Future Directions in Metadata Remediation for Metadata Aggregators

February 2009

Greta de Groat

DIGITAL LIBRARY FEDERATION

First Digital Library Federation electronic edition, February 2009

Digital Library Federation [ISBN-13: 978-1-933645-07-0]

www.diglib.org.

This publication was funded by generous support from the Gladys Krieble Delmas Foundation.

Contents

About the Author

Greta de Groat

Greta de Groat is a Discovery Metadata Librarian at Stanford University Libraries. She is the past ALA Committee on Cataloging: Description and Access liaison from the ALA Association for Library Collections and Technical Services Networked Resources and Metadata Interest Group and the Online Audiovisual Catalogers Cataloging Policy Committee. For the past decade she has taught cataloger training workshops, particularly in cataloging Internet resources. Before coming to Stanford, Greta was the Authority Control Librarian and head of the Quality Control Unit at WLN. She is also a silent film historian and serves as the librarian at the Niles Essanay Silent Film Museum, as well as maintaining a website on silent film actresses.

Acknowledgments

This report was commissioned by the Digital Library Federation (DLF) through the DLF Aquifer initiative. Chair of the Aquifer Metadata Working Group Jenn Riley and DLF Aquifer Director Katherine Kott offered oversight and guidance. San José State University School of Library and Information Science master's degree student Bonnie Fullerton tested software and assisted with the glossary. Scott Van Duyne of Stanford University Libraries ran the Stanford Named Entity Recognizer program. David Newman of the University of California, Irvine provided the Statistical Topic Model information for **Appendix A**. Annie Wu provided editorial assistance through the Stanford Alumni Association volunteers program, in addition to DLF Program Manager Barrie Howard. DLF gratefully acknowledges support for this study from The Gladys Krieble Delmas Foundation.

[This page left intentionally blank.]

Executive Summary

DLF Aquifer, a Digital Library Federation initiative, focuses on making digital content—especially cultural heritage materials pertinent to American culture and life—easier for scholars to find and use. One avenue to providing better access to digital collections is by including the collections in aggregations that are promoted and exposed through commonly used channels such as commercial search services.

Successful aggregation depends on robust, consistent metadata. While data providers may strive to include all applicable fields for their chosen metadata format in newly created records, records that have been mapped from legacy data in other formats will seldom be optimized in their new home, and the creators of these records may not have the resources to augment these records in any more than the simplest ways. Remediation tools to improve the quality of metadata for improved services are therefore highly desirable.

With support from The Gladys Krieble Delmas Foundation, the Digital Library Federation embarked on a project to inventory existing tools and services for metadata mapping, remediation, and enhancement. Once identified, tools were evaluated for general applicability across digital library and other cultural heritage environments. The results of the research show that a handful of tools are usable as-is, but many tools need more work to be generally applicable in a variety of environments and significant development would be required to create a robust and well-defined set of metadata remediation services. Key points of note:

- Relatively few tools are available that can work directly on metadata records rather than full text, and those that are available need to be customized for each aggregator.
- Workable tools are available for date normalization, and also for normalizing and matching coordinates to U.S. geographic names.
- A statistical topic model program for subject clustering has been developed.
- Both named entity and topical keyword extraction remain problematic, with a fairly high percentage of errors.
- Authority files may be used to break up pre-coordinated Library of Congress subject strings into topical, name, geographic, temporal, and genre facets to improve searching.
- Mappings between different thesauri, which should allow for better search processing in aggregations containing multiple subject vocabularies, are still under development.
- Infrastructure for work collocation, appropriate to aggregators with significant published materials, is still underdeveloped and will probably need to wait for the widespread adoption of the new standard for resource description, Resource Description and Access (RDA).
- Unambiguous identifiers for entities such as names and works would be useful when the community infrastructure is developed, but are not yet supported by most metadata formats.
- Unambiguous, machine-actionable rights statements are also an area where the community infrastructure is still under development.

The report is organized by categories of service that could be enabled with better metadata. Each section includes an inventory of available tools, assessment of those tools, and an evaluation of what might be accomplished in the future. The information gathered here provides a reference framework for members of the cultural heritage community to use when considering their own tool development priorities, and a road map for areas that would benefit from collaborative efforts.

Introduction

One avenue to providing better access to digital collections is to include the collections in aggregations that are promoted and exposed through commonly used channels, such as commercial search services. Successful aggregation is assisted in large part by robust, consistent metadata.

While data providers may strive to include all applicable fields for their chosen metadata format in newly created records, records that have been mapped from legacy data in other formats will seldom be optimized in their new home, and the creators of these records may not have the resources to augment these records in any more than the simplest ways. Aggregators that provide value-added services are in a better position to programmatically remediate large quantities of records in order to enhance their services. As metadata for different collections vary widely, it is difficult to provide a "one-size-fits-all" approach to remediation, but it is clear that remediation on a collection-by-collection basis is not sustainable. Better metadata leads to better services, and the richer the initial metadata, the better remediation strategies are likely to work.

This report will detail the current state of the art of remediation efforts, describe the additional services that aggregators could offer if the metadata were there to support them, and identify the types of tools that are needed to remediate the metadata in order to achieve the desired level of service. The report is aimed toward designers of metadata aggregations, including programmers, project planners, and metadata specialists. Knowledge domains such as computer science, informatics, information retrieval, information science, and library science are within the scope of the report. It is assumed that remediation efforts will be focused on working with the metadata itself, as many aggregators do not have access to the raw digital item. Some processing related to the raw digital item would be ideal for certain purposes and is mentioned in passing, but it is considered out of scope for this report. It is assumed that some of the metadata may be newly created, but much of the metadata may be legacy data. For example, legacy data may include library cataloging data mapped from older MARC bibliographic records, which would likely use LC subject headings and a combination of controlled and uncontrolled names. Other metadata may have been created by other academic departments or institutions, perhaps stored in databases using locally developed thesauri and names in nonstandard forms.

Many of the following ideas are not new. Some facilitate the basic search and browse functions that we are accustomed to in the library world, while others enable new services that take advantage of Web 2.0 technologies. Some of these remediation techniques may be constrained, depending on the schema of the metadata. For example, the lack of granularity in simple Dublin Core will not support some techniques that could be used with a more granular schema such as MODS.

Many of these remediation technologies involve data mining in some form. Typically, this has been done with full text documents, but many aggregations are made up of primarily photographs, maps, ephemera, and other types of materials with no text to mine. For aggregators, textual information would need to be extracted from existing metadata records.

Processes would extract entities such as personal and geographic names and topical words, cluster and normalize variant forms, search and match against external authority files, and insert a correctly encoded element back into the metadata record. This is a complicated multistep process that will have potential problems at almost every step. These technologies are far from perfect, and in some cases the needed support infrastructure in the larger library world is still underdeveloped. Metadata enhanced in this way will be messier and "uglier" than handcrafted metadata, and will have some degree of clutter and redundancy, as well as a certain percentage of erroneous data. Service providers will need to decide if the gain in added access is significant enough to outweigh these disadvantages. Librarians will need to give up some degree of control over neat and precise metadata records (which are far less neat and precise than we would like to believe).

The following report will discuss user tasks and services, the necessary metadata support to complete those tasks and services, and the existing tools and proposed tools used to remediate the metadata. The report is divided into sections based on broad metadata elements (i.e., topical subjects, genre, names, geographic information, dates, title information, type of resource, addressable raw object, rights, identifiers). Within each metadata element there is a description of certain desired services, each of which further discusses metadata support, existing tools, and desirable tools for attaining that service.

The report also includes a glossary of technical terminology. The appendices contain the results for informal testing conducted on three different metadata remediation techniques involving the statistical topic model, entity extraction software, and foreign title translation.

A. Topical Subjects

Summary of desired services:

- Cluster topically similar records to enable high-level browse and filtering
- Increase the number of records searchable by subject by automatically assigning subject headings, preferably from a recognized vocabulary
- Use library-generated subject headings to their greatest potential
- Use subject terminology consistently on all the metadata records in an aggregation
- Increase available search terms in records by incorporating user-suggested tags into metadata

Desired service: Cluster topically similar records to enable high-level browse and filtering

Metadata support: Rich descriptive information would be needed in the metadata record, or the tool would have nothing to work with. The tool could examine specific fields containing content likely to be of high value, such as titles, descriptions, abstracts, and notes. (See also **Section F—Title Information**, which discusses titles for remediation to that field, which would also support this function). Machine translation could be done for non-English titles and the translated title added to the metadata record to provide words for the tool. User tags might also be considered. If available, a field for classification could also be helpful to use in this process, particularly when there is ambiguity in the subject words. Collection-level information could also be used, but if the scope of the collection-level information is broader than that of the individual record, one runs the risk of associating the record with overly broad or irrelevant names and subjects. However, as with classification, collection-level information could be useful to provide context for otherwise ambiguous words. Ideally, the metadata format should have some way to identify fields containing machine-generated data added by the aggregator, such as a "machine generated" attribute.

Existing tools: Statistical topic model program for subject clustering used for the OAIster collection of metadata through the DLF portal (see http://quod.lib.umich.edu/i/imls/), developed by David Newman of the University of California, Irvine. An example of how the program works and instructions for obtaining the code from the developer are available in **Appendix A**.

Desired tools: The tool developed by David Newman is currently designed to be used with Dublin Core and an in-house subject vocabulary. It would be useful to adapt it for use with MODS and other formats. It would be technically possible to use LCSH (or more likely FAST) as the clustering vocabulary, or to use whatever vocabulary is most useful in the context of the particular aggregator. In order for an aggregator to take advantage of this tool, the search

screens need to be optimized for broadening and narrowing searches by category. This tool could be used in conjunction with translation tools detailed in the section on titles in this report, and perhaps even with harvested user tags to increase the amount of useful text for the tool to use.

Literature: Hagedorn, K., S. Chapman, and D. Newman. "Enhancing Search and Browse Using Automatic Clustering of Subject Metadata," *D-Lib Magazine* 13 no. 7/8 (July/August 2007), http://www.dlib.org/dlib/july07/hagedorn/07hagedorn.html.

Newman, D., K. Hagedorn, C. Chemudugunta, and P. Smyth. "Subject Metadata Enrichment Using Statistical Topic Models," in "Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries," JCDL 2007, Vancouver, British Columbia, Canada, June 18–23, 2007: 366–375, http://doi.acm.org/10.1145/1255175.1255248.

Comments: This is a promising methodology for enhancing subject access. It is designed to work with existing metadata records rather than full text, and was used in a test database consisting of the whole corpus of metadata from the OAIster repository (at that point, nine million records). The first step in this process is to run it against the whole aggregation of metadata records to gather and cluster terms from meaningful fields like title, description, and subjects, and from that process determine the topics associated with each record. The process is labor-intensive and involves a great deal of processing time. For the OAIster test, a team of manual reviewers looked at the clusters and assigned high-level subjects, but one could theoretically automate part or all of this process or choose a different method to describe the topics. Once that is done, this part of the program may not need to be run for a number of years (basically not until new terms are needed, which is likely if there are collections in new subject areas). Cluster labels, including broad classification terms, can be added back to the individual records. This part of the process is re-run each time a collection is added or updated.

The subjects assigned to OAIster were from a local classification scheme, but a subset of basic LCSH/FAST terms could be used. This subset would need to be manually developed by each aggregator for its own use, depending on the subject focus.

While the researchers on this project judged that the clustered topic was mostly useful and was correctly assigned to enough records to make it a worthwhile process, it was in an aggregation with a large percentage of scientific materials, which are the most amenable to such processing. Records with very minimal metadata, such as the title of a photograph that is simply the name of a person depicted in the photograph, leave very little for the program to work with, and results can have significant errors. The tool should be further tested by different aggregators with different types of materials and feedback from users should be solicited as to whether the usefulness of the appropriate additions outweighs the inconvenience and the potential unreliability of the results.

