organization and description of information, those who organize and manage information still create structures, identify relationships, and learn from user-collection interactions.

Caprio and Sanders emphasized that now, more than ever before databases drive the management of information. Databases are representations of real-world objects and events which have characteristics or attributes. The structure of the database by necessity puts constraints on which attributes of the real-world entities are described. This idea is not new — card catalogs and online catalogs are databases that do the same thing. A catalog is a database that serves as an interface between a library's collection and its users. Databases create logical structures for objects and data. Logical structures used in library databases are AACR, ISBD, and MARC. With the growth of digital information and different types of databases to manage this information, other structures – XML for example – have come into play. The structure used (how data are represented in databases) should meet the needs of the community working with the information.

In this new environment, data structures do not need to conform to a single standard, so long as these diverse data structures can be mapped to one another. Crosswalks and namespaces are ways that data structures can be matched up or merged as needed. Namespaces on the semantic Web use the infrastructure of the Web to represent agreements on how to refer to a particular entity. For example, in a data structure for biographical data, the data element "bio:title" might refer to a title such as Mr., Ms., or Dr., while in a data structure for bibliographic data, the data element "bib:title" might refer to the title of a book. Namespaces define what "title" means in each type of database and prevent collisions between similarly named data elements when data from different databases are merged.

Rules can also be created between systems that enable one system to evaluate the legitimacy of information coming from another system through built-in semantic Web structures. This process is known as reification. Similarly, systems can make inferences about relationships between or among objects. In the same way call numbers are related in a library catalog, ontologies, or relationships, are created through Resource Description Framework (RDF) triples. For example, a statement can be constructed using RDF that creates the following logic: "If Alice is the mother of Mark, then Alice is Mark's parent." The property of "mother" is a subproperty of "parent." This is a formula for building an ontology, which can be defined as "a rigorous and exhaustive organization of some knowledge domain that is usually hierarchical and contains all the relevant entities and their relations." Ontologies are key components of the semantic Web.

After introducing the terminology of the semantic Web, Caprio and Sanders attempted to demystify these concepts for their librarian audience by illustrating that, conceptually, catalogers also create relationships between entities in the form of subject headings, call numbers, and uniform titles, which are only meaningful in relationship to one another in totality. Catalogers work with multiple languages for structuring data, including AACR2, ISBD, LCSH, LCC, and MARC.

The second half of the breakout session was an open discussion relating the data structures of the semantic Web to the work of librarians and the activities of library users. Caprio and Sanders framed the discussion with the Functional Requirements for Bibliographic Records (FRBR) user tasks of "find," "identify," "select," and "obtain," asking audience members for examples of familiar tasks that fell under each category. They stressed that the FRBR user task model is not linear and that the FRBR concepts of work, expression, manifestation, and item emphasize that bibliographic records in library catalogs need to be thought of, not individually, but in relationship with other records in the database.

Following from this idea, Sanders made the point that what is changing in the cataloging world is that catalogers are now working more and more with large groups of bibliographic records rather than creating and modifying records individually. We have already created a structure for library records. Our current task is to create systems that work better with the data we have, transforming records system-wide as needed, as opposed to updating records individually. We are in the position where we can share our data on a large scale with each other, encoding

decisions in the system about sources of data we deem trustworthy. By allowing other systems to access our data through RDF and related protocols, we expose the data more broadly, making it more useful. For example Innovative Interface's Encore search interface includes a geographic scope, exposing data that were previously hidden.

Caprio and Sanders concluded that librarians' roles are getting bigger, not smaller. Catalogers need to start learning new tools to make library data available in new ways. Cataloging is shifting from a focus on individual records to cataloging large sets of records. The intellectual work required of catalogers will remain the same, but the languages/tools and technology used will be different.

Tradition, Transition, and Transformation A Look at Next Generation Library Systems

In another morning breakout session, John Larson (requirements analyst, Ex Libris) introduced the Ex Libris Unified Resource Management (URM) framework. Larson framed his presentation by stating that the goal of any future library system should be to help libraries render services, not just manage materials. Larson set the stage by examining the three "T's" of a library system: tradition, transition and transformation. Any successful library of the future will need to address these three T's in a dynamic way. The first "T" is "tradition:" libraries should build on their strengths by creating greater efficiencies in carrying out today's services and processes. This includes basic functions like circulation, ordering, and invoicing. The second "T" is "transition,"

which refers to finding ways to enhance today's processes by utilizing changing technology to shift workloads so that staff can be more effective. Examples of this are cooperative collection development, smart fulfillment, interoperability of resources, and a greater role of users in library decisions. The final "T" is "transformation," which deals with supporting the evolving role of the library in its future performing institutional information management activities. For example, library systems must support e-research, institutional repositories, and online services on global, as well as local, levels. Here the librarian's role changes from the gatekeeper of information to more of a central player in the research process.

