Finding a Catalog

Generating Analytical Catalog Records from Well-Structured Digital Texts

David Mimno
Tufts University
Perseus Project
Medford, MA 02155
david.mimno@tufts.edu

Alison Jones
Tufts University
Perseus Project
Medford, MA 02155
alison.jones@tufts.edu

Gregory Crane
Tufts University
Perseus Project
Medford, MA 02155
gregory.crane@tufts.edu

ABSTRACT

One of the criticisms library users often make of catalogs is that they rarely include information below the bibliographic level. It is generally impossible to search a catalog for the titles and subjects of particular chapters or volumes. There has been no way to add this information to catalog records without exponentially increasing the workload of catalogers. At the same time, well-structured full-text XML transcriptions of printed works are becoming increasingly available. This paper describes how existing investments in full text digitization and structural markup combined with current named-entity extraction technology can efficiently generate the detailed level of catalog data that users want, at no significant additional cost. This system is demonstrated on an existing digital collection within the Perseus Digital Library.

Categories and Subject Descriptors

H.3.7 [**Information Systems**]: Information Storage and Retrieval—*digital libraries*

Keywords

analytical cataloging, information extraction, library automation

1. INTRODUCTION

Researchers at the Perseus Digital Library (PDL) have spent the past four years developing a 55 million word collection of works from the period of the American Civil War. This collection now represents approximately three hundred bound volumes, proofread to a high level of accuracy and carefully tagged with XML conforming to the guidelines of the Text Encoding Initiative (TEI). As a digital collection, the Perseus Civil War-era documents are a significant resource, representing a significant investment of time and resources. However, from the perspective of any academic or even public library, the collection is not large. The physical books fit comfortably along three walls of a small room. The catalog records for this collection, when composed to the level of specificity customary to works of this nature, would barely be noticed in a typical library's online public access catalog (OPAC). In this paper, we will show how we were able to automatically generate 60,000 catalog records for distinct sub-sections of these documents based on the inherent structure of the XML documents combined with named entity extraction systems. The resulting catalog better reflects the number of intellectual works that have been digitized and organized in the creation of this collection.

The benefits of digitization such as searchability and wide dissemination are sufficient in themselves to justify the expense of creating digital documents. However, digitization can also have profound impacts on the quality and quantity of data available within a library catalog. Much research has been done recently on automatic metadata generation based on digital documents, often loosely structured web pages [15, 34]. In this paper, we are exploring automatic metadata generation in a large, well-structured corpus. In particular, we will focus on identifying "works within works" such as chapters within books or articles within serials (also called "analytical" or "component" cataloging), identifying people, places and other named entities that appear in those sections of documents, and following cross-references between documents. In the past, cataloging in these areas has been strictly limited by cost considerations. Although the average cost of cataloging has been declining steadily over the past several decades, researchers at the Iowa State University library estimate that it is currently $16.25 per title [23]. Analytical cataloging in particular has been constrained by practical considerations. Cataloging a ten volume work as ten individual records significantly increases the cost of cataloging those items. In addition to the cost of additional cataloging, the decision of how to catalog multi-volume works has implications for the physical arrangement of books in the library. Do patrons expect to find an individual volume under the classification of the volume itself, or alongside the other volumes in the work? Finer subdivisions, such as chapters within a single volume, are rarely reflected in a typical library catalog beyond a brief table of contents [30].

Digital documents, particularly XML documents, have a very different relationship to the catalog. The very nature of a well-structured XML document makes it easy to describe the parts of a work down to a very narrow level of specificity. Moreover, it is practical to display small segments of larger works and "collocate" related document sections in the online digital library interface.

JCDL'05, June 7–11, 2005, Denver, Colorado, USA.

In this paper, we will first describe the history and current state of the technologies that underly our catalog generation system. We will then introduce the Perseus Civil War-era collection and the automatic named entity extraction systems that have been developed to annotate it. Next we will describe the process of automatically generating catalog records from the collection. Finally, we will explore and evaluate different interfaces for the automated catalog records and consider the implications of large-scale automatic catalog generation.

2. RELATED WORK

2.1 Analytical cataloging

Analytical cataloging involves creating records for parts of a work for which there also exists a catalog record for the whole. One common form of analytical cataloging is component cataloging in serials, where every article in a journal has a catalog record as well as the journal itself. Although component cataloging was practiced in the nineteenth century, it has become extremely rare because it is expensive and often duplicates records in commercially available abstracting and indexing services [27]. Nevertheless, many researchers have emphasized the value of analytical catalog records in an integrated library catalog environment. Zahiruddin Khurshid reported on a project to provide analytical MARC records for electronic journals on CD-ROMs in the collections of King Fahd University of Petroleum and Minerals [18]. Although the process was time-consuming, the resulting catalog records were found to be very useful. Users who would otherwise have been unaware of the availability of those journals were able to find them easily. Khurshid also notes that maintaining the same information outside of the catalog, for example on a separate web page, was not nearly as helpful as records that were integrated into the catalog. As early as 1985, Herbert H. Hoffman anticipated that computerized catalogs would be capable of handling "work-level" records as well as the typical "item-level" records [10]. Hoffman has since successfully implemented such a system at Santa Ana College [11]. He has observed that analytical cataloging increases the recall of title searches, as well as offering more specific subject headings [12].

2.2 Metadata extraction

There is a large body of research showing that automatic metadata extraction systems can produce accurate, reliable metadata. Researchers at OCLC created Scorpion, a system that automatically assigns DDC numbers to Web documents, which was eventually integrated into OCLC Connexion and WebDewey [4]. Research conducted by the Metadata Generation Research Project at the School of Information and Library Science at the University of North Carolina at Chapel Hill found that automatic metadata generation can often produce results of similar quality to those produced by professional metadata creators [8].