Desired service: Increase the number of records searchable by subject by automatically assigning subject headings, preferably from a recognized vocabulary

Metadata support: See Metadata support for the previous entry.

Existing tools: There are many commercial products attempting to analyze text, extract meaningful words, create an ontology, or map them to an existing ontology. One example is Kea (http://www.nzdl.org/Kea/), a key-phrase extractor that is distributed under the GNU General Public License and has been used as a component in several tools. The documentation says that Kea will index controlled vocabularies in SKOS format, and already does MeSH and others. If plans to make the Library of Congress Subject Head available in SKOS come to pass, this vocabulary could potentially work with Kea.

Desired tools: One needs an effective way to assign topical subject headings, preferably from a recognized subject authority such as LCSH, FAST, and the like. Terms would have to be extracted and/or matched from the metadata record itself (using meaningful fields such as title, abstract, notes, and possibly refining by classification), except in the case of textual materials, which could have the full text mined if the digital item is available to the aggregator. The terms could then be queried in openly available machine-readable authority files. If the queried terms match authorized terms or their cross references in the chosen thesaurus, these authorized terms could be imported back into the record as appropriately coded metadata elements.

Literature: Bamman, D, and G. Crane. "Building a Dynamic Lexicon from a Digital Library," in "Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries," JCDL 2008, Pittsburgh, Pennsylvania, United States, June 16–20, 2008: 11–20, http://doi.acm.org/10.1145/1378889.1378892.

Medelyan, O., and I. H. Witten. "Thesaurus Based Automatic Keyphrase Indexing," in "Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries," JCDL 2006, Chapel Hill, North Carolina, United States, June 11–15, 2006: 296–297, http://www.cs.waikato.ac.nz/~olena/publications/jcdl06_kea.pdf.

Witten, I. H., K. J. Don, M. Dewsnip, and V. Tablan. "Text Mining in a Digital Library," *International Journal on Digital Libraries* 4, no. 1 (August 2005): 56–59, http://www.dcs.shef.ac.uk/~valyt/download/greenstone-gate.pdf.

Comments: There are many projects, both in the commercial and academic sectors, attempting to use various text mining methodologies or taxonomic type-indexing for assigning topical information for this Holy Grail of automated metadata services. Unlike the statistical clustering tool by Newman previously described, they have generally been used on full-text materials rather than metadata records, and it is not clear how successful they would be on the much sparser information contained in metadata. Some rely on matching phrases and words to an existing ontology, which is similar to the latter part of the processing of the statistical clustering tool. Most of these involve human intervention at some scale, usually in choosing ontology terms or deciding on matches/non-matches that the program has found difficult. As many of these are designed for narrow topical areas, they often prove difficult to scale, and topically heterogeneous collections make relevant vocabulary harder to pin down and disambiguate. The programs need training data for the program to identify relevant words and phrases and "learn" how to assess their relevance. Sometime this is in the form of a large corpus of text, sometimes it can learn from human-coded data. Some of the relevance algorithms rely on frequency of a term's occurrence in the text, which would not be applicable when extracting terms from a metadata record. There is usually a large amount of processing time involved, particularly during the early stages of the process.

This method would be particularly useful for records with no subject headings, as well as records with inadequate subjects. It would not be very useful on records with nondescript titles and no abstracts. Despite these disadvantages, there is a large amount of work going on in this field, and this area bears watching for future developments. However, it is probably too underdeveloped at this point to be useful for metadata remediation.

Ideally the ontology from which the subject terms are assigned should be an authorized vocabulary such as LCSH, though it may need to be augmented by a great many more lead-in terms to be useful in this context.

Desired Service: Use library-generated subject headings to their greatest potential

Metadata support: Split un-subfielded, precoordinated LCSH strings derived from non MARC metadata into separate fields, subfields, or subelements (such as MODS <subject> subelements or Dublin core fields <subject> <coverage> and <type>).

Existing tools: Library authority control vendors (e.g., Backstage Library Works) have custom algorithms to locate subordinate data in unsubfielded MARC records and correct the tagging. For matching purposes, Library of Congress name and subject authority files are openly available, and some libraries and authority control vendors have manually created lists of authorized LCSH subdivisions. In addition, the OCLC Terminologies project (http://tspilot.oclc.org/resources/index.html) has LCSH and FAST records available in a SRU searchable form.

Desired tools: A tool to take un-subfielded, precoordinated LCSH strings and parse them into subelements that could be appropriately tagged, and insert them back into the metadata record is desired. The LCSH-derived subelements could be matched against an openly available machine-readable LCSH authority file to identify the type of string (e.g., topical, geographic, genre, temporal, name). The appropriately encoded strings should then be reinserted into the metadata record and, where supported by the metadata format, an indication as to the source of the vocabulary term should be made.

Literature: Backstage Library Works. MARS Authority Control planning guide (http://www.marclink.com/MARSguide1.pdf).

LTI. Authority Record Matching (http://www.authoritycontrol.com/A-MATC-D.html).

Comments: Library of Congress Subject Headings make up the bulk of subject headings in aggregated collections that consist largely of library legacy data, such as American Social History Online. Many records contain precoordinated strings of the type (e.g., Topic—Place—Time period—Genre). Metadata records derived directly from MARC records could have the v, y, and z subfields, as well as headings encoded as 600, 610, and 651 mapped into appropriate elements when the metadata format supports this granularity. However, when they are mapped into another metadata format, those elements sometimes have ended up in a single field, making it difficult to parse them into data to support faceted searches. Records deriving from databases where there was less granularity in their subject fields may also have the entire string residing in the same field, usually subdivided with a dash, or a dash with a space on either side. In addition, LC did not implement the MARC genre subfield v until 1999 and has only just started implementing the MARC 655 field (and has not yet for all formats). Therefore, many records do not have genre elements explicitly coded, and these usually end up mapped into topical subject fields.

This seems like low-hanging fruit as far as remediation is concerned, since this is a long-standing problem, the vocabulary of the strings is finite and relatively well controlled, and the machine processing is relatively uncomplicated. There would be a measurable benefit to parsing these strings and making them available to a search interface. Search functions can use explicitly encoded geographic information for faceted searching for place names, as well as for map views. Genre is an underutilized element in traditional library catalogs, as well as in aggregators, partially due to past MARC and Library of Congress practices. The data currently buried in the precoordinated strings and topical fields could be brought out and used in a genre search or limit. LCSH temporal data tends to be rather broad, but it could be used to inform other date information in the record. Names used as subjects are often coded as topical subjects, making them unavailable for a name search. Processes for names, genres, and places detailed elsewhere in this report are partly or fully dependent on a tool to enable these elements to be explicitly identified.

Authority control vendors have long had manually created lists of valid LCSH subfields, as have some libraries, and LC has in the last few years provided subfield authority records and has begun adding references from the "indirect" form of geographic names found in LCSH subdividisions (e.g., California—Los Angeles). Versions of LCSH are available from the OCLC Terminologies project (http://tspilot.oclc.org/resources/index.html) in an SRU searchable form, and at http://lcsh.info. FAST, based on uncoupled LCSH strings, is also available from the same source.

Desired service: Consistent use of subject terminology on all the metadata records in an aggregation

Metadata support: For metadata formats that support it, controlled subject fields should have appropriate identification of the thesaurus to which the terminology belongs in order to drive the mapping between the thesauri. This information could be provided by the data provider, programmatically determined by the service provider, or found on an as needed basis by an API.

Existing tools: The OCLC Terminologies project (http://tspilot.oclc.org/resources/index.html) is attempting to bring together several vocabularies in an SRU searchable form, but at this point there does not seem to be a way to search all vocabularies at once and some major terminologies (most obviously the Art and Architecture Thesaurus) are available only by license agreement. The crosswalks between the thesauri are under development, but not yet completed. Within a single thesaurus, broader and narrower terms are available for programmatic use in searching.

Desired tools: From the metadata record, take an uncontrolled term or terms from a thesaurus other than the one the aggregator wants to use, and generate terms from the desired thesaurus for them. When a metadata record is preprocessed by the aggregator, the terms would be queried in multiple thesauri until a match is found. The thesauri may reside in a local database, or they may reside in an offsite service, such as the OCLC Terminologies Service. The query should identify the thesaurus to which the term belongs and retrieve equivalent terms from desired thesaurus. These terms would then be inserted in the metadata record. This process might be facilitated in metadata formats that have a means to identify the authority source for a subject field. In this case, the aggregator preprocessing stream could determine whether that identification is present, and query that specific thesaurus. Another method might be to match all the vocabularies in preprocessing and insert all applicable terms into the metadata record. This approach may take more preprocessing and bloat the record. It would also be desirable to insert the highest level term in the term hierarchy into the metadata record to facilitate query expansion.

Literature: Vizine-Goetz, D., C. Hickey, A. Houghton, and R. Thompson. "Vocabulary Mapping for Terminology Services." *Journal of Digital Information* 4, no. 4 (March 2004), http://jodi.tamu.edu/Articles/v04/i04/Vizine-Goetz/.

Comments: When metadata records come from a variety of sources, each source may have used different subject thesauri for their subject headings. Users who search one term may fail to retrieve appropriate records that have subject headings from a different thesaurus. It would be useful for an aggregator to take subject terms from other thesauri and substitute the equivalent term from their chosen thesaurus. Some individual vocabularies are available publicly or through the OCLC Terminologies service. Mappings between them are still under development and need to be openly available for queries so that developers can use them in the preprocessing and/or searching functions. Some vocabularies, like the Art and Architecture Thesaurus, are only available by license agreement, that may hinder the functionality of records that have those headings. This would particularly affect aggregations containing some museum data amidst a much larger percentage of library data. While mappings between vocabularies remain underdeveloped or unavailable, there will not yet be a community infrastructure to support cross-mapping functionality.

Desired service: Increase available search terms in records by incorporating user-suggested tags into metadata

Metadata support: Tags generated by users could be used to enrich existing metadata. For metadata formats that support it, tags could be harvested and inserted into subject fields, marked with their source. Alternatively, the terms could be housed separately from the metadata but aggregated together. This

might be a useful way to augment other processes, such as topic clustering or entity extraction, particularly for text-poor resources with minimal metadata such as photographs.

Existing tools: HarvANA (see Hunter article in Literature below for contact information), Steve.Museum tagger (http://www.steve.museum/).

Desired tools: Adapt the above tools or develop a similar one to collect user tags (encouraging them to use meaningful tags) and associate them with or add them to the metadata record.

Literature: Hunter, J., I. Khan, and A. Gerber. "Harvana: Harvesting Community Tags to Enrich Collection Metadata," in "Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries," JCDL 2008, Pittsburgh, Pennsylvania, United States, June 16–20, 2008: 147–156. http://doi.acm.org/10.1145/1378889.1378916.

Trant, J. "Social Classification and Folksonomy in Art Museums: Early Data from the Steve.Museum Tagger Project," 17th SIG/CR Classification Research Workshop, November 4, 2006, http://www.slais.ubc.ca/USERS/sigcr/sigcr-06trant.pdf.

Comments: The HarvANA tool is intended to be open source, but it is still in development and is not yet ready for open release. There are three parts to this tool: the client plug-in for annotating web resource, the annotation server, and the OAI-PMH harvesting layer on the annotation server. The tagging is not done through the search interface, but through an IE plug-in that the users need to install. Specific tagging ontologies are developed by knowledge domain experts for a particular implementation and are exposed to the users through a drop down menu (an "ontology-directed-folksonomy" approach). Users may use one of the provided tags or add their own. They are also able to make free-text annotations. The source of the annotations is authenticated to help protect against malicious tagging and tag spam. The annotations are then converted into RDF, which are harvested via OAI-PMH. They are aggregated with the original metadata in a centralized metadata store. Information can be found on the AUS-e LIT project website at http://www.itee.uq.edu.au/~eresearch/projects/aus-e-lit/.