Larson explained that the Ex Libris URM will allow for a smoother workflow among librarians by giving them access to the information they need in one place. This is accomplished by utilizing an open platform approach with Service Oriented Architecture, documented Web services, and interoperability, which will reduce local IT requirements. The system will be managed centrally and accessed through the Web. Network-level deployment options will facilitate cooperation and community, for example, through forums to discuss issues, trends, and solutions to problems members might face. The URM will also be modular and extensible in order to accommodate future needs, such as new metadata schemas and resource types. Bibliographic control using the Ex Libris URM will be facilitated by a Metadata Management System (MMS). The MMS streamlines bibliographic control by creating a community around metadata and cataloging activities as well as taking advantage of new technologies to facilitate library management activities, as with linked data concepts. The MMS will consist of a "Library Zone" for control of local versions of metadata records and a "Community Zone" for shared descriptive metadata. This will reduce the need to store and manage data locally. A set of

centralized services for authority control and record improvement will be available, and collaboration with fellow catalogers will be possible.

Larson explained that the URM will bring innovation to the selection and acquisition processes as well. By bringing better information about available resources to selectors and users, the URM will support user-driven selection, and the flow of information from selection to acquisition will be streamlined. E-resource management processes will be unified, and acquisitions processes will be automated whenever possible to reduce cost and effort. In terms of new library roles in the research process, Larson noted that the URM will facilitate offering services that can directly aid researchers using digital institutional repositories Ex Libris has just finalized the design of the URM with its development partners. Version 1 of the URM will be beta tested in the second half of 2011. Larson expects that Version 1 will be available generally in the first half of 2012.

Metadata 101

In one of three afternoon breakout sessions, Kelcy Shepherd gave an introduction to the topic of metadata. Shepherd is the digital interfaces librarian at the University of Massachusetts, Amherst Libraries and is an Adjunct Professor at the Simmons College Graduate School of Library and Information Science in Boston. Shepherd explained that metadata are often defined

as “data about data”; however, a more thorough definition is provided by the National Information Standards Organization, which describes metadata as “structured information that describes, explains, locates or otherwise makes it easier to retrieve, use, or manage an information resource.” (1) In today’s digital information age, metadata allow librarians to better describe resources, particularly electronic resources, in order to ensure that the user receives desired information.

Shepherd described the basic functions of metadata. She stated that metadata are digital objects that describe another object and that can be embedded within a digital resource or stored separately. Metadata should follow standards that have been established to support interoperability, as it is of vital importance that multiple systems be able to exchange metadata with minimal data loss. As metadata support the navigation of digital objects, they expand access by allowing for greater discovery.

Multiple components of metadata include elements and structure, content, and format. Metadata consist of data elements, or fields that can be easily indexed and automatically retrieved and the relationships between them. Examples of standards for metadata structure are Metadata Object Description Schema (MODS), Encoded Archival Description (EAD), and Machine Readable Cataloging (MARC). These elements, in turn, follow standards for content with which librarians have become familiar, such as AACR, Resource Description and Access and Cataloging Cultural Objects. Metadata content follows data value standards, or controlled lists of values for specific fields, in order to facilitate searching across a variety of systems. Two examples of controlled

vocabulary data are the Library of Congress Subject Headings and the Union List of Artists' Names. Finally, metadata must follow a format for data encoding, such as ISO 2709 or Extensible Markup Language (XML).

Shepherd explained that there are four types of metadata: descriptive, structural, use, and administrative. Descriptive metadata are information about the intellectual content and physical format of the object. This includes such basics as title, author, size, and contributor and is vital in supporting discovery, identification, and selection of the information. Descriptive metadata follow standards including Dublin Core, MARC, MODS, and EAD.

Structural metadata are information about the relationships among individual components of a complex digital object in order to support navigation. This structure includes data related to individual files, physical and/or intellectual structure, and behaviors. Shepherd gave an example of a twelve-page diary with multiple images. Structural metadata relate the pages to one another so that they can be displayed in order and connect the images to the pages on which they belong. In addition, the diary may have multiple views (e.g., PDF, thumbnail images, and transcript). The systems that are used to retrieve the components of the diary need to know how to identify each component and how each component relates to the others. Structural metadata relate the multiple views to one another so they will display properly, thus supporting navigation of complex digital objects. The standard used for structural metadata is Metadata Encoding and Transmission Standard (METS).