Research that focuses on the use of natural language processing and machine learning techniques have shown comparatively precise and relevant metadata [9, 7, 33]. Machine learning and natural language techniques have also been used for named entity recognition in digital library collections. The CLiMB (Computational Linguistics for Metadata Building) project at Columbia University has worked towards creating a named entity tool that can "identify references to a single art object (e.g. a particular building) with high precision in text related to images of that object in a digital collection" by matching an authoritative list of art objects with variants of these objects in the text [3]. They have also created tools to help image catalogers in the semi-automatic creation of metadata for images [2]. Similar work with authority lists and name disambiguation was done by researchers at the Digital Knowledge Center of the Sheridan Libraries at Johns Hopkins University. They created an automated name authority control system (ANAC) that was developed to identify Library of Congress (LC) authorized names against the names in the descriptive metadata of the Lester S. Levy collection of sheet music [26]. A number of projects have also explored the automatic creation of subject and keyword metadata [13, 16].

Researchers at Syracuse University's Center for Natural Language Processing have developed a tool called MetaExtract that automatically assigns Dublin Core and GEM Metadata to educational resources using extraction techniques from their natural language processing research. They found that except for the Title and Keyword elements, professional educators found no significant differences between the quality of automatic and manually entered metadata [34]. The Infomine Scholarly Internet Resource Collection uses expert intervention in the creation and checking of metadata, but they have developed a number of automatic web crawlers that generate thematic data about websites including keyphrases and LCSH [15].

2.3 Authority list creation

In order for effective metadata creation based on named entity recognition to occur, an appropriate controlled vocabulary must exist. Name authority files are of particular importance for disambiguation and can enhance access to large digital collections. In producing analytical component catalog records for the Civil War-era collection, we frequently were not able to align extracted named entities with existing authority files such as the Library of Congress Name Authority File because they do not focus on the same historical period. The creators of one digital historical collection suggest that one of the greatest challenges in digitizing and transcribing historical materials is "the high degree of personal name variation which demands authority control for insuring accurate transcription and effective discovery of the descriptive or intellectual content contained in these primary resources." [5] As a result, we are researching ways to create authority lists for historical named entities and share the results of our authority work with other institutions.

While many suggestions have been proposed no standard solution has yet been implemented. In 1995, NACO (Name Authority Cooperative Program) was formed by the LC and other partners, which allowed select institutions to add records to the LCs Name Authority File. One group of researchers proposed the use of a new standard number for author names, the INSAN (International Standard Author Number), that would serve as a general link to all variants of an author's name in different national authority record files [28]. The FRANAR (Functional Requirements and Number of Authority Records) Working Group is currently examining the "functional requirements of authority records" and the possible creation of an ISADN or International Standard Authority Data Number as a way of linking together disparate authority records [25].

Other research has focused on the creation of international authority files. The European LEAF (Linking and Exploring Authority Files) Project has attempted to create a framework that will support international collaborative work in the area of name authority control. The LEAF project reported that "a common format is not yet available for the exchange of authority data within the archival community or for the cross domain exchange of such data between libraries, archives, museums and related institutions." [17] OCLC is also currently working on a Virtual International Authority File project with the LC and the German National Library to test automatic linking of LC-NAF and German Personal Name Authority File through OCLCs matching algorithms [32]. In a recent article, the LCs Barbara Tillett has also suggested the possible creation of a "virtual union authority file" that would link all national authority files to a central server, each authority file would be hosted locally and this server would simply harvest data as requested by the user [31].

3. THE CIVIL WAR-ERA COLLECTION

The United States nineteenth century materials collection in the PDL represents a test bed to explore:

1. Surveying document types and genres. Our focus has been upon historical materials with numerous references to people, places and organizations but we have literary materials, prose and poetry, in preparation.

2. Exploring markup schemes. We are surveying typical document structures and searching for particular patterns that warrant markup: examples include indices, short phrase summaries at the start of chapters, block quotes with and without modern citations, and a range of embedded documents (such as quoted letters).

3. Evaluating services. Data structures are important in that they enhance the services that digital library systems are able to provide. We encode structural markup in XML so that the mediating system can accurately display the chapters, sections and other units of a document. Digital library systems such as CiteSeer [1], on the other hand, do a surprisingly good job mining textual citations from unstructured texts. What services should digital libraries support and how can these services be made feasible?

4. Comparing our results with other digital library systems and collections. We create our own digital resources so that we can try out numerous organizational schemes without offense to others who have carefully arranged their own resources in the way they wish. We compare what we can and cannot do when we have complete freedom to what we can do when we are constrained by what materials are available.

5. Identifying and developing authority lists for historical materials. Many of the people, organizations and events that are referred to in our documents are of minor historical importance. Many of the places referred to either no longer exist or have new names. We do not know of any existing electronic authority lists that provide comprehensive coverage for named entities in our cultural heritage documents. We identify potential print sources for authority lists such as encyclopedias. Where possible, we convert these into usable electronic resources.

Most of our effort has concentrated on named entity identification and builds on earlier work with other collections. Our approach falls mid way between a classic computer science approach (which aims at completely general tools that can work with any corpus from any domain) and traditional editing (in which human beings determine all annotation by hand). We have described this intellectual space as corpus editing. Corpus editors apply automated tools to corpora that are finite but far too large for manual processing. Corpus editors aim to provide statistical measures of accuracy for the corpus as a whole rather than checking every decision by hand.