The Steve.Museum tagger is a collaboration between several museums to research the application of semantic tagging and folksonomies for online museum collections. They have developed a suite of open-source tagging tools available on SourceForge (http://sourceforge.net/projects/steve-museum). It is geared toward art museums, but could be applicable to other projects with large numbers of images. The Institute of Museum and Library Services (IMLS) is funding additional projects that will build on the Steve tagger, so additional tools may be forthcoming.

Since the same digital objects are being harvested multiple times for different aggregators, and each could theoretically be providing the option of user tagging, the same object may be gathering different tags in different aggregations. Perhaps it would be useful if there was a process to feed the tags back to the original metadata provider so the tags can be shared across multiple aggregations. It remains to be seen whether users will actually get very far in tagging records that number into the thousands.

B. Genre

Desired service: Ability to accurately and consistently search by genre when appropriate

Metadata support: Genre terms consistently encoded as genre elements within each metadata format (e.g., in MODS <genre> element or <subject><genre> subelement, or Dublin Core <type>). Content should be from an identifiable authority when possible, and in metadata formats that support it, the specific authority should be identified.

Existing tools: As previously discussed, the OCLC Terminologies project (http://tspilot.oclc.org/resources/index.html) is attempting to bring together several vocabularies, including genre vocabularies, in an SRU searchable form, but at this point there does not seem to be a way to search all vocabularies at once. The OCLC FAST project page (http://www.oclc.org/research/projects/fast/) has a list of valid LCSH form subdivisions that may be downloaded in a spreadsheet.

Desired tools: A tool to identify genre information within the metadata and explicitly code it so that it can be used in faceted searching is desired. Where topical subject strings have not been split into subelements, use LCSH and FAST terminology to parse and code the subelements (see the discussion in **Section A—Topical Subjects**). For some records, this will reveal genre subelements where none were previously explicitly tagged. Where subjects are already split into subelements, check topical subelements for possible genre headings that have been incorrectly identified as topical. Match any genre elements and sublelements against one or more genre authority files. Where the metadata format supports it, add an indication of the source of the authority.

With the source of an authority identified (or matched to a vocabulary term) and crosswalks existing between the vocabularies, it should be possible to use the desired tool described in **Section A—Topical Subjects, Desired service: Consistent use of subject terminology on all the metadata records in an aggregation**. In addition to topics, this tool could help retrieve genre information from various vocabularies. Since it is not unusual to have multiple genre vocabularies within an aggregator, this could be a great aid to searching.

It should also be possible to generate at least high-level categories (books, photographs, etc.) where none exist, using clues from elsewhere in the record, such as physical description.

Comments: Genre is a vexing issue because the library world, particularly the Library of Congress, has traditionally avoided the use of genre headings with books, either not assigning them at all or mixing elements of form and genre in with subject headings. There were concerns about redundancy in non-book materials. For example, a picture is obviously a pictorial work, so the genre heading "pictorial works" is never added to "images" because users should know to limit to images if they want them. These subfields of the subject string were not even identified as genres in the MARC

format before 1999. When these older MARC records are mapped into another metadata format, the genre elements are mapped either into a general subject field or into a topical subelement, thus hiding them from a search based on a genre tag. LC has only recently begun moving genre information in some formats (moving image, radio, and soon music) into explicit genre tags in the MARC record, but in other domains (as well as vast legacy data), genre headings in MARC continue to be encoded as topics. There has been ongoing discussion over many years as to the difference between form and genre and whether that distinction is important. Because of the lack of leadership in this area, genre and form remain ill-defined, and many different groups, including divisions within the Library of Congress, have developed their own genre authority lists and their own practices for them. It would be theoretically possible to break form/genre facets out of LCSH strings and move them to the genre field. The FAST records available from the OCLC Terminologies project can be a source of information. However, since the Library of Congress did not begin issuing genre authority records (MARC coded 155) until 2007 and is far from completing this project, the genre information in FAST is quite incomplete. In some formats, such as MODS and MARC, there is a genre subelement within the subject element to which elements from LCSH may have been mapped. This approach should also work as long as both the genre subelement and the genre element are searched when a user performs a genre search.

Generating a genre element where none exists in the record would be very difficult, as this is not the kind of information one can usually extract from words in the text. How high a level should one go? Are genres such as "Book," "Pamphlet," or "Photograph," useful? Genres could be useful if they were applied consistently, but specific knowledge communities have their own detailed vocabularies and the aggregator would need to decide on one and map any existing terms to the target vocabulary. This information is seldom in the original metadata, and a tool to generate this data would need to be very sophisticated. Although it might be possible to design such a tool in the future, it may not be feasible at this time.

C. Names

Summary of desired services:

- Enable the user to retrieve all relevant items associated with a person or group
- Enable the user to retrieve all relevant items associated with a name regardless of the fullness or spelling of the person or group
- Enable names to be browsed by either last name or first name but displayed in natural order

Desired service: Enable the user to retrieve all relevant items associated with a person or group

Metadata support: Names should be contained within fields that identify them specifically as names, so that users may specify a search for a name. Names that are not in name fields, but are in titles, notes, and abstracts, could be programmatically extracted and added back into the metadata record in an appropriate field. Ideally, they should be added back in an authorized form with the appropriate authority identified (if supported by the metadata format).

Existing tools: OpenCalais (http://www.opencalais.com/) is a free service, but not open source (it is provided by Thompson/Reuters). The program is given text and it returns extremely verbose RDF (this can be done via an online form or API). The API version that can be downloaded from their site can produce microformats and may also give a simpler output format than the online form version. OpenCalais claims to recognize names, companies, movie titles, and other entities. Extracted personal names are in direct order. It does not recognize names that are in indirect order.

Stanford Natural Language Processor Tools (http://nlp.stanford.edu/software/index.shtml). The Stanford Named Entity Recognizer is an entity extraction program that examines text strings for words that might be personal or company names.

GATE, available on SourceForge (http://sourceforge.net/projects/gate), and is part of the suite of tools available with Greenstone (http://www.greenstone.org/).

ANAC (automated name authority control system) for the Levy Sheet Music Collection at The Johns Hopkins University.

The Perseus project developed its own named entity extractor optimized for Civil War–era names. Contact information for the developers can be found in the Mimno paper cited in the Literature section below.

Among commercial services, BasisTechnology's Rosette entity extractor (http://basistech.com/entity-extraction/?gclid=CIDWos-s0pUCFQkiagod4VGujA) is particularly strong in extraction in a multilingual environment.

Desired tools: After extracting named entities from the record, the tool should attempt to match names against NAF or other authority files, and import the controlled names back into the metadata record. Source should be recorded when using a metadata format that supports this approach. Matching names should have the source recorded while uncontrolled names would not. If the aggregation contains textual materials and the aggregator has access to them, names could be extracted from the full text. A complication is whether to map the names to a name field associated with creators or with subjects. It would be very difficult for a tool to tell to which field they belonged. Ideally, a search engine would allow name searches in both fields, as well as either creator name and subject name fields. If terms are mapped back into the metadata, it would be useful to identify the field content as machine generated, if supported by the metadata format. Applicable fields for extracting names would be title, abstract, subject.

Literature: DiLauro, T., G. S. Choudhury, M. Patton, and J. W. Warner. "Automated Name Authority Control and Enhanced Searching in the Levy Collection," *D-Lib Magazine* 7, no. 4 (April 2001), http://www.dlib.org/dlib/april01/dilauro/04dilauro.html.

Mimno, D., A. Jones, and G. Crane. "Finding a Catalog: Generating Analytical Catalog Records from Well-structured Digital Texts," in "Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries," JCDL 2005, Denver, Colorado, United States, June 7–11, 2005. JCDL '05. ACM, New York, New York: 271–280, http://doi.acm.org/10.1145/1065385.1065448.

Comments: This crucial field is required when applicable to enable the user to search and browse creator names. Names as subjects or names of persons depicted are often missing from metadata in photograph collections, particularly from the Library of Congress, or from metadata that originated from systems that did not distinguish name subject from topical subjects and lacked the granularity of MARC. Metadata formats that support the identification of a specific authority as the source of the name seldom contain this information, as it is difficult to map them from other metadata schemes, including MARC.

There has been a great deal of work on automated name extraction and disambiguation, but it has been concentrated in scientific literature, which tends to have very formalized citation practices and is usually associated with a very specific knowledge domain. A promising early tool, the Levy Sheet music tool (automated name authority control system, or ANAC), which extracted names from statements of responsibility, disambiguated, and matched to NAF, is no longer being used by its developers. It was used for a narrow body of materials, which facilitated the use of clues in NAF for fuzzy matching. It still took a fair amount of human intervention, and it proved difficult to scale. A more recent experiment took place at the Perseus Digital Library, where full-text documents relating to the U.S. Civil War were mined for personal and geographic names, which were matched against both NAF and locally constructed authorities. Matching against NAF was done manually with student labor, and unmatched headings were used to search in other sources to create a local authority file. There was an initial investment in determining pattern matching rules, but the developers stated that these rules scale well. Verifying the correctness of these automatic tags is significantly more expensive in terms of time and labor. However, the developers found that the quality of the initial automatic pass was sufficient to produce useful personal, corporate, and geographic subject headings. The developers are also developing an authority file based on their work, which would be of great benefit to any other project with overlapping interests.

How well would these tools do without human supervision in the authority matching? Determining the answer to this question would require testing. Exact machine matching to NAF would have little success. The experimental Virtual International Authority File (http://viaf.org/), housed at OCLC, contains not only NAF but much supplementary information as well, including additional variants, names of works, and other associated names. It also has a 930 field "alternate form of author's name" with the name in direct order, derived from usage information in bibliographic records. This file is planned to be available through an OAI interface, and there is already SRU access to the file. With this file it might be possible to use clues in the

metadata record to make a fuzzy match with better accuracy than could be made with NAF alone. For uncertain matches or no matches, it would probably be better to simply insert those headings back into the metadata record, preferably with an indication that they are machine-generated, where this is supported by the metadata format. In a further process, it might be possible to cluster these unmatched names to see if they are possible variants for the same person. If a limited human review was possible, perhaps only these possible clusters could be sent to a reviewer to settle on an authoritative form. Any work done at this level would be best shared with the originator of the metadata, so that the enhancement would not have to be redone when the records are reharvested.

The extraction processes do make a lot of mistakes, and the question of whether the clutter of irrelevant and inaccurate names is worth the effort would need to be calculated. The results of an informal experiment can be found in **Appendix B—Informal Test of Entity Extraction Software**. Ten records were taken from the Library of Congress photo and ephemera collections, and the title, abstract, and note fields were run through the OpenCalais viewer (http://sws.clearforest.com/calaisviewer). The OpenCalais viewer did identify some names accurately, but the rest of the names were either misidentified or completely missed. The same ten records were submitted to the Stanford Named Entity Recognizer service with mixed results. Some names and locations were identified accurately, but others were not.

Note fields are tricky, because many relate not to the individual item but to the collection or hosting institution (e.g., Library of Congress, George Grantham Bain). However, some records also have abstracts and other significant explanatory information in general note fields, so eliminating that field for matches could mean the loss of many significant names. Perhaps as a collection is preprocessed, collection-like names could be identified and suppressed, if this was considered desirable.

Although machine-generated names also generate a lot of noise, it must be remembered that even in human-created metadata, the tracing of names is not perfect. With the entrance of the commercial sector into this field and the high level of resources being poured into this area, it is perhaps time to take a serious look at this function. In addition to Basis Technology, tools include:

- Attensity (http://www.attensity.com/)
- Clarabridge (http://www.clarabridge.com/content_mining_platform_services.aspx)
- BusinessObjects (http://www.businessobjects.com/product/catalog/text_analysis/)
- Linguamatics (http://www.linguamatics.com/)
- Aerotext (http://www.lockheedmartin.com/products/AeroText/index.html)
- NetOwl (http://www.sra.com/netowl/)

Desired service: Enable the user to retrieve all relevant items associated with a name regardless of the fullness or spelling of the person or group

Metadata support: Names in an authorized form in appropriate metadata elements. If it is supported by the metadata format, the source of the name authority should be identified. In the future, it would be helpful if metadata formats carried the numeric identifier or URI of the named entity, rather than relying on the text string of the name itself as an identifier.