Use metadata consist of information about how and how much a digital object has been used. Examples of use metadata include familiar information such as circulation statistics as well as data such as user tracking, search logs and “hits” to Web sites. There has been a move toward common standards for collecting this information through the Standardized Usage Statistics Harvesting Initiative (SUSHI) and Counting Online Usage of Networked Electronic Resources (COUNTER).

Shepherd divided administrative metadata into three categories: technical metadata, preservation metadata, and rights metadata. Technical metadata consist of information about the technical processes used to produce a digital object, or required to use it. Technical metadata support quality assessment regarding how a digital object was produced, accurate rendering of the digital object, and its preservation. This category includes data related to the hardware and software used to create the data or needed to render it, digitization protocols, and authentication and security details. Standards for technical metadata include Metadata for Images in XML (MIX) and Technical Metadata for Text (textMD).

The second category of administrative metadata, preservation metadata, pertains to information about the preservation management of digital materials in order to support long-term retention and accessibility. Preservation metadata includes data on how the digital material was created, including what programs were used, any changes to the digital object or its chain of custody, and technical requirements for accessing the information. This is necessary for being able to access the data now and in the future.

Rights metadata contain information about rights related to access and use of information in support of appropriate use by end users. This includes data related to copyright, licensing, and terms and conditions of use. Rights metadata follows the standards CopyrightMD and METSRights.

After outlining the four basic types of metadata, Shepherd discussed descriptive metadata in more detail, addressing four standards for descriptive metadata currently in use: Dublin Core, MODS, VRA Core, and EAD.

Dublin Core began as a basic fifteen-element system to describe Web-based documents. It remains very general and supports a wide range of resources. Every element in Dublin Core is optional and repeatable, and the standard is not format specific. Two online projects that use Dublin Core metadata are the Maine Memory Network and Connecticut History Online.

MODS is another descriptive metadata standard. It consists of a subset of MARC elements, but it is less granular than MARC and employs user-friendly, language-based tags instead of numerical tags. MODS was designed to be particularly applicable to digital resources and follows the common XML format. The Library of Congress uses MODS in its online I Hear America Singing project, as does the University of California's Calisphere.

A third descriptive metadata standard presented by Shepherd was the Visual Resources Association's VRA Core, which is used to describe visual materials, including works, images, and collections. VRA Core 4.0 also uses the XML format. Finally, Encoded Archival Description (EAD) is another descriptive metadata standard used for archival materials. EAD supports complex hierarchical structures. It, too, uses the XML format. Examples of online resources using EAD metadata are the Northwest Digital Archive and the Online Archive of California.

Shepherd spoke further about the structural metadata standard METS. This schema is designed for the management, exchange, and display of digital objects and has the capacity for descriptive, structural, and administrative data, incorporating other metadata standards. METS metadata can be used for images, transcripts, descriptions of objects, and multiple thumbnails all on one XML document. Projects using METS include the Brown University Library Center for Digital Initiatives, University of Florida Digital Collections, and DRAM (Database of Recorded American Music (DRAM).

Shepherd concluded with the point that the confusing number of metadata standards in existence is a result of the different ways metadata are used. Metadata standards exist for different functions, purposes, formats, audiences, and communities.

Go Fish! How to Catch and Clean MARC Records Using Z39.50 and MarcEdit

In another afternoon breakout session, Benjamin Abrahamse (head, Serials Cataloging Section, Massachusetts Institute of Technology) gave a breakout session entitled, "Go Fish! How to Catch and Clean MARC Records Using Z39.50 and MarcEdit." He presented a method he has developed for assembling MARC record sets using non-MARC publisher-supplied metadata. This method uses MarcEdit to retrieve bibliographic records from OCLC using Z39.50, spreadsheets to examine the set of bibliographic records for good and bad data, a text editor that supports regular expressions, and MarcEdit for final editing of sets of bibliographic records before loading them into the catalog. He said that the skills necessary for his procedure include knowing how to form basic Z39.50 queries, how to use regular expressions, and how to use sorting and filtering functions in spreadsheets.