At present we identify roughly 5 million named entities in a corpus of 55 million words – one named entity for every eleven words of running text. Most named entities span more than a single word (e.g., "Boston, Ma.," "George Washington" and "the New York Herald" count as one named entity each). The density of coverage achieved by the system is substantial.

The architecture of the system is fairly standard and breaks down into two basic parts. In the first stage, we maximize precision and mine the documents for well marked entities. In some cases, we can use an external authority list to check for plausible patterns. Thus, we can scan text for patterns such as POSSIBLE PLACE NAME + DELIMETER + REGION (e.g., "Boston, Ma.," "Berlin, Germany"). This works well in many cases but is not foolproof. In some cases, phantom patterns emerge: there is a "Philadelphia" in New York, for example, and the string "Philadelphia, New York" can erroneously resolve to a reference to that town rather than to two major US cities. With historical documents, of course, problems are greater, as new towns are founded and existing place names change over time: "Berlin, Md" uniquely designates different towns in 1860 and 1870.

Where authority lists are lacking, some standard textual patterns provide sufficient accuracy: "Lieutenant + NAME" describes a person. In other cases, promising patterns are less satisfactory: General Assembly, General Committee, General Announcement, and similar patterns make the pattern "General + NAME" problematic. Such troublesome patterns seem to grow in number but at a rate less than the size of the overall corpus. Doubling the size of the corpus may require only a modest increase in exception lists. Moreover, since troublesome prefixes such as "General" are finite, there are ways to minimize their impact. Terms such as "General" vary from corpus to corpus. Analyzing patterns of how such phrases grow in different corpora is an on-going subject of research.

The first stage of analysis generates a tagged corpus with 3.2 million proposed named entities. Although we have not yet carefully measured the precision of these results, qualitatively the precision is sufficiently high and the corpus sufficiently large that we can extract useful statistics. The initial training set contains the following classes of data:

1. Frequency of string in class. In the 55 million word corpus, Washington appears 3,910 time as a recognizable place name, 605 times as surname (e.g., "George Washington") and 95 times as a forename (e.g., "Washington Irving").

2. Trigrams preceding word classes. The phrase "the vicinity of" precedes place names 592 times and forenames just 5 times. We allow for apostrophes so that phrases such as "in the vicinity of Smith's farm," where Smith is a person and "vicinity" modifies "farm" are properly tagged.

3. Trigrams following word classes. The phrase "has gone" follows surnames 38 times and place names just twice.

4. Matches to real world entities. This records how often texts identifiably mention Washington, D. C. (3,064), versus Washington, GA (143) and so forth.

The second major stage uses the statistics collected above to classify underspecified terms. It calculates the probability that a given Washington is a place or a person and, if a place, which Washington it is. This stage looks at three types of evidence:

1. Specific annotations to particular passages. If we encounter "Washington," we check to see if we have seen a marked instance of Washington (e.g., President Washington or Washington, D.C.) within the last ten pages. If so, we default to that entity.

2. Usage patterns for a particular document. We examine the statistics for "Washington" within the current document as a whole. If Washington appears, we combine the frequency from this document with corpus-wide trigram statistics to classify it as person or place.

3. If the term does not appear in this document, we check the statistics for the corpus as a whole to classify it as person or place.

At this point, we link particular mentions to specific entities. In some cases, (e.g., "the New York Herald," "Washington, D. C."), we have been able to link phrases to real world entities early on in the process. In many cases, though, we may be confident that a particular Washington is a place but need to determine which "Herald" or which "Washington" is meant. Different classes of entity pose different problems: personal names are particularly tricky since they comprise and open set and are variable in form: we need to determine how likely it is that "Lee," "Robert E. Lee," "R. E. Lee," and "Robert Edward Lee" all designate the same person.

Much of our work has gone into developing adequate authority lists for this corpus. As mentioned above, even in the best case, where we can use the Getty Thesaurus of Geographic Names, we encounter serious problems identifying places from older documents. We have thus entered the 1855 Harper's Gazetteer of the world, which lists 60,000 places and provides us with our best model of American geographic naming conventions in the mid 19th century. In other cases, we have lacked such comprehensive resources and been even more dependent upon converted print materials: thus we have include works such as Dyer's Compendium of the Civil War for its list of Union regiments and roster of Union Generals, Rowell's 1870 Newspaper guide etc.

4. CATALOGING THE CIVIL WAR-ERA COLLECTION

We set out to catalog the Civil War-era collection in the following way. First, we collected as much cataloging data as possible from existing sources. This process involved collecting typical bibliographic data from the print volumes and using that information to query the Library of Congress catalog. We then analyzed the XML structure of the digital document in order to create analytical entries for every major section of the document. The choice of what constitutes a major section was largely based on the nature of the document. We looked for the smallest sections that had consistently useful headings. Empirically, this usually corresponds to the chapter level. Next, we extracted named-entity information from each document, connected each reference to available authority records, and recorded each distinct named entity as a subject heading for the document. Finally, we extracted cross references and other relationships between works. All of this information was then written out into a series of metadata records. This process was completely automated, with no human interaction required. Cataloging the entire 55 million word corpus took a matter of hours on a single fairly ordinary server.

As a container for our metadata records, we chose the Metadata Object Description Schema (MODS) [22] recently developed by the Library of Congress. MODS has several advantages for our purposes. Perhaps the most important benefit is that it was created within the context of traditional library cataloging. MODS is a direct descendent of the MARC format. Crosswalks are available between MARC and MODS that offer nearly lossless round trips. Although MODS does have somewhat reduced specificity compared to MARC, conversions between the two metadata container formats are much more robust than conversions from either format to Dublin Core. The fact that MODS is an XML format means that it is possible to use widely available software tools for editing, searching, and formatting metadata records. These include implementations of the XSL stylesheet language and XML databases such as eXist [6].