Existing tools: See the tools described in the previous section.

Desired tools: A tool that will effectively cluster like names, match them against NAF or other authorities, return the authorized form of the name, and identify the source of the name authority when possible in the metadata format. In aggregations using metadata formats that support authority attributes, these attributes may have their own authority control problems (e.g., NAF), and should either be normalized in processing or the tools should be able to cope with the variants.

Comments: This is similar to **Desired service: Enable the user to retrieve all relevant items associ-**

ated with a person or group, but dealing only with names already encoded as names or subjects in the metadata. Topical subject fields should be checked, as many names have been mapped to topical fields in the transformation process from another format, or are in formats that do not allow the type of subject to be specified, as in simple Dublin Core. These should be encoded as name elements before continuing with the processing. Usage of authorized forms of names is desirable to collocate names that may appear in different forms and to disambiguate persons with the same names. This is a difficult process, as name forms vary widely, and are often insufficient even for manual identification. Disambiguation will be a problem. The proposed Virtual International Authority File (http://viaf.org/) will be helpful here, as it might be possible to use clues in the metadata record to make a fuzzy match with better accuracy than could be made with NAF alone. Because many aggregators deal with records for photographs, gray literature, and ephemera, a high percentage will not be in LC NAF or in the Virtual International Authority File. Providers may have to develop some ancillary authorities which should be shared with the community at large. Projects that attempt to match every name to NAF have thus far relied on heavy manual intervention. Creating local authority files would also require human intervention.

In the long run, it is desirable to use unambiguous, machine-actionable identifiers instead of text strings to represent authorized name headings. However, development work needs to be done by the community at large before this technology can be fully implemented by service providers, and metadata formats would need a way to incorporate this data.

Desired service: Enable names to be browsed by either last name or first name but displayed in natural order

Metadata support: Where the metadata format supports additional granularity (e.g. MODS, EAD), divide names into separate subelements for the given name and surname, as well as date, titles, etc.

Existing tools: There are existing Java and PHP and probably other scripts that will split names into first and last name, although it is not clear how well they perform on names with separate prefixes or names in languages where the surname precedes the forename. Some Javascript examples can be found at http://homepage.ntlworld.com/kayseycarvey/jss4p1.html, http://osdir.com/ml/lang.moto.user/2003-03/msg00016.html, and http://ocw.mit.edu/NR/rdonlyres/Health-Sciences-and-Technology/HST-950JMedical-Computing Spring2003/080BDFD4-C0B9-4732-854F-5122C24D-B40A/0/englishname_java.txt. A PHP example may be found at http://www.php.net/manual/en/function.split.php.

Desired tools: It would be helpful if this data was available in NAF (similar to the way indirect geographic names are referenced in the 781 field of the MARC authority record) or the proposed Virtual International Authority File (http://viaf.org/). The latter does have a 930 field "alternate form of author's name" with the name in direct order, but it would also be helpful to have the name broken up into tags for the name parts, which is not currently supported by MARC21 (although it was previously supported by UNIMARC). The ideal tool would simply take the authorized name (either already in the record or discovered through one of the previously described processes), query NAF or the VIAF, return the split-up form of the name stored there, and translate it into the appropriate format to be inserted in the metadata. Lacking that infrastructure, a tool would need to be developed to split the name and insert the element tags.

Comments: Search tools can usually provide basic searches of names in direct or indirect order, but it is more precise and elegant to be able to display the name either way depending on context or the choice of the user. Browsing is particularly problematic for names that are entered in direct order. Some users may prefer to browse in direct order, while others may prefer indirect order. It would be useful to offer the users a choice in browsing mode.

D. Geographic Information

Desired service: Enable searching/ limiting by geographic place, either directly by name or by plotting the place on a map

Metadata support: Accurate use of the geographic element tags, sometimes supplemented by place names occurring in publication or creation information. Where the metadata format supports identification of the subject authority, this should be included to enable accurate identification of the place name, particularly when disambiguation is needed. Import geographic coordinates into the metadata record to support advanced mapping. For large places, include geographic coordinates for area rather than point.

Existing tools: The Geo-gazer processing tool developed for the DLF Aquifer American Social History Online project matches records that have a <subject><geographic> field against the USGS Gazetteer (http://geonames.usgs.gov/domestic/download_data.htm) and then takes other non-geographic subject fields into account in the search for a place name. The right-most words in subjects are often more geographically significant, and thus are more heavily weighted in the match. Weighting is also used to disambiguate places like New York, which has multiple levels of jurisdiction, or Washington, which refers to various places and is also a personal name. Populated places get a higher weight than civil districts, geographic features, and so on. Geographic coordinates are retrieved from the USGS Gazetteer service and added to the item records. This approach, which supports a map mashup, is currently hosted with the rest of the American Social History Online code on SourceForge (http://sourceforge.net/projects/dlf-aquifer/), but it can be downloaded via SVN and run as a separate project. It can also be downloaded separately from RubyForge at https://rubyforge.org/projects/geogazer/.

Desired tools: Tool to identify geographic names in meaningful fields such as title, abstract, notes, and unfaceted subjects, and match them to the data in the USGS Gazetteer database or other geographic databases. (See the discussion on splitting up LCSH headings into their applicable facets in preprocessing in **Section A—Topics.**) This tool would provide geographic elements for the tool to work with when they already exist in the subject string. For records with no geographic facet, attempt to match geographic names to title, abstract, notes, publication, or creation information. Most entity extraction tools also look for geographic names (and appear to be somewhat more successful with them than with personal names), but if existing tools are sufficient for geographic name identification, it should not be necessary to rely on the same kind of entity extraction tools that might be useful for other names.

Comments: A user searching for information associated with a place may want either information about the place itself, or want a resource that was published or created in a place. Simple searching/ limiting by place is ambiguous as to what is being represented. Sometimes the only place element in a record is in publication or creation date fields, but this can be misleading—sometimes the place

of origin is related to the intellectual content, but sometimes it is merely the location of a commercial publisher and has nothing to do with the content. It would be hard to programmatically determine when it is significant. Would this process be best applied only to certain types like images and ephemera? For books and music, the publication information generally applies to a commercial publisher and is likely not relevant to the content of the item.

Matching geographic names to an authority that has a hierarchy of place names could enable a feature to expand a query to the next larger jurisdiction or narrow it to a more specific place. Geographic coordinates can enable such services as allowing a user to draw a bounding box or polygon on a map and retrieve all things with points inside the polygon. They could also allow for users to produce statistical or other types of map views of the data.

E. Dates

Desired service: Accurately allow searching and limiting by date, including a time-line display

Metadata support: Inclusion of a machine-readable date field in a standardized scheme. Use of a key date indication in metadata formats which support it, which could be automatically added when there is only one date.

Existing tools: CDL Date Normalization Utility (http://www.cdlib.org/inside/diglib/datenorm/).

Desired tools: Further refinements of this tool to correct remaining problems, in order to work with a variety of metadata formats, and to distinguish between coverage dates and creation/publication dates where this is significant or desired.

Comments: Dates may appear in records in a variety of formats, and with or without associated textual elements such as "c" or "ca." In metadata formats that allow it, date encoding is often undeclared, and the date fields can have different meanings for different materials and the distinction between creation and publication is not necessarily stated in the record. As with places, dates may be associated with coverage of content as well as with creation or publication, and it is not clear in most search interfaces for which context the date information applies. Many metadata records contain dates of digitization or record update, which may further muddy the waters.

There is an excellent tool for working with these dates: the Date Normalization tool developed at the California Digital Library, also used by DLF Aquifer. It was originally designed for use with Dublin Core metadata. It works quite well, although it cannot generate a date where none was provided. Outstanding issues include dates in the title in the "mm/dd/yy" form when either 1900s or 2000s dates are possible; distinguishing "c" for circa and "c" for copyright (it currently interprets it as latter); the "baseball score" problem, which could interpret a sports score of 19-1 as a 20th century date; and four-digit identifiers which may be mistaken for dates. In the DLF Aquifer version (optimized for MODS), the date normalization software looks first at the originInfo dates in MODS records. A problem with using this field is that it sometimes coincides with the content (as in primary source materials), and sometimes the content is for a different period and the date refers to publication information (as in historical materials). For aggregators rich in primary source materials, this may not be a problem, but for aggregations largely composed of secondary sources, it would be much less reliable. There is also the issue of collections, which only give the date of digitization, although the date normalizer tries to account for those by removing the most recent dates from consideration. It looks elsewhere in the record for date information if it does not find adequate information in the creation or publication field, but it does not look at the temporal information in the subject fields, which tend to be overly broad (e.g., 19th century). Still, it might be useful to see if there is an obvious discrepancy between the collected dates and the creation/publication fields to identify whether the creation date and coverage dates are significantly different.

F. Title Information

Summary of desired services:

- Meaningful titles
- Search terms in the language of the aggregator's primary user base for titles originally in different languages
- Predictable and consistent alphabetical browse

Desired service: Meaningful titles

Metadata support: A brief title in an appropriate field in the metadata record.

Existing tools: The enhancement tool (see Foulonneau in the Literature subsection below) developed for a Committee on Institutional Cooperation (CIC) project and subsequently used in aggregation projects in the University of Illinois at Urbana-Champaign library, is currently undergoing extensive revision. This tool originally consisted of a series of XSLT style sheets that normalized both item- and collection-level metadata. For example, an OAI-PMH <provenance> node was added to each item-record to identify the OAI–PMH repository it was harvested from, and based on that information and other characteristics of the item records, information such as collection title was added to each item record. The search interface must be optimized to take advantage of the added data fields. It is not known at this time when the revision will be completed.

Desired tools: Adapt the above tool for a variety of metadata environments.

Literature: Foulonneau, M., T. W. Cole, T. G. Habing, and S. L. Shreeves, "Using Collection Descriptions to Enhance an Aggregation of Harvested Item-level Metadata," in "Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries," JCDL 2008, Denver, Colorado, United States, June 7–11, 2005: 32–41, http://doi.acm.org/10.1145/1065385.1065393.

Comments: Most metadata records have at least one title field. However, not all titles have turned out to be meaningful. Although problems with titles are well known and partly accounted for in search and browse algorithms, they are still worth noting.

When metadata is taken out of its native context (often a project on a specific topic or collection), it may leave titles that are cryptic or meaningless in isolation. While great efforts are being made to educate metadata providers and potential providers of this issue, the existence of much legacy data, particularly in databases, will make out-of-context titles a problem for some time to come. Information about context could be taken from a collection record for that collection, if one exists and is provided to the aggregator. Information may also be available in an element for a related title, particularly if the metadata format allows the identification of the type of relationship as a "host" or "series." These fields could be included in search algorithms and entity and topical extraction programs, along with the title. However, the scope of the collection-level information is often broader than

that of the item-level record, so it is rather a crude methodology to use. More promising are the methods proposed by Foulonneau et al. for ranking and weighting the collection data.

Desired service: Search terms in the language of the aggregator's primary user base for titles originally in different languages

Metadata support: Indication of language of the title in the record. For metadata formats that support it, an indication that a title is a translation, particularly a machine translation, would be desirable.

Existing tools: Google (http://www.google.com/language_tools) and Babel Fish (http://babelfish.yahoo.com/) open Web forms could be used for very small projects, but would be impractical for a large collection. The APIs currently built on these tools seem to be geared toward Web page translation. There is a program built to query the Google Translator hosted by CodeProject (http://www.codeproject.com/KB/IP/GoogleTranslator.aspx), and perhaps one could be built specifically to query XML data. There are several commercial software options, such as Babylon (http://www.babylon.com/), Systran (http://www.systransoft.com/translation/translation-products/), and many others.