Abrahamse's method begins by harvesting from a provider's Web site basic bibliographic data (e.g., ISBN, LCCN, DOI, as well as complete title information and URLs) for the set of e-books to which he has gained access through subscription or purchase. He opens this information in a spreadsheet, selects the fields to query, and exports it to a text editor. After converting the data to a Z39.50 query, he saves it as a text file. He brings this file into MarcEdit and then uses its Z39.50 function to retrieve bibliographic records from OCLC. After reconverting the bibliographic data from OCLC to a tab delimited text file, he imports it into a spreadsheet. Once there, he sorts it by shared values to remove duplicates and to filter out unwanted records. For

instance, he uses data such as the encoding level and number of holdings in OCLC to choose which record to keep in his file.

Once he has identified which bibliographic records to add to the catalog, he reimports the tab delimited file into MarcEdit and, again, uses Z39.50 to retrieve the needed full bibliographic records by the OCLC numbers. When he has the file of records in MarcEdit, he performs functions such as adding or removing whole fields, editing subfields (such as removing 300 delimiter "c" to convert print records to e-resource records), and editing indicators.

Abrahamse finished his presentation by outlining best practices. These included how to name files and how to form queries. He recommended OCLC, Library of Congress, and the catalogs of Harvard, MIT, and the University of California as places from which sets of bibliographic records may be harvested.

Creating a Trillion-Field Catalog: Metadata in Google Books

The afternoon keynote speaker was Jon Orwant (engineering manager, Google Books, Google Magazines, and Google Patents). Orwant outlined Google's techniques for correcting metadata for Google Books. These metadata originate from over a hundred sources, all of which are

incomplete, inaccurate, and ill-formatted. Orwant described the intention behind the Google Books project and the process of scanning books. In 2005, Google announced its intention to scan every book in the world as part of their mission “to organize the world's information and make it universally accessible and useful.” Thus far Google has scanned twelve million books, 10 percent of the works printed since Gutenberg invented the printing press. Twice a week, using mathematical models, Google attempts to count all the books in existence. Their current calculation estimates that there exist 174 million manifestations of 120 million works and that there are four billion metadata records describing them.

To scan books, Google works with both publishers and libraries. From most publishers, Google receives a hard copy book. Before the books is scanned, its spine is sliced off, after which the book’s pages are fed through a sheet-fed scanner. This process was not acceptable for library books, so Google developed a non-disruptive procedure in which people turn the book’s pages as cameras photograph them. After a book is scanned, its image must be processed: The image is cropped to the size of the page and an algorithm is applied to remove any warping in the text caused by the curvature of the page. The image is further processed to remove dirt and any stray images of the page turner’s fingers. The book is then submitted for optical character recognition (OCR) and metadata are created for the book. Finally, the book is ranked and indexed in books.google.com.

After a book is scanned, metadata are created that identifies the work, expression, and manifestation. Google relies on metadata from many sources, including OCLC, library catalogs,

the Library of Congress, Bowker's *Books in Print*, and the book scan itself. These sources are combined to create a "best" record for each expression or manifestation. The generation of metadata is entirely automated, and the algorithms that drive it are continually revised as Google discovers systemic errors and peculiarities. Orwant provided examples of the metadata-creation process.

One problem Orwant described was creating accurate metadata for sets (e.g. *Lord of the Rings*), series (e.g. *The Hardy Boys*), serials, and multivolume works with different titles for each volume. These types of materials are difficult because there is little uniformity in multivolume work cataloging. For example, one OCLC number might cover volumes with multiple ISBN's. Google spent about a year creating rules about which metadata sources to trust, which fields in MARC records to rely on for what data, and how to combine these fields to create an accurate Google Books metadata record. For some multivolume books, Google examines the physical description (300) field in order to detect "multi-volumeness," as well as ISBN (020), title (245), and contents (505) in order to extract metadata for the scanned item and cluster related titles together in the case of books in a series. In another example, Google looks for the word "set" in the bibliographic record. Google has translated "set" into a number of languages so it can be detected in bibliographic records in multiple languages.

In the case of serials, libraries use a single bibliographic record for each serial and a single barcode for each bound volume. Google, however, wants to create separate metadata for each issue of a periodical. They have created a probabilistic framework to detect and create metadata

for individual periodical issues. As a result, periodicals are now listed in Google Books with the volume number as part of the title.

An additional problem mentioned by Orwant involves ISBNs. In some countries, publishers assign random ISBN's to make their books look more "valuable." Google has discovered that as a result hundreds or thousands of books share the same ISBN. Once Google realized this was happening, they wrote rules into their metadata-creation software to ignore ISBNs of books published in certain countries, relying instead on the OCLC number to harvest metadata from bibliographic records. This is an example of Google being able to identify systematic problems and to write software to correct problems based on analyzing large data sets.