MODS is also well suited to the content of the collection. Although the tools and interfaces through which one might interact with documents in the Perseus Digital Library are profoundly different from those in which one might approach a row of bound books on a shelf, the materials within our collections are not yet significantly different. Our Civil War-era collection is primarily composed of scanned books. The fields in MODS and the practices of AACR2 are well suited to describing the materials that form the source of this collection.

Furthermore, MODS records containing catalog data for the books that make up our collection are readily available. The Library of Congress has recently begun providing catalog records in MODS format through an SRW wrapper around the Z39.50 interface to their Voyager catalog [29]. Although it appears that the path of a catalog record through this system is complicated, this public interface is remarkably simple, robust, and easy to use. Particularly compared to the complexity and instability of many Z39.50 clients, the systems overhead required to interact with the LC catalog using SRW is extremely low. These records form a strong starting point on which we can base our attempts to extend the catalog data.

4.1 Analytical cataloging

The analytical cataloging of the Civil War-era texts is founded on four TEI tags, <pb> for page breaks, varia-

tions of for divisions such as chapters, <head> for the headings of chapters, and <argument> for chapter arguments. An argument is a short summary of the main topics of a chapter, usually in the form of short phrases separated by em-dashes. These tags form the most basic level of markup in our TEI documents. They are fairly inexpensive to add. The page break tags are created automatically as a result of the OCR process. Annotations for chapter divisions, headings, and arguments can be efficiently added by editors with little specific training. A typical volume in the Civil War-era collection contains between 20 and 30 chapters. Entry-based works such as encyclopedias typically contain several thousand individual entries.

```
<div1 id="c.21" type="chapter" n="21">
<pb id="p.306" n="306"/>
<head>Chapter 20:  military situation in <placeName
key="tgn,7007255" reg="Kentucky"> Kentucky
</placeName>. </head>
<argument>
<p>
<list type="simple">
<item><persName key="n0003.0021.00306.03023"
n="Johnston, General,,,,"> <roleName n="General">
General </roleName> <surname> Johnston </surname>
</persName>'s arrival in <placeName> Nashville
</placeName>. </item>
<item>personal reminiscences, the defense of
<placeName key="tgn,7007825" reg="Tennessee"> Tennessee
</placeName>. </item>
<item><persName key="n0003.0021.00306.03024"
n="Johnston,General,,,,"> <roleName n="General">
General </roleName> <surname> Johnston </surname>
</persName>'s resources and theory.</item>
<item>letter to <persName key="n0003.0021.00306.03025"
n="Davis, President,,,,"> <roleName n="President">
President </roleName> <surname> Davis </surname>
</persName>. </item>
```

Figure 1: An example of the TEI markup that supports automatic catalog generation. All personal and geographic name tags were added automatically.

It should be noted that the tags that allow the analytical cataloging described here require a very low level of investment in digitization. The Gutenberg distributed proofreading system [24] is one example of an efficient process that could provide the level of annotation required to generate analytical records. Assuming that quality OCR output is available, the single largest cost of digitizing a text is fixing recurring problems in the body of the text. This process includes correcting OCR errors, moving footnotes, and marking paragraphs. Adding chapter divisions and correcting and annotating the headings that go with them is a tiny fraction of this work, which can be done long before a text is fully proofread. In addition, chapter divisions and headings tend to be typographically distinct from the body of a text, so it is likely that structural divisions and headings could be reliably extracted by automated methods. Even digital libraries that rely on page images and uncorrected OCR could thus still generate high-quality component catalog records of this type with little extra investment.

Once a text has been marked up with structural tags, it is a simple process to extract a list of document sections. This task can be accomplished with any XML parser. Once we have processed the XML document, we have some part of the following information:

1. A title derived from the <head> element.
2. A summary derived from the <argument> element.
3. The extent in pages of the section, based on the <pb> elements that fall within the document section.
4. Metadata for the work as a whole.

We have encoded this information as follows. We place the title of the section in the <titleInfo> element. We chose to place any chapter arguments in the <abstract> element, although they could arguably be considered tables of contents. The most important section of the new MODS record is the <relatedInfo> element. This element allows us to specify the relation of the current document section to the document as a whole. MODS defines many types of related info, including "host", which describes the relation of a part to a whole. The MODS <part> element allows us to specify the precise type and location of the current section. We used the "type" element of the tag that specifies the document section, the order of that tag within the document, and the page extent to generate this information.

MODS allows any element to occur in a <relatedItem> tag, so we could include the entire MODS record for the host document. This information makes it simpler for the online catalog interface to display to the user from which document the current record is drawn. In order to reduce redundancy, however, we have chosen to represent only the title, author, and editor elements of the host document.

There are two additional elements, the <identifier> tags. These are internal identifiers within the Perseus Digital Library. The identifier within the <relatedInfo> element is the Perseus ID of the host document, for example "Perseus:text:2001.05.0029". The identifier for the current section of that document is the ID of the document, concatenated with a "query" describing the location of the current section, for example "Perseus:text:2001.05.0029". Either identifier is sufficient to call up the full text of the document or the document section within the Perseus Digital Library interface.

Note that although we have so far only cataloged one level of document sections, there is no reason that we could not generate records for arbitrary layers of document hierarchy. A chapter record could refer to the record for a particular volume, which could itself refer to the main record for the entire work.