Desired tools: In an English language context, a tool could flag non-English records (which hopefully would have an element identifying them correctly as such, although it may be feasible without it). A process would run them through a machine translation, and add the translation back to the record as another title element. For metadata formats that support it, some way of indicating that it is a translated title would be desirable. This approach should work equally well for other language environments.

Comments: Aggregated records may not all be in the same language. Such records will often fail a keyword search, which retrieves records in the main language of the aggregation. In addition, processes such as subject clustering and entity extraction may fail when the record is not in the expected language. It should be possible for the aggregator to identify these records and supply a translated title to support keyword searching and extraction processes. This would need to be done before any extraction or clustering processes are completed. Machine translation is far from perfect and often grammatically laughable, but it has improved greatly in the past few years and provides results good enough to be potentially useful for generating keywords for other processes. One may not want to display such a title to the user.

See Appendix B—Foreign Title Translation Test for an informal test of some metadata records with the language coded as Spanish. This test showed some pitfalls, such as place names that are not necessarily useful to translate, some words that the translator could not identify, and many records alleged to be in Spanish which did not actually have Spanish titles. However, many titles had acceptable translations, and as the information would be supplemental to the actual title and the process should be relatively simple to implement, it appears that the information gleaned by translation would do more good than harm.

Desired service: Predictable and consistent alphabetical browse

Metadata support: In metadata formats that support them, correct use of any mechanisms for indicating initial articles which are not to be filed upon.

Existing tools: Many programs (such as iTunes) already automatically identify and process initial articles in English.

Desired tools: For metadata formats that have a mechanism for identifying non-filing initial articles and characters, a tool could coordinate with language or language of title elements to use appropriate initial articles for that language. It would also need to indicate characters not to be filed upon, such as brackets. For metadata formats with no such mechanism, a tool could either ignore the non-sort elements within the browse function, or, alternatively, it could insert another title element into the

metadata record with the initial article stripped out. This approach would cause the record to show up in a browse both under the initial article and under the first significant word. This result is redundant, but perhaps more flexible and user-friendly.

Comments: There are many programs which, in an alphabetical browse, ignore the English initial articles, and a few also ignore Spanish or French initial articles. However, when different language titles exist in the same collection, these programmatic browses may be confused, such as mistaking the word "Los" in Los Angeles as an initial article to be ignored, or mistaking the French word "A" as the English initial article. Some metadata formats (e.g., MODS, MARC) have mechanisms to explicitly identify initial articles to be ignored in sorting, which may not be used when mapping records from a format with no such mechanism (e.g., Dublin Core). If the aggregation consists of records with the appropriate granularity, it may wish to remediate these by looking at the element for language (or better yet, language of title), matching against the appropriate set of initial articles, and demarcating the initial article appropriately so that it will function properly in a browse. While this is not foolproof (indeed many records have language inappropriately coded or not coded at all), it should in normal circumstances achieve a high level of accuracy. Coordinating with translation software and lists of geographic places may help to minimize the confusion of place names and numbers with non-sort articles.

Metadata formats which do not support this approach will need to rely on the algorithm that controls the browse function to create an effective browse list. It should be able to look at the language of title (or, lacking that, language) element to determine the language of the title and apply the appropriate set of initial articles. Where the language is lacking, it could default to the dominant language, or, if the software can determine the language by analyzing the words in the title, it could apply the appropriate set.

G. Type of Resource

Desired service: Accurately search and filter resources by type of material

Metadata support: Accurate and consistently coded fields indicating the type of the resource. Use of the explicitly coded fields for MIME type in metadata formats which support that element.

Existing tools: University of Michigan's DC Type Normalization style sheet is bundled with DLXS Open Source OAI tools. The transform works with Dublin Core records and creates DLXS bibclass records. The table of values the transform uses is available at http://www.oaister.org/docs/normal_types.txt. (JHOVE http://hul.harvard.edu/jhove/.)

Desired tools: A tool that will determine the type of resource, either by inferring it from the metadata record alone, or by an analysis of the digital file itself, taking into account not only the MIME type but the actual content. Services such as JHOVE can determine the MIME type of a digital object, but cannot directly work from the URL in the metadata record. The service provider would need to crawl the URLs to examine the raw item data and record the MIME type, and use this information to deduce the correct type.

Comments: This element is essential for enabling a user to search and filter by certain high-level material types such as text, image, and so on. There is currently very high compliance in this area in some metadata formats (e.g., MODS, MARC), but other formats (such as Dublin Core), may vary widely. There are currently some tools used by OAI institutions that may be of use here.

The metadata service provider could add a particular type programmatically across all records in a collection if it was clear that all items were of the same type. Collection documentation provided by the data provider could help in this regard. If the data provider uses types that are documented and used consistently, they could be mapped or clustered with like types, as in the University of Michigan DC type normalization table. The service provider could also sample records, but this approach risks missing a small number of records that might be assigned an inappropriate type. The service provider could try to determine the MIME type of the digital object and use this information to deduce the correct type by examining the raw item data, recording the MIME type, and using this to deduce the correct Internet media type. JHOVE supposedly has this capability, but it would need to be coordinated with the crawled bitstream in order to deduce the MIME type. This approach would also not be accurate in many cases, since many text and notated music items have been digitized as image files, and notated music and maps have been digitized as PDF files. Unless there is some programmatic way to discern, for example, an image of text from an image of music from other type of images, relying on the MIME type to deduce the correct resource type could lead to very misleading results.

H. Addressable Raw Object

Desired service: Facilitate Web 2.0 functionalities for end users to manipulate raw data

Metadata support: Where supported by the metadata format, some mechanism to determine when a URL is for the raw item rather than the item in context—preferably both URLs should be in the record with appropriate identification.

Existing tools: Protocols such as OAI-ORE and Asset Actions are being developed to address the need for direct manipulation of digital objects. However, it is too early to assess the impact these protocols may have on metadata remediation and aggregation.

Desired tools: When there is no URL in the record for the raw object, use a crawler to identify the raw object bitstreams and import the URL back to the record. Do not show the direct URL to the end user.

Comments: Some aggregator guidelines require that each record contain at least one URL that leads to the item in context. The reason for this requirement is that users are not directed to a page containing the item itself without any supporting context. In some metadata formats, it is also possible for data providers to provide a second URL and indicate that this is for the "raw object." By providing a link to the raw digital object in addition to an item in context, a user could take action on the bitstream by using tools provided by the aggregator without necessarily showing them the URL. The process should coordinate with fields for rights to prevent access where rights do not support such usage. This approach is very desirable for processing on the part of the service provider, but few data providers provide the second URL and even fewer indicate that the URL leads to the raw object.

A difficult category of materials is complex items embedded in a Web page. There may be multiple pages and multiple MIME types. This process would be more difficult to enable on such resources.

I. Rights

Desired service: Ability to convey to the user the rights and restrictions associated with a resource and to enable or deny add-on services appropriately

Metadata support: Consistent use of fields designated for rights information, preferably using a standard license and or rights expression language such as Open Digital Rights Language (ODRL).

Existing tools: None identified.

Desired tools: For records that do not have a rights statement, a tool to extract this information from the collection record is desired. Translate into machine-actionable rights when possible and insert that field—if a machine actionable field is already there, then translate it into a human-understandable field for the user. A tool to effectively use Creative Commons data would be particularly desirable.

Comments: Many metadata records lack explicit information on rights and restrictions, or the data providers only provide this information in the file or collection information or on their Web site. Information that is provided is usually not in machine-readable form. Users appreciate a human-readable rights statement so that they know up front what they can do with a digital object. A machine-actionable statement would be necessary for tools to be able to enforce restrictions on use. The community infrastructure here is emerging, mostly through Open Digital Rights Language (ODRL), which can express Creative Commons information. What is needed is authoritative standardized rights statements that are both actionable and eye readable, or mechanisms to resolve one to the other.

Sometimes rights data may be in collection-level records only. This information could be extracted and added to the item-level records, or it could be inherited from the collection record. A blanket opt out could be applied, or the lack of a rights statement can be made the equivalent of considering the resource public domain with unrestricted rights.

J. Identifiers

Desired service: Unambiguously identify a particular digital object

Metadata support: Use of metadata elements for identifier information such as the MODS and Dublin Core <identifier> fields.

Existing tools: While some work has been done in this area (e.g., DOIs, The Handle System, PURLS, NOIDs, OpenURL and related Link Resolver systems, proposed registries connected with RDA development), none of these systems provide a widely applicable solution to the identifier problem.

Desired tools: When a metadata record contains a standard identifier, a tool could match this identifier to a database of identifiers. The tool could check the metadata record and add missing information about that item that was retrieved from the database, such as authorized creator names, variant titles, subjects, year of creation, and so on. The tool could also do the reverse: when a standard identifier is lacking, it could search the database for a matching item and retrieve and insert the identifier into the metadata record. Taking either a standard identifier or identification information from title/author/publisher/date/edition fields, a tool could do a match against the database to identify potential duplicates.

Comments: Identifier types are diverse and relatively few records have identifiers that are meaningful outside of the local environment. Standard identifiers are not as applicable to the ephemeral and unique materials that make up many online collections, but could be useful for published materials and mass-produced materials such as sheet music. De-duplication of identical materials is already an issue, particularly in the case of sheet music, music, and commercial films. Sheet music plate numbers, for example, could be very useful to match and correctly identify plate numbers in records (usually in a note field) and move them to an identifier field, and to provide a source of additional information that might be added to the metadata records for music. This could be valuable for searching and de-duplication, since there is a wide range of variation in music cataloging practices and the same publication could be cataloged in radically different ways. Conceivably, a tool could match publisher, date, creator, edition, and title for other types of material. There could be varying degrees of fuzziness applied to the matching algorithm and different weights could be given to different types of matches to assign various levels of confidence as to whether the item is really a duplicate.

Glossary

Sources:

Authority Control: A Basic Glossary of Terms, http://ublib.buffalo.edu/libraries/units/cts/ac/def.html

Basis Tech, http://www.basistech.com/entity-extraction/

Boutell.com WWW FAQs, http://www.boutell.com/newfaq/definitions/mimetype.html

British Library. Redefining the Library: Glossary, http://www.bl.uk/aboutus/stratpolprog/redeflib/glossary/index.html

Digital Libraries Glossary, http://www.cs.cornell.edu/wya/DigLib/MS1999/Glossary.html

DLF Mission Statement, http://www.diglib.org/about/dlfmission.htm

FAST: Faceted Application of Subject Terminology, http://www.oclc.org/research/projects/fast/

Glossary of Terms, Aquifer Context, http://wiki.dlib.indiana.edu/confluence/download/attachments/24288/GlossaryVersion7x.pdf?version=1

Handle System, http://www.handle.net/

Inside CDL. NOID (Nice Opaque Identifier): Minter and Name Resolver, http://www.cdlib.org/inside/diglib/noid/

International DOI Foundation (IDF), http://www.doi.org/

MIC. Glossary of Cataloging & General Terms, http://gondolin.rutgers.edu/MIC/text/how/catalog_glossary.htm

National Library of Australia: Present Identifiers, http://www.nla.gov.au/initiatives/persistence.html

Online Dictionary for Library and Information Science (ODLIS), http://lu.com/odlis

PURLS, http://purl.oclc.org/

SRU (Search/Retrieval via URL), http://www.loc.gov/standards/sru/index.html

UIUC DLI Glossary, http://dli.grainger.uiuc.edu/glossary.htm

Vocabulary Definitions for the OCKHAM Reference Model, http://wiki.osuosl.org/display/OCKPub/ORMDefinitions