Google has also made great strides in author disambiguation based on publication date, title, and other metadata. For example, Google used to list any contributor to a book as an author, but now if their software finds the contributor listed as an editor in any of the bibliographic records they examine, the contributor will not default to "author" in Google's metadata. In another example, Google used to have difficulty distinguishing between different versions of the same author, e.g. "Mao," "Chairman Mao," and "Zedong Mao," and treated them as separate authors. Now, string comparison techniques combine them as one author. Google can also recognize the same name in different scripts and looks for names in different parts of a MARC record depending on the country of origin of the record.

Orwant provided examples of how Google has learned to handle dates of publication. Certain combinations of titles and dates are blacklisted; for example any edition of *Our Bodies, Ourselves* published prior to the twentieth century. In some instances, incorrect dates resulted when Google interpreted dates expressed in the Islamic calendar as dates expressed in the Gregorian calendar. Google's algorithms monitor the distribution of books by publication year, so they can investigate any unusual spikes in publication that could indicate errors in publication dates. Google is very careful about publication dates, as an incorrect date of publication could result in fines for copyright infringement since date of publication drives whether or not a book is in the public domain and therefore available to Google Books viewers in its entirety.

Orwant next discussed how Google can infer the subject of a book. Google draws in metadata from library records, as well from BISAC (Book Industry Systems Advisory Committee) subjects provided by publishers. Google Book searchers tend to be seeking broader subject terms than those provided by Library of Congress Subject Headings, so Google is able to generate BISAC subjects from LC subjects. Sometimes this doesn't work smoothly, as when a scholarly work about spiders was erroneously classified as juvenile literature.

To wrap up his talk, Orwant shifted his focus from metadata to the future of Google Books according to the terms of the proposed Google Books Settlement Agreement. He explained that Google intends to continue scanning books. They will remove any book from Google Books at the request of a copyright holder and will display only 20 percent of the text of any book not in the public domain for which the copyright holder cannot be found. Subscriptions to English-

language materials will be available to libraries, and Google will place one terminal in every public library building in the United States with which patrons can access the full text of Google Books materials. Google will sell access to Google Books content on behalf of rights holders. They will sell access to books not in print for which the copyright owner is unknown, with the proceeds of the sale held in escrow in the case the rights holder is located. Google will provide technologies that make books more accessible to the disabled and will fund a separate, non-profit organization to search for rights holders of material Google has scanned.

Orwant emphasized the potential of Google Books as a research corpus of all books ever published. For example, Google engineers worked with researchers to examine word use throughout the Google Books database. They examined frequency of word use and changes in the past tense of verbs over time in order to predict when a verb's past tense will become regularized and formed by adding the suffix "-ed." For example, the past tense of the word "sneak" has shifted from "snuck" to "sneaked." Google's goal is to have anyone be able to use Google Books to perform this kind of analysis. Orwant provided another example in which commonly used patterns of three words, or trigrams, could be used to estimate when a book was published.

In a final example, Orwant described how Google Books was used to test the "great man" theory of history. Researchers attempted to answer the question of whether developments in modern calculus were "in the air" and simply recorded by "great men," such as Sir Isaac Newton and Gottfried Leibniz, or whether these men made unique contributions not arrived at by others.

Researchers wrote a program to translate calculus concepts into multiple languages and then searched Google Books for the frequency of these terms by date. They determined that at the time that Newton and Leibniz were writing about calculus, many others were writing similar material. By this analysis, Albert Einstein does appear to be a "great man," as he was the only person at the time writing about theories of relativity.

Contact info:

Andrée J. Rathemacher

Associate Professor
Head, Acquisitions
University Libraries, University of Rhode Island
15 Lippitt Road
Kingston, RI 02881-2011
Phone: (401) 874-5096
Fax: (401) 874-4588
E-mail: andree@uri.edu
http://www.uri.edu/library/

Martha Rice Sanders

Knowledge Management Librarian
The HELIN Consortium
15 Lippitt Road
Kingston, RI 02881-2011
Phone: (401) 874-4951
Fax: (401) 874-4588
E-mail: msanders@etal.uri.edu
http://library.uri.edu/screens/libinfo.html

Michael A. Cerbo II

Assistant Professor
Bibliographic Access & Resource Management Librarian
University Libraries, University of Rhode Island
15 Lippitt Road
Kingston, RI 02881-2011
Phone: (401) 874-5967
Fax: (401) 874-4588
E-mail: mcerbo@uri.edu