One of the most dramatic examples of the power of digitized documents to provide detailed cataloging information is the case of encyclopedias. In a typical cataloging environment, no one would consider creating a record for each entry – the cost, compared to any other book, would be immense. For a digitized encyclopedia, the procedure is straightforward. Each individual entry is a division, just like a chapter in a typical monograph. Although the significant expense of digitizing and tagging an encyclopedia is justified in itself, the secondary benefits to the catalog are profound. Cataloging the entries in a print edition of Harper's Encyclopedia of United States History would not only be prohibitively expensive but also most likely not particularly useful. Once we

```
<?xml version="1.0"?>
<mods xmlns:xsi="http://www.w3.org/2001/
XMLSchema-instance" xmlns="http://www.loc.gov/
mods/v3" version="3.0" xsi:schemaLocation=
"http://www.loc.gov/mods/v3 http
://www.loc.gov/standards/mods/v3/mods-3-0.xsd">
<titleInfo> <title>Chapter 49</title> </titleInfo>
<abstract>
From the North.  -- rumored defeat of Gen.  Early.  --
panic among officials.  -- moving the archives.  --
Lincoln's inaugural.  -- victory in North Carolina.  --
rumored treaty with France.  -- Sheridan's movements.
-- letter from Lord John Russell.  -- Sherman's
progress.  -- desperate condition of the government.
-- Disagreement between the President and Congress.
-- Development of Grant's combination.  -- assault at
Hare's Hill.  -- departure of Mrs.  President Davis.--
</abstract>
<subject><name type="personal" authority="naf">
<namePart>Sheridan, Philip Henry</namePart> <namePart
type="date">1831-1888</namePart> </name> </subject>
<subject><name type="personal"
authority="naf"><namePart>Lee, Robert
E. (Robert Edward)</namePart> <namePart
type="date">1807-1870</namePart></name> </subject>
<subject><name type="personal"
authority="naf"><namePart>Grant, Ulysses
S. (Ulysses Simpson)</namePart> <namePart
type="date">1822-1885</namePart></name></subject>
...
<subject> <geographic>North Carolina</geographic>
<geographicCode authority="tgn"> tgn,7007709
</geographicCode> </subject>
<subject> <geographic>Richmond</geographic>
<geographicCode authority="tgn"> tgn,7013964
</geographicCode> </subject>
<subject> <geographic>Charlottesville</geographic>
<geographicCode authority="tgn"> tgn,7013585
</geographicCode> </subject>
...
<relatedItem type="host">
<titleInfo> <nonSort>A </nonSort> <title>rebel war
clerk's diary at the Confederate States capital</title>
</titleInfo>
<name type="personal">
<namePart>Jones, J. B. (John Beauchamp)</namePart>
<namePart type="date">1810-1866</namePart>
<role> <roleTerm authority="marcrelator"
type="text">creator</roleTerm> </role>
</name>
<part> <detail type="chapter" order="49"/> <detail
type="volume"> <number>2</number> </detail>
<extent unit="page"> <start>436</start>
<end>463</end> </extent> </part>
<identifier>Perseus:text:2001.05.0022</identifier>
</relatedItem>
<identifier>Perseus:text:2001.05.0022:
chapter=49</identifier>
</mods>
```

Figure 2: An abbreviated example of a MODS record derived from a single chapter. This entire record was generated automatically.

have a digital copy, however, the situation is completely different. Automatically generated analytical catalog records for encyclopedia entries, created for no additional investment, allow users to follow links to those entries directly from the OPAC.

Although we started with approximately two hundred catalog records for the entire Civil War-era collection, by examining the structure of our XML records we have automatically generated 60,000 records. The difference in scale is similar to the difference between the library of a medieval monastery and a modern public library.

4.2 Named Entities

After our first phase of catalog metadata extraction, we have tens of thousands of MODS records, each describing a chapter or entry in a larger work. The metadata in the records is, however, still fairly sparse. In the next phase of metadata extraction, we begin to fill out the individual records with subject headings. There are many types of subject heading. For the purposes of this paper, we have limited our automatic subject analysis to concrete named entities such as personal, corporate, and geographic names. In the final MODS records, we establish a one-to-one correspondence between named entities identified in the original document section and subject headings added to the generated catalog record.

As was previously discussed, automatic named entity extraction requires a significant initial investment in the creation of collection-specific gazetteers and pattern matching rules. It is difficult to estimate the cost of developing named entity extraction systems for the same reason: performance is dependent on how well-tuned existing systems are to the names and patterns in a new document collection. However, once this investment has been made, the system scales well. The cost of running an information extraction system of this nature over one million words is similar to the cost of running the same system over a 55 million word corpus. Verifying the correctness of these automatic tags is significantly more expensive in terms of time and labor, but the quality of the initial automatic pass is sufficient to produce useful subject headings.

Our named entity identification routines scan for a range of entities, including monetary sums, measures of distance, weight and other scales, dates, times, established phrases referring to historical events, etc. Evaluation of patterns within the 19th century English corpus is a major focus of our work (e.g., how often does the pattern General + NAME designate a person or something such as "General Assembly") and will be reported separately. For the purposes of this paper, the most important categories are personal names, places, and organizations. Of these, places are the easiest to evaluate because of the Getty Thesaurus of Geographic Names (TGN). All results vary by genre – newspaper advertisements, for example, are particularly hard to analyze because they contain many short phrases and they use clues such as capitalization inconsistently. In one typical four volume history of the Civil War, however, we classify 29,266 phrases as possible placenames, with precision currently at 95%. For these phrases, we were able to assign TGN numbers to 21,759 of these (75%), with accuracy currently at 90%. Typical errors include reference to "Frederick," denoting Frederick the Great, in a context where we would expect Frederick, MD. Place names for which no

TGN number was assigned typically refer to local points (e.g., "Smith's farm") that would not appear in a gazetteer such as the TGN. Our own results reflect substantial tuning of the links between TGN and historical collection (for example, the TGN has no entry for "Fort Sumter" so we link "Fort Sumter" to Charleston, SC). While we have much work to do, the current system already provides usable input to the cataloging process described here.