Term	Definition
AAT	Art & Architecture Thesaurus (AAT). A structured vocabulary for describing and indexing works of visual art and architecture. Initially developed by the Getty Information Institute, the AAT is made available through the Getty Research Institute. [ODLIS]
Aggregator	A service that gathers information published by different sources and organizes it under a common search interface. The aggregator may also license access to a collection of journals from many different publishers. [British Library]
API	Application Programming Interface (API). A set of functions, procedures, or classes that an operating system, library, or service provides to support requests made by computer programs. [Wikipedia]
Authority control	The procedures by which consistency of form is maintained in the headings (names, uniform titles, series titles, and subjects) used in a library catalog or file of bibliographic records by applying an authoritative list (called an authority file) to new items as they are added to the collection. Authority control is available from commercial service providers. [ODLIS]
Classification	A list of classes arranged according to a set of pre-established principles for the purpose of organizing items in a collection, or entries in an index, bibliography, or catalog into groups based on their similarities and differences, to facilitate access and retrieval. In the United States, most library collections are classified by subject. Classification systems can be enumerative or hierarchical, broad or close. [ODLIS]
Clustering	Data clustering is a data analysis technique that involves partitioning a data set into subsets with elements that share common traits. For example, semantic clustering is the clustering of objects based on a semantic proximity. [Wikipedia]
Collection-level information	Information provided by a metadata provider to describe a digital collection. This information (e.g., title of the collection, rights information) pertains to the collection as a whole, as opposed to metadata for each individual item.
Controlled names	See **Authority control**.
Controlled vocabulary	An established list of preferred terms used by a cataloger or indexer assigning subject headings or descriptors in a bibliographic record to indicate the content of the work in a library catalog, index, or bibliographic database. Synonyms are included as lead-in vocabulary, with instructions to see or use the authorized heading. For example, if the authorized subject heading for works about dogs is "Dogs," then all items about dogs will be assigned the heading "Dogs," including a work titled *All About Canines*. A cross-reference to the heading "Dogs" will be made from the term "Canines" to ensure that anyone looking for information about dogs under "Canines" will be directed to the correct heading. Controlled vocabulary is usually listed alphabetically in a subject headings list or thesaurus of indexing terms. The process of creating and maintaining a list of preferred indexing terms is called vocabulary control. [ODLIS]

Term	Definition
Data mining	Data processing using sophisticated data search capabilities and statistical algorithms to discover patterns and correlations in large pre-existing databases; a way to discover new meaning in data. [WordNet]
Disambiguate	Disambiguation is the process of resolving conflicts when different entities share the same name or label. For example, two or more people may share the same name, the same initials may refer to more than one organization, different concepts may be referred to by the same term, or different intellectual works may have the same title. Examining the name or label in context is one way to disambiguate a term.
DLF	The Digital Library Federation (DLF) is an international association of libraries and allied institutions. Its mission is to enable new research and scholarship for its members, students, scholars, lifelong learners, and the general public by developing an international network of digital libraries. [DLF Mission Statement]
DOI	The Digital Object Identifier (DOI) System is for identifying content objects in the digital environment. DOI names are assigned to any entity for use on digital networks. They are used to provide current information, including where they (or information about them) can be found on the Internet. Information about a digital object may change over time, including where to find it, but its DOI name will not change. [The International DOI Foundation (IDF)]
Dublin Core	*Dublin Core Metadata Element Set* (Dublin Core Metadata Initiative) (http://dublincore.org). A standard set of 15 elements (e.g., title, creator, subject), with optional qualifiers and community-specific extensions. All elements are optional and repeatable within an application profile used to structure data elements into records customized for specific audiences. Dublin Core is used to structure descriptive information about a resource and to map readily to other descriptive schema to facilitate sharing information across different metadata schemas and user communities. First developed in the mid-1990s, and originally intended for use in describing Web sites and Web pages, Dublin Core is now used also for describing physical and digital collections in museums, libraries, archives, and other repositories. [MIC glossary]
EAD	Encoded Archival Description (EAD). The EAD Document Type Definition (DTD) is a nonproprietary standard for encoding in Standard Generalized Markup Language (SGML) or Extensible Markup Language (XML) the finding aids (e.g., registers, inventories, indexes) used in archives, libraries, museums, and other repositories of manuscripts and primary sources to facilitate use of their materials. EAD was developed in 1993 on the initiative of the UC Berkeley Library and is maintained by the Library of Congress, in partnership with the Society of American Archivists. [ODLIS]
Elements	Portions of metadata that refer to distinct properties of an object. Elements are usually named "tags" in XML metadata. In MARC, elements are values coupled with MARC codes. In a table, elements would be labeled columns, along with their semantics. [Vocabulary Definitions for the OCKHAM Reference model]. In a library/metadata context, different elements will reside in different fields.
Entity extraction	The process of identifying names, places, dates, and other words and phrases that establish the meaning of a body of text from large amounts of unstructured data coming from sources such as e-mail, document files, and the Web. [Basis Tech]
Faceted searching	Faceted search enables users to navigate a multidimensional information space by combining text search with a progressive narrowing of choices in each dimension. [Wikipedia]

Term	Definition
FAST	Faceted Application of Subject Terminology (FAST). An OCLC program to adapt the LCSH with a simplified syntax to retain the very rich vocabulary of LCSH while making the schema easier to understand, control, apply, and use. The schema maintains upward compatibility with LCSH, and any valid set of LC subject headings can be converted to FAST headings. [FAST: Faceted Application of Subject Terminology]
Fields	An individual item of information in a structured record, such as a catalog or database record. [Digital Libraries Glossary]
Folksonomy	Folksonomy (also known as collaborative tagging, social classification, social indexing, and social tagging) is the practice and method of collaboratively creating and managing tags to annotate and categorize content. Folksonomy has become a popular term to describe the bottom-up classification systems that emerge from social tagging. [Wikipedia]
Granularity	The level of descriptive detail in a record created to represent a document or information resource for the purpose of retrieval. For example: whether the record structure in a bibliographic database allows the author's name to be parsed into given name and surname. [ODLIS]
Handle System	The Handle System is a general purpose distributed information system that provides efficient, extensible, and secure HDL identifier and resolution services for use on networks such as the Internet. It includes an open set of protocols, a namespace, and a reference implementation of the protocols. The protocols enable a distributed computer system to store identifiers, known as handles, of arbitrary resources and resolve those handles into the information necessary to locate, access, contact, authenticate, or otherwise make use of the resources. This information can be changed as needed to reflect the current state of the identified resource without changing its identifier, thus allowing the name of the item to persist over changes of location and other related state information. The original version of the Handle System technology was developed with support from the Defense Advanced Research Projects Agency (DARPA). [Handle System Web site]
Harvest	The process of gathering data from Web pages and other Internet sources and sending it back to a central site for indexing. In the Open Archives Initiative (OAI), metadata is harvested from distributed repositories, such as e-print servers and library catalogs. [ODLIS]
Infrastructure	The structural elements that provide the framework for an entire structure. The term has diverse meanings in different fields, but it is perhaps most widely understood to refer to roads, airports, bridges, and utilities. [Wikipedia]. For the purposes of this report, infrastructure refers to authority files, thesauri, vocabularies, registries, and the like that are publicly available in machine-readable form and their associated delivery services, that can be used to support various automated metadata validation and enhancement activities.
Knowledge domain	The content of a particular field of knowledge. [WordNet]
LCSH	Library of Congress Subject Headings (LCSH). A thesaurus of standard terms created by the Library of Congress that is used to locate resources on a specific topic. LCSH is used by most libraries in the United States.

Term	Definition
Legacy data	Information stored in an old or obsolete format or computer system that is, therefore, difficult to access or process. [BusinessDictionary.com] In a library and metadata context, legacy data usually refers to older cataloging and metadata (often in MARC format, but sometimes in stand-alone databases) that may or may not map well to newer metadata formats.
MARC	Machine Readable Cataloging (MARC). An international standard digital format for the description of bibliographic items developed by the Library of Congress during the 1960s to facilitate the creation and dissemination of computerized cataloging from library to library within the same country and between countries. By 1971, the MARC format had become the national standard for dissemination of bibliographic data, and by 1973, an international standard. [ODLIS]
MeSH	Medical Subject Headings (MeSH). The thesaurus of controlled vocabulary used by the National Library of Medicine (NLM) of the United States. MeSH subject headings are used in the NLM's MEDLINE database (available on the Web as PubMed), Index Medicus, and bibliographic cataloging records. MeSH headings are published in print by the NLM in an alphabetically arranged annotated list and in tree structures. [ODLIS]
Metadata Remediation	Correcting or improving existing metadata. [Glossary of terms, Aquifer context]
MIME type	Multimedia Internet Mail Extensions (MIME). MIME types are used to identify the type of information contained in a file. [Boutell.com WWW FAQs]
MODS	Metadata Object Description Schema (MODS). An XML schema developed by the Library of Congress for representing MARC-like semantics in the XML markup language. MODS can be used to carry selected data from MARC 21 records or for creating original resource description records according to a specification richer than Dublin Core, but less complex than full MARC. MODS cannot be used for the conversion of MARC to XML without loss of data (MARCXML was designed for that purpose). [ODLIS]
NAF	Library of Congress Name Authority File (NAF). The authority file for the Library of Congress. Most libraries in the United States base their authority work on this file. [Authority Control: A Basic Glossary of Terms]
NOID	The NOID software tool *mints* (generates) opaque identifiers and tracks information to help them remain unique, stable, and closely connected to the objects that they identify. These identifiers should be opaque enough to age and travel well, but should easily *resolve* (connect you) to objects and to their descriptions. [Inside CDL]
Normalization	A process by which data is transformed to make it more consistent. In a library/metadata setting, normalization is usually carried out on test strings and is often performed before text is processed in some way, such as searching or matching against another text string.
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). A specification that defines a mechanism for data providers to expose their metadata.
OAIster	A service that harvests collection data from a large variety of institutions. Data is made available to any interested end user for searching. A product of the University of Michigan Digital Library Production Service. [Glossary of terms, Aquifer context]
Ontology	A formal specification of how to represent the objects, concepts, and other entities that are assumed to exist in some area of interest, and the relationships among them. [UIUC DLI glossary]

Term	Definition
Open Digital Rights Language (ODRL) Initiative	The Open Digital Rights Language (ODRL) Initiative is an international effort aimed at developing and promoting an open standard for rights expressions. ODRL is intended to provide flexible and interoperable mechanisms to support transparent and innovative use of digital content in publishing, distributing, and consuming of digital media across all sectors and communities. [ODRL Web site: http://www.odrl.net/]
OpenURL	The OpenURL standard is designed to support mediated linking from information resources (sources) to library services (targets). A "link resolver" or "link-server" parses the elements of an OpenURL and provides links to appropriate services as identified by a library. A source is generally a bibliographic citation or bibliographic record used to generate an OpenURL. A target is a resource or service that helps satisfy user's information needs. Example targets include full-text repositories; abstracting, indexing, and citation databases; online library catalogs; and other Web resources and services. [Wikipedia]
Persistent identifiers	An persistent identifier is a name for a resource which will remain the same regardless of where the resource is located. Therefore, links to the resource will continue to work even if it is moved. [National Library of Australia: Persistent Identifiers]
Pre-coordinated LCSH strings	Library of Congress subject headings are structured as text strings containing various facets. In the MARC format, these facets were designed to be input in distinct subfields of the subject heading field but to display as a text string with the facets connected with hyphens. For example: Motion pictures—Germany—History—20th century—Dictionaries. Some metadata formats or legacy databases do not support the faceting of a single string, so the entire text string may be entered as a single subject field. Such subject heading strings are called precoordinated, as opposed to vocabularies such as FAST, which break the facets into separate fields.
PURL	Persistent Uniform Resource Locator. Functionally, a PURL is a URL. However, instead of pointing directly to the location of an Internet resource, a PURL points to an intermediate resolution service. The PURL resolution service associates the PURL with the actual URL and returns that URL to the client. The client can then complete the URL transaction in the normal fashion. In Web parlance, this is a standard HTTP *redirect*. [PURLS Web site]
RDF	Resource Description Framework (RDF). A family of World Wide Web Consortium (W3C) specifications, originally designed as a metadata data model, which has come to be used as a general method of modeling information through a variety of syntax formats. [Wikipedia]
Rights expression language	A syntax that provides information about how an object can be used, who owns the copyright, if it is in the public domain, and the like. [Glossary of terms, Aquifer context]
Search	To enter a query in an online textbox, submit and receive targeted results from a Web site's databases. [Glossary of terms, Aquifer context]
SKOS	Simple Knowledge Organisation Systems (SKOS). A family of formal languages designed for representation of thesauri, classification schemes, taxonomies, subject-heading systems, or any other type of structured controlled vocabulary. SKOS is built upon RDF and RDFS, and its main objective is to ease the publication of controlled structured vocabularies for the Semantic Web. SKOS is currently developed within the W3C framework. [Wikipedia]

Term	Definition
SRU	Search/Retrieve via URL (SRU). A standard search protocol for Internet search queries, utilizing CQL (Common Query Language), a standard query syntax for representing queries. SRW (Search Retrieve Web Service) is a companion protocol to SRU. The Library of Congress serves as the maintenance agency for these standards. [SRU, Search/ Retrieval via URL, (Library of Congress)]
Subfields	A division of a field. In the MARC format, fields are divided by subfield delimiters, such as the ones that divide the facets of a precoordinated LCSH string. [Glossary of Terms, Aquifer Context glossary]
Tagging	A term used to describe human indexing of material. In a library/metadata context, tagging may refer to users supplying keywords to Web resources (see "Folksonomy"), or it may refer to the semantic markup of text.
Thesaurus	A controlled vocabulary with a syndetic structure within a circumscribed subject field used to organize material or information. [UIUC DLI glossary]
Tools	Software utilities used to facilitate development and testing of software products and services. Tools can also refer to software that enables content consumers and content providers to perform specific activities (e.g., the UVA "Collector tool"). [Glossary of Terms, Aquifer Context]
Uncontrolled	Data which does not or is not known to conform to a controlled vocabulary.
Web 2.0	A living term describing changing trends in the use of World Wide Web technology and Web design that aims to enhance the creativity, information sharing, collaboration, and functionality of the Web. Web 2.0 concepts have led to the development and evolution of Web-based communities and hosted services, such as social-networking sites, video-sharing sites, wikis, blogs, and folksonomies. [Wikipedia]

Appendix A—Statistical Topic Model

The topic model works on a collection of text documents. It produces two things

- A list of topics that together describe the collection
- A list of the topics for each document in the collection

A self-contained package for David Newman's topic model code is available at http://www.ics.uci.edu/~newman/ (use the 'code' link). You can download it and run it using "run.sh".

The input is a file (docs.txt) that contain the filenames of each document in the collection:

```
example1/20000101.0001.txt
example1/20000101.0002.txt
example1/20000101.0003.txt
example1/20000101.0004.txt
example1/20000101.0005.txt
example1/20000101.0006.txt
example1/20000101.0007.txt
example1/20000101.0008.txt
example1/20000101.0009.txt
example1/20000101.0015.txt
... (etc.)
```

The topic model runs just using this input (docs.txt). This then produces two files: topics.txt and topicsin-docs.txt.

The first output file (topics.txt) contains the top words in each topic. In the example, this file is:

```
[t1] going thing home think lot school big job told smith ...
[t2] country nation war group camp government economy tutsi ...
[t3] millennium times square 2000 midnight celebration crowd ...
[t4] week league star point left look giant free return big ...
[t5] city york end firework small air million morning call ...
[t6] president national percent say campaign bradley political ...
[t7] team game season player coach play games yard run say football ...
[t8] problem y2k computer system 2000 saturday government ...
[t9] nyt york putin russia yeltsin russian times cox service ...
```

```
[t10] american century 000 sport book number million building ...
```

The second output file (topicsindocs.txt) contains the topics in each document:

```
<example1/20000101.0001.txt> t7 t4 t1
<example1/20000101.0002.txt> t6 t9 t3 t10
<example1/20000101.0003.txt> t3 t5 t1
<example1/20000101.0004.txt> t8 t5 t3
<example1/20000101.0005.txt> t10 t3 t6 t9
<example1/20000101.0006.txt> t3 t5 t8
<example1/20000101.0007.txt> t3 t5 t8
<example1/20000101.0008.txt> t3 t5 t2 t10
<example1/20000101.0009.txt> t3 t5 t9
<example1/20000101.0015.txt> t3 t1
... (etc.)
```

This second file is the "enhanced" metadata that users can access to find individual documents (e.g., we may likely label [t7] as "sports"), so searching for items tagged with "sports" would return `<example1/20000101.0001.txt>` (and others results down the list).

Note that automatically learned topics are usually more interpretable that the ones in the above toy example (which just contains 228 documents).

Appendix B—Informal Test of Entity Extraction Software

Test file consisted of 10 LC item records with no subjects or only geographic subjects from the Bain and Ephemera collections. An identifier is given for each record with the text of the field's title, abstract, and notes. The text of these fields for each record was pasted into the OpenCalais viewer (http://sws.clearforest.com/calaisviewer/), and the same text was submitted to a test iteration of the Stanford Named Entity (NER) Software.

Identifier	Text entered	OpenCalais Results	Stanford NER Results
oai:lcoa1.loc.gov:loc.pnp/cph.3a01286	Elijah W. Halford Half lgth., seated at desk, facing right. Title and other information transcribed from unverified, old caption card data and item. George Grantham Bain Collection (Library of Congress). No known restrictions on publication	Facility: Library of Congress Organization: Congress Person: George Grantham Bain, Elijah W. Halford Half lgth	ORGANIZATION: Congress PERSON: Elijah W. Halford, George Grantham Bain Collection LOCATION:

Identifier	Text entered	OpenCalais Results	Stanford NER Results
oai:lcoa1.loc.gov:loc.pnp/cph.3a01426	Run on 19th Ward Bank Crowd outside bank, New York City From George Grantham Bain, 82 Union Square East, N.Y. Photoprint by Bain News Service, N.Y.C. Title and other information transcribed from unverified, old caption card and item. George Grantham Bain Collection (Library of Congress). No known restrictions on publication.	Facility: Library of Congress Organization: Congress, Bain News Service IndustryTerm: outside bank Company: Ward Bank City: New York City, Union Square East Person: George Grantham Bain ProvinceOrState: New York	ORGANIZATION: Bain News Service, Congress, Ward Bank PERSON: George Grantham Bain, George Grantham Bain Collection LOCATION: Square East, N.Y.C., N.Y., New York City
oai:lcoa1.loc.gov:loc.pnp/cph.3a02680	Sacred bullocks, India G.G. Bain. restricted permission to use to be obtained from D.J. Culver. George Grantham Bain Collection (Library of Congress). This record contains unverified, old data from caption card. No known restrictions on publication.	Facility: Library of Congress Organization: Congress Country: India Person: George Grantham Bain	ORGANIZATION: Congress PERSON: D.J. Culver, George Grantham Bain Collection, G.G. Bain LOCATION: India

Identifier	Text entered	OpenCalais Results	Stanford NER Results
oai:lcoa1.loc.gov:loc.pnp/cph.3a02681	The Wind's Palace, Jeypore, India Permission to use to be obtained from D.J. Culver Title and other information transcribed from unverified, old caption card data and item. George Grantham Bain Collection (Library of Congress). No known restrictions on publication.	Facility: Library of Congress, Wind's Palace Organization: Congress Country: India City: Jeypore Person: George Grantham Bain	ORGANIZATION: Congress PERSON: D.J. Culver, George Grantham Bain Collection LOCATION: India, Jeypore
oai:lcoa1.loc.gov:loc.pnp/cph.3a02715	Coolie woman, Darjeeling, India This record contains unverified, old data from caption card. George Grantham Bain Collection (Library of Congress). No known restrictions on publication.	Facility: Library of Congress Organization: Congress Country: India Person: George Grantham Bain, Darjeeling-	ORGANIZATION: Congress PERSON: George Grantham Bain Collection LOCATION: India, Darjeeling
oai:lcoa1.loc.gov:loc.rbc/rbpe.00000700	Amherst Express. Extra. Williams and Amherst base ball and chess! Muscle and mind!! July 1st and 2d, 1859 These two contests mark the beginning of competition between student teams of the two colleges. Both were played in Pittsfield, Mass., and both ended in Amherst victories. The ball game lasted 3-1/2 hours, with a final score of 73 to 32.; Imprin Available also through the Library of Congress Web site in two forms: as facsimile page images and as full text in SGML.	SportsGame: chess Facility: Library of Congress Organization: Congress City: Pittsfield ProvinceOrState: Massachusetts Technology: SGML	ORGANIZATION: Congress, Amherst Express, SGML, Amherst PERSON: Williams, Muscle LOCATION: Mass., Pittsfield, Amherst, Imprin

Identifier	Text entered	OpenCalais Results	Stanford NER Results
oai:lcoa1.loc.gov:loc.rbc/rbpe.00000800	War department, Washington. April 20, 1865. $100,000 reward! The murder of our late beloved president, Abraham Lincoln, is still at large Broadside advertising reward for capture of Lincoln assassination conspirators, illustrated with photographic prints of John H. Surratt, John Wilkes Booth, and David E. Herold.; Lincoln, Abraham, 1809-1865—Assassination—Washington (D.C.); Boot Available also through the Library of Congress Web site in two forms: as facsimile page images and as full text in SGML.	Facility: Library of Congress Organization: Congress, War department IndustryTerm: advertising reward Currency: USD City: Washington Person: John H. Surratt, John Wilkes Booth, Abraham Lincoln, David E. Herold ProvinceOrState: Washington Technology: SGML	ORGANIZATION: Congress, SGML PERSON: John Wilkes Booth, John H. Surratt, David E. Herold., Abraham Lincoln, Boot, Lincoln LOCATION: D.C., Washington, Lincoln, Abraham
oai:lcoa1.loc.gov:loc.rbc/rbpe.00100200	To the friends of our country. [Alabama 1824?] Regarding Andrew Jackson's land speculation.; Jackson, Andrew.; Imprint 2. Available also through the Library of Congress Web site in two forms: as facsimile page images and as full text in SGML.	Facility: Library of Congress Organization: Congress Person: Andrew Jackson ProvinceOrState: Alabama Technology: SGML	ORGANIZATION: Imprint, Congress, SGML PERSON: Andrew Jackson LOCATION: Alabama, Andrew., Jackson
oai:lcoa1.loc.gov:loc.rbc/rbpe.00100400	Report of the committee appointed to examine the state bank. Jan. 8th, 1827 Available also through the Library of Congress Web site in two forms: as facsimile page images and as full text in SGML	Facility: Library of Congress Organization: Congress IndustryTerm: state bank Technology: SGML	ORGANIZATION: Congress, SGML PERSON: LOCATION:

Identifier	Text entered	OpenCalais Results	Stanford NER Results
oai:lcoa1.loc.gov:loc.rbc/rbpe.00100500	An abstract of the census of Alabama, taken in 1827 Census; Imprint 2 Available also through the Library of Congress Web site in two forms: as facsimile page images and as full text in SGML.	Facility: Library of Congress Organization: Congress ProvinceOrState: Alabama Technology: SGML	ORGANIZATION: Imprint, Congress, SGML PERSON: LOCATION: Alabama

Appendix C—Foreign Title Translation Test

The titles for fifty metadata records from two American Social History Online collections, for which the language field indicated that the titles were in Spanish, were selected for title translation testing. The title translation testing consisted of entering the original titles into the Google translation tool and recording the results of the translation. The results of the title translation testing are provided in the table below. It is worth noting that some original titles have nonuseful or erroneous elements in them, and the titles were not edited before they were submitted to the tool. Also, despite the language indication in the metadata record, certain selected titles are not in Spanish.