In the Civil War-era collection, named entities are explicitly marked and disambiguated by the automatic information extraction system. Entities are marked using the TEI <rs> (reference string), <name>, and <orgName> tags. The type of entity (ship, newspaper, regiment, etc.) is noted in an XML attribute on each tag. In cases where the system is able to resolve a named entity reference to a particular authorized name such as a TGN number, that information is also stored in an attribute.

Just as in the case of structural markup, extracting named entity information from the XML text into an automatically generated catalog record is as simple as parsing the XML. Encoding named entity information as subject headings in MODS is straightforward. Each subject is enclosed in a single <subject> element. Within a given subject element, personal names are encoded in <name> elements. Place names are encoded as <geographic> or <hierarchical Geographic> elements.

Of the 60,000 MODS records that we generated, around 40,000 contained between zero and five subject headings. 9,000 had between six and 15. 4,400 records had between sixteen and thirty distinct named entity references. An additional 8,000 records had more than thirty. The distribution of particular subject headings is uneven. Of the 140,000 distinct personal names detected by the named entity extraction system, 15 people (generals and presidents) occur as subjects in more than 1,000 records. Just over 500 people occur in more than 100 records. Fewer than 9,000 people occur more than ten times. The distribution of places is similar. 54 of 27,000 distinct geographic subject headings occur in more than 1,000 records. 638 occur in 100 or more records. 4,500 geographic names occur in ten or more records.

A catalog is most useful when it is combined with an authority file – a controlled list of authorized name and subject headings. Personal names that have not been aligned to standard authorized forms are useful within the catalog of the Perseus Civil War-era collection. They are sufficient to collocate records that refer to the same named entities. However, linking our named entity records to authorized versions will offer profound benefits. As library catalogs become progressively more connected, using authorized forms for subject headings will not only allow users of our catalog to extend their searches outside our library, but will also allow users of other catalogs to find records in our catalog. This is not an abstract consideration. Standards for library federation such as OAI metadata harvesting and SRW search services are making catalogs increasingly connected.

Our ability to connect named-entity references with entries in authority lists varies considerably. In some cases, we have been able to make highly accurate connections cheaply. In other cases we have not been able to make connections at all. The variables are the availability of widely-used authority lists and the ambiguity between entities.

The best case is the names of places. We have made use of the Getty Thesaurus of Geographic Names (TGN). The TGN is an exhaustive list of placenames throughout the world. It contains geographic coordinates, the type of place (city, town, geographic feature), and any alternate or past names. As discussed above, we have had a good rate of success in automatically resolving place name references to TGN entries. In cases where we have been able to identify a TGN number for a place name reference, we have included a <geographicCode> element within the subject heading along with the <geographic> element.

People provide a more difficult case. Clearly the best authority list for the purposes of the Civil War-era collection is the Library of Congress Name Authority File. Headings exist for most of the major people mentioned in the texts. A student worker was able to connect extracted names from the collection to LC NAF entries using the LC authority web search interface[21] at a rate of approximately 200 names per hour. This rate could be increased with a webservice-based interface to the NAF.

The majority of common personal names in the collection have authorized LC forms. However, just as with place names, the majority of the unique personal names extracted from the collection, especially those that occur infrequently, are not in the NAF. In those cases, we have often been able to extract standard forms of names opportunistically from encyclopedia entries and lists of commanders.

The situation for organizations, corporate names, and events is similar to that for personal names. In many cases we have been able to take advantage of contemporary reference works to construct local authority lists. These entities include newspapers, military units, and battles. More work is needed to study the process of extending existing authority files based on historical reference works, merging controlled vocabularies from multiple sources, and finally incorporating local authority files into general authority files as it seems appropriate. Our experience suggests that a distributed model of authority control as suggested by several researchers [20, 17, 31] could provide a standardized infrastructure for this process.

Current and developing metadata formats make the process of progressively associating named entity subject headings with authorized forms simple. MODS is careful to provide catalogers with space to declare the authority under which a subject heading is provided. This level of annotation provides a useful "hook" by which previously recorded subject headings can be progressively regularized as local authorities are aligned with more commonly used headings. We expect that the developing MADS standard will enable us to create, process and override authority records in the future.

4.3 Cross-references between documents

Another important feature of detailed markup is the ability to mark cross references between documents. In TEI, this is accomplished through the use of <cit> and <bibl> tags. When searching the contents of a library, it is helpful to understand the interrelations between works, authors, and schools of thought. Especially in the case of historical documents, where the intellectual context of a work may not be clear to the reader, it is important to note the fact that one document references, or is referenced by, another work in the collection. One of the criteria by which the works in the Civil War-era collection were selected was the extent

to which they reference each other. For example, the Memoirs of Ulysses S. Grant have been digitized along with a response by Adam Badeau. Badeau frequently quotes long passages of Grant's memoirs.

To the extent possible, in the process of tagging the text of both works all quotes were marked with specific TEI tags that identify not only the extent of the quote, but the source of the quote as well. Just as with the named entity tags, these citation tags can be read by any XML parser. The result is that for every section of a document, we can generate a list of documents referenced by that section. Assuming that the referenced section is also in the collection, these references can then be added to both the source record and the target record within the MODS <relatedItem> field.