ID	Original Title	Translated Title
1	Instruccion formada en virtud de real orden de S.M., que se dirige al señor comandante general de provincias internas don Jacobo Ugarte y Loyola para gobierno y puntual observancia de este superior gefe y de sus inmediatos subalternos	Instruction formed under royal command of His Majesty, which is addressed to Mr commander general internal provinces don Jacobo Ugarte and Loyola for government and timely compliance with this higher GEF and their immediate subordinates
2	Ynstrucciones y reglamentos de Yndias	Ynstruccion and regulations Yndias
3	El cólera Asiático: reseña sobre esta epidemia e instrucciones hijienicas para evitarla : comprende la cartilla del doctor Primavera, los preceptos hijiénicos de la Junta de Sanidad de Madrid i las recetas del doctor Castañé	Cholera Asia: review on this epidemic and instructions hijienicas to prevent it: the book includes doctor's Spring, hijiénicos the precepts of the Board of Health Madrid i prescriptions from the doctor Castañé
4	Tarifas, reglamento i clasificacion para el trasporte de carga en Los Vilos, Coquimbo i Huasco	Rates, rules i clasificacion to transport cargo in Los Vilos, Coquimbo i Huasco

ID	Original Title	Translated Title
5	Parallel histories [electronic resource]: Spain, the United States, and the American frontier = Historias paralelas : España, Estados Unidos y la frontera Americana	Parallel histories [electronic resource]: Spain, the United States, and the American frontier = Parallel Stories: Spain, the United States and the American border
6	Tirana Española	Tirana Spanish
7	La pasadita, a satirical Mexican song	The pasadita, a satirical song Mexican
8	Emblemas de la libertad y de la humanidad [graphic]: La Cruz Roja, Madre de todas las naciones	Emblems of freedom and humanity [graphic]: The Red Cross, Mother of all nations
9	La marina britanica en la guerra	The British navy in the war
10	Tirana Española Externallinks	Tirana Spanish Externallinks
11	Map of America by Diego Ribero 1529	Map of americas by Diego Ribero 1529
12	[Map of California shown as an island]	[Map of California shown as an island]
13	Mapa de una parte de la America Septentrional	Map of part of the Northern americas
14	Mapa de la America Septentrional dividido en dos partes: En la primera se descriven sus provincias segun los derechos que piensa tener a ellas la corona de Francia: en la segunda, segun las pretensiones de la Inglaterra	Map of the americas Northern divided into two parts: The first descrive their provinces in the rights that they intend to take the crown of France: in the second, according to the pretensions of England
15	Descripcion de la costa de Tierra Firme desde el Rio de la Empalizada hasta Cavo de Clara: Por las latitudes y longitudes de Dn. Bartolome de Rosa	Description of the coast of Tierra Firme from the Rio de la fence until Cavo Clare: For latitudes and longitudes of Dn. Bartolome de Rosa
16	Plano del archipielago de Clayocuat situada su boca mas O. llamado Puerto de Sn. Rafael por los 49º 20′ de latd. N. y en la longd. de 20º 55′ y la mas E. nombrada de Clayocuat por los 49º 7′ de la misma especie y 20º 22′ al O. del meridiano de Sn. Blas [?]conocidas sus bocas y descubiertos todos sus brazos e islas interiores por el Thente. de Navio de la Rl. Armada Dn. Francisco de Eliza Comandte. del Paguebot de S.M. nombrado San Carlos y Goleta Sta. Saturnina (alias la Orcasitas) en este presente año de 1791	Drawings of the archipelago located Clayocuat of his mouth more O. called Port of Sn. Rafael by 49 º 20′ of latd. N. and the Longde. 20- º 55′ and more E. Clayocuat appointed by the 49 º 7′ of the same species and 20 º 22′ to O. the meridian of Sn. Blas [?] Known their mouths and discovered all his arms and islands by the interior Thente. Navio of the Rl. Navy Dn. Francis Eliza Comandte. Paguebot of the S.M. Goleta appointed San Carlos and Sta. Saturnino (aka the Orcasitas) in this current year of 1791

ID	Original Title	Translated Title
17	Plano de la Bahia de la Ware y entrada de Filadelfie	Flat Bay Ware and the entry of Philadelphia
18	[Mapa del Golfo y costa de la Nueva España: desde el Río de Panuco hasta el cabo de Santa Elena ...]	[Map and the Gulf coast of New Spain, from Rio de Panuco to Cape St. Helena ...]
19	Descripcion geographica de la parte que los españoles poseen actualmente en el continente de la Florida del Del Dominio en que estan los ingleses con legitimo titulo solo en virtud del tratado de pases del año de 1670 y de la jurisdicion que indevidamente an ocupado despues de d[ic]ho tratado en que se manifiestan las tierras que usurpan y se definen los limites que deven prescrivirse para una y otra nacion en conformidad del derecho de la Corona de España	Geographer description of the party who currently possess the Spaniards on the continent of Florida's domain that are the English title only with legitimate under the treaty passes the year of 1670 and the jurisdiction that indevidamente an occupied after d [ic] ho treaty to express usurping land and defines the limits deven prescrivirse for another nation and in accordance to the right of the Crown of Spain
20	Plano y costa de la Palisada o de Misipipi zituada, su entrada o Cabo de Lodo en 29 gs. 17 ms. de lattud. norte y en longd. de 385 gs. 3 ms. segun Tenerife	Map and expense of Palis or Misipipi zituen, check or Cape Mud in 29 gs. 17 MS. of lattud. North and Longde. GS-385. 3 MS. according Tenerife
21	Descripsión de la costa de la Luciana y entrada en el Río de Micisipi con sus zonds. y baxos, nuebamte. correjido y enmendo por los pilos. de la Armada, el año 1769	Descripsión the coast of Luciana and entry into the River Micisipi with their zonds. and Bax, nuebamte. correct and amended by the batteries. of the Navy, the year 1769
22	Plano del desembocadero del Río Misipipi en el seno Mexicano con parte del territorio de la Movila, el qual incluien los Franceses en la provincia que han nombrado, la Luisiana	Map desembocadero River Misipipi within Mexican territory with part of mobility, qual including the French in the province that have appointed, LA
23	Plano. I descripcion de la costa, desde el Cavo Cañaveral, hasta cerca de la boca de la Vir[g] inia: contando, costa de Florida, Georgia y Carolinas del S, y N, con todos sus puertos, este[ros ...]letas, baxos, islas y rios; segun las vlti[mas not]icias, hata [sic] oy Octubre de 1756	Plano. I description of the coastline from the Cavo Canaveral, until near the mouth of Vir [g] INI: counting coast of Florida, Georgia and the Carolinas S and N, with all its ports, this [ros ...] Leto, Bax, islands and rivers, according to the vlti [not more] ICI hat [sic] oy October 1756
24	Descripcion de la Bahia de Santa Maria de Galve, y Puerto de Sn. Miguel de Panzacola con toda la costa contigua y las demas bahias que tiene en ella, hasta el Rio de Apalache	Description of the Bahia de Santa Maria de Galve, and Port of Sn. Miguel de Panzacola with the entire coast and the other adjacent bays that it takes, until Rio de Appalachian

ID	Original Title	Translated Title
25	Plano de la bahia de Pansacola	Map of the bay of Pansacola
26	Plano numero 1. de la barra, y Rio de San Juan desde su entrada hasta dos millas mas arriba del paso de San Nicolas, manifestandose en su curso todos los baxos, sacatales, caños, y ys. las que comprehende, y tambien la de la barra chica, situacion do los reductos, y colocacion de los barcos para su defenza, y caminos que deven tomarse para la retirada los defensores	Plane number 1. of the bar, and Rio San Juan since its entry to two miles above the passage of San Nicolas, manifesting itself in all its course Bax, sacatales, pipes, and ys. which comprehende, and also that of the bar girl, a situation do the holdouts, and placement of ships for its defence, and roads that deven taken for the withdrawal defenders
27	Bahia de Tampa	Tampa Bay
28	Plano de la ciudad y puerto de San Agustin de la Florida	Map of the city and port of San Agustin, Florida
29	Plano del Pto. de la Movila situado en la latd. N. de 30º 10′ tomado á los Ings., el día 14 de marzo de 1780	Map Pto. of mobility in the latd. N. 30º 10′ took the Ings., on March 14, 1780
30	La Luisiana cedida al Rei N. S. por S. M. Christianisima: con la Nueva Orleans, è isla en que se halla esta ciudad. Construida sobre el mapa de Mr. d'Anville	The LA ceded to Rei N. S. S. M. Christianisima: with the New Orleans, è island that is this city. Built on the map of Mr. d'Anville
31	[Map of Las Ormigas Grant, Sabine and DeSoto Parishes, Louisiana]	[Map of Las Ormigas Grant, Sabine and DeSoto Parishes, Louisiana]
32	Mapa topográfico de la provincia de Texas	Topographic map of the province Texas
33	Guía de las regiones de trabajos agrícolas en los estados del oeste	Guide regions of farm work in the western states
34	Derrotero hecho por Antonia Vélez y Escalante, misionero para mejor conocimiento de las misiones, pueblos de indios y presidios que se hallan en el Camino de Monterrey a Santa Fe de Nuebo Mexico	Route by Antonia Velez and Escalante, a missionary for better understanding of the missions, presidios and peoples of Indians who are in the Way of Santa Fe de Monterrey in Mexico Nuebo
35	[Map showing Caribbean area including West Indies and Gulf of Mexico	[Map showing Caribbean area including the West Indies and Gulf of Mexico

ID	Original Title	Translated Title
36	Mapa maritimo del Golfo de Mexico e islas de la America: para el uso de los navegantes en esta parte del mundo, construido sobre las mexores memorias, y observaciones astronomicas de longitudes, y latitudes	Map sea of Gulf of Mexico and islands of the americas: for use by sailors in this part of the world, built on mexores memories, and astronomical observations of heights and latitudes
37	Descripcion de las costas, islas placers, i bajos delas, Indias Occidentales	Description of the coasts, islands pleasures, i low delas, West
38	Descripcion de la costa de Tierra Firme desde el Cavo de la Agusa hasta la Barra de Palmas diga de la Trinidad: Con todas las yslas, bajos, arresifes & ca. Leho por las latitudes y longitudes de Dn. Bartolome de la Rosa	Description of the coast of Tierra Firme Cavo from the Agusa until Bar Palmas says of the Trinity: With all yslas, low arresifes & ca. Leh by latitudes and longitudes of Dn. Bartolome de la Rosa
39	Mapa y plano del Seno Mexicano: Contodas las costas, de tierra firme ẽyslas de barlovento consus adyacentes, recopiladas, sus-latitudes y longitudes en el puerto de la Havana con junta de primeros y segdos. pilotos de la esquadra y segun el neuvo padron	Map and flat Breast Mexican: Contodas the coast of mainland yslas windward consus adjacent compiled, their-latitudes and longitudes in the port of Havana with the board first and segdos. esquadra the pilots and the second roll neuvo
40	Mapa, que comprende la Frontera, de los Dominios del Rey, en la America Septentrional	Map, which includes the Border, the domains of the King, in the Northern Americas
41	Mapa de toda la frontera de los dominios del rey en la America	Map of the entire border of the king's dominions in the americas
42	Carta general de la República Mexicana	Charter general of the Mexican Republic
43	Mujeres en mi Vida, Film Poster for	Women in my Life, Film Poster for
44	(title unknown)	(Title unknown)
45	Calaca Huesuda—El Día de los Muertos Exhibition	Calaca bony—The Day of the Dead Exhibition
46	El Año de los Deiz [sic] Millones, Announcement Poster for	The Year of deiz [sic] Millions, Poster for Announcement
47	January Calendar	January Calendar
48	Salvador Allende Memorial Reading—Pablo Neruda, Announcement Poster for	Salvador Allende Memorial Reading—Pablo Neruda, Poster for Announcement

ID	Original Title	Translated Title
49	May Calendar	May Calendar
50	Galería Calendario Exhibition	Exhibition Gallery Calendar