The primary difficulty with extracting cross-reference metadata from the Civil War-era collection has been the lack of standard citation schemes within the corpus. Many of the works in the collection have been reprinted several times, occasionally under slightly different titles. It is not immediately clear how to uniquely identify an intellectual work so that any edition of that work will benefit from the cross-reference notation. Existing identifiers, such as LC control numbers, ISBNs, and OCLC accession numbers all refer only to a particular edition of a work. Within the collection this is less of a problem, as we have only digitized one copy of each work. Recording cross-references with standard identifiers is a practical solution for local functionality. As digital library catalogs become increasingly federated, however, it will become very useful to be able to identify documents in such a way that the digitization of a work at another library will, without any human intervention, cause a new link to be created at the Perseus Digital Library. It is our hope that as IFLA's Functional Requirements for Bibliographic Records (FRBR) [14] recommendations are implemented, practices for uniquely identifying intellectual works will become available.

5. DISTRIBUTING THE CATALOG

Ultimately, the value of cataloging is based on the extent to which it enables users to search collections. The methods described above enable a well-structured digital library to provide an unprecedented level of access to its collections. On the other hand, the huge volume of data created can also be an impediment. It is important to consider how to make the data present in the Perseus catalog data available to the largest audience without overwhelming users with sparse, repetitive data.

Emerging technical standards have made it increasingly simple to provide access to catalog data. At the most basic level, the MODS standard serves as an excellent platform for digital cataloging. It is easy to integrate MODS records into a variety of applications. Records can be transported easily across the web using existing, well-tested software. Technology for parsing, searching, and displaying MODS XML is readily available. We are providing three primary methods of access for catalog data.

1. An XML-based online public access catalog

   The simplest and most powerful way to search the collections is a custom web application provided at the Perseus Digital Library. This application is based on the eXist XML database[6]. The eXist database ships with an example web application that performs most of the search and display functions of a typical OPAC. The web application itself is a single script, barely 250 lines long. Every other aspect of the system's functionality is represented either by the MODS record itself, the inherent search capabilities of the XML database, or a set of XSL stylesheets that format the MODS records into HTML web pages. After some small modifications to the default eXist interface, we had a functional online catalog.

2. Harvesting records from an OAI provider

   Another method for disseminating catalog data is through an Open Archives Initiative data provider [19]. Implementations of the OAI protocol are widely available. OAI providers are capable of distributing metadata records in several formats, including MODS and Dublin Core. Crosswalks exist from MODS to Dublin Core. However, it should be noted that the hierarchical nature of the analytical cataloging in this collection would be almost completely erased in the transition to unqualified Dublin Core. The syntax of the DC Relation field carries less information than the <relatedInfo> section of a MODS record. It has been our experience that the difference in difficulty between creating MODS metadata records and Dublin Core records is negligible and easily offset by the enhanced functionality offered by the better structured fields. We expect that as OAI providers increasingly choose to offer MODS-formatted records along with Dublin Core-formatted records, the functionality of services based on harvested metadata will increase proportionally.

   In order to prevent harvesters of the Civil War-era collection from being swamped with records, we make use of OAI sets. All work-level records are in one set, all analytical records in a separate set. In this way we can still provide a listing of works that are available for harvesters who do not want to deal with the additional 60,000 records.

3. Federation through SRW interface

   One important tool for digital library federation is the recently developed Search/Retrieve Webservice protocol (SRW) [29]. SRW is a simplification of the Z39.50 information retrieval protocol. The advantages of the HTTP-based SRW over Z39.50 have already been discussed in this paper. SRW differs from OAI in that clients can perform searches as needed, without harvesting the entire contents of a repository. Implementations of SRW and the related CQL query language are available from several providers, including OCLC Research.

The second challenge in disseminating automatically generated metadata, that of not overwhelming users, is more difficult to address. One obvious question is whether the vastly increased number of subject headings will actually improve the user's ability to find relevant documents. Catalog records are limited in that there is currently no way to distinguish a chapter that mentions Ulysses S. Grant in passing from one that is primarily about Ulysses S. Grant. In the past, this was not a problem because only a small number of subject headings were added to each record. Further research is needed on the most effective ways to present

large numbers of catalog records to users. One possible direction would be to record the number of times that a given person, place, or organization is mentioned in a document in the <subject> element, and sort search results accordingly.

It may prove impractical from an interface standpoint to include every named entity reference in the collection in cataloging data. However, it is useful to consider the distribution of personal names in the Civil War-era collection. There are approximately 140,000 unique names. Of these, only about 100 occur more than one thousand times in the running text of the entire 55 million word corpus. An additional 20,000 names occur 10 or more times. The statistics for subject headings are even more tractable. Of the 60,000 MODS records generated from the collection, "Grant, Ulysses Simpson" occurs as a subject heading in approximately 1400. A more typical example, "Marblehead, Mass." occurs in 80 records. It is unlikely that the majority of those references to Marblehead would be included in any current catalog. The presence of those more marginal subjects in automatically generated catalog records allows users to perform searches within the library OPAC that would previously only have been possible using free-text search engines.

Another aspect to consider when analyzing the utility of large numbers of analytical catalog records subject headings is the physical distance between the catalog and the collections. The catalog system described above is closely integrated with the rest of the Perseus Digital Library. Each record includes an active hyperlink to the digital resource described. This feature makes it easier for researchers to analyze and evaluate large numbers of search results significantly more efficiently than they could in a standard library OPAC.

6. CONCLUSION

We have shown how an initial investment in digitization and markup can have benefits beyond simply making a text available online. Even based on a relatively low level of tagging and OCR correction, automatically generated catalog records can significantly increase the presence of digital objects within familiar, widely used library systems. The addition of authority control and emerging standards for catalog interoperability allows the benefits of text digitization to be felt far beyond individual digital libraries. Catalogs of digitized texts can efficiently display the interconnections within and between works, based on both direct references and topical similarity. Users can explore digital library collections with unprecedented levels of detail and specificity, without significant additional investment in cataloging.

7. ACKNOWLEDGMENTS

This work was made possible by the NSF Digital Library Initiative Phase 2 (NSF IIS-9817484), with particular support from the National Endowment for the Humanities. The authors would also like to thank Candy Schwartz, Briana Kula, Gwynne Langley, Lisa Cerrato, Anne Sauer, and Jennifer Mimno.

8. REFERENCES

[1] CiteSeer: Scientific literature digital library. Available online at: `http://citeseer.ist.psu.edu/`.

[2] CLiMB toolkit. Available online at: `http://www1.cs.columbia.edu/~delson/CLiMB/CLiMB/Toolkit/`.

[3] P. T. Davis, D. K. Elson, and J. L. Klavans. Methods for precise named entity matching in digital collections. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 125–127. IEEE Computer Society, 2003.

[4] L. Dempsey, E. Childress, C. J. Godby, T. B. Hickey, D. Vizine-Goetz, and J. Young. Metadata switch: thinking about some metadata management and knowledge organization issues in the changing research and learning landscape. Available online at: `http://www.oclc.org/research/publications/archive/2004/`, August 2003. Forthcoming in LITA guide to e-scholarship (working title).

[5] N. P. Ellero. Panning for gold: Utility of the world wide web for metadata and authority control in special collections at the claude moore health sciences library. *Library Resources & Technical Services*, 46(3):79–84, 87–91, July 2002.

[6] eXist: Open Source Native XML Database. Available online at: `http://exist.sourceforge.net`.

[7] G. Giuffrida, E. C. Shek, and J. Yang. Knowledge-based metadata extraction from postscript files. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries*, pages 77–84. ACM Press, 2000.

[8] J. Greenburg. Metadata extraction and harvesting: A comparison of two automatic metadata generation applications. *Journal of Internet Cataloging*, 6(4):59–82, 2004.

[9] H. Han, C. L. Giles, E. Manavoglu, H. Zha, Z. Zhang, and E. A. Fox. Automatic document metadata extraction using support vector machines. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 37–48. IEEE Computer Society, 2003.

[10] H. H. Hoffman. Future outlook: Better retrieval through analytic catalogs. *Journal of Academic Librarianship*, 11(2):151–153, 1985.

[11] H. H. Hoffman. Library practices with NINO. *Technicalities*, 22(2), 2002.

[12] H. H. Hoffman. Better recall of exact work titles in online catalogs. *Public Libraries*, pages 344–346, November/December 2003.

[13] C.-C. Huang, S.-L. Chuang, and L.-F. Chien. Using a web-based categorization approach to generate thematic metadata from texts. *ACM Transactions on Asian Language Information Processing (TALIP)*, 3(3):190–212, 2004.

[14] IFLA Study Group on the Functional Requirements for Bibliographic Records. *Functional Requirements for Bibliographic Records: Final Report*, volume 19 of *UBCIM Publications-New Series*. K.G.Saur, München, 1998.

[15] iVia under the hood: Harnessing new technologies in support of collaborative service design and amplification of expert effort. On the Web, July 2004. Retrieved from `http://infomine.ucr.edu/iVia/newtech.shtml/`.

[16] C. Jenkins, M. Jackson, P. Burden, and J. Wallis. Automatic RDF metadata generation for resource discovery. In *WWW '99: Proceeding of the eighth international conference on World Wide Web*, pages 1305–1320. Elsevier North-Holland, Inc., 1999.

[17] M. Kaiser, H. J. Lieder, K. Majcen, and H. Vallant. New ways of sharing and using authority information: The LEAF project. *D-Lib Magazine*, 9(11), November 2003.

[18] Z. Khurshid. Analytical cataloging of full-text journal databases at a middle east university. *Cataloging & Classification Quarterly*, 32(2):81–89, 2001.

[19] C. Lagoze and H. V. de Sompel. The Open Archives Initiative: Building a low-barrier interoperability framework. In *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries*, pages 54–62, 2001.

[20] K. T. Lam. XML and global name access control. *OCLC Systems & Services*, 18(2):88–96, 2002.

[21] Library of congress authorities. Available online at: `http://authorities.loc.gov/`.

[22] S. H. McCallum. An introduction to the Metadata Object Description Schema (MODS). *Library Hi Tech*, 22(1):82–88, 2004.

[23] D. E. Morris, C. B. Hobert, L. Osmus, and G. Wool. Cataloging staff costs revisited. *Library Resources & Technical Services*, 44(2):70–81, 2000.

[24] G. B. Newby and C. Franks. Distributed proofreading. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 361–363. IEEE Computer Society, 2003.

[25] G. Patton. FRANAR: A conceptual model for authority data. *Cataloging & Classification Quarterly*, 38(3/4):91–104, November 2004.

[26] M. Patton, D. Reynolds, G. S. Choudhury, and T. DiLauro. Toward a metadata generation framework: A case study at Johns Hopkins University. *D-Lib Magazine*, 10(11), 2004.

[27] A. J. Schimizzi. Enhancement of research library print material through the use of component cataloging: An oclc user's perspective. *Cataloging & Classification Quarterly*, 38(1):65–86, 2004.

[28] M. M. M. Snyman and M. J. van Rensburg. Revolutionizing name authority control. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries*, pages 185–194. ACM Press, 2000.

[29] SRW - Search/Retrieve Webservice. Available online at: `http://www.loc.gov/z3950/agency/zing/srw/`.

[30] A. G. Taylor. *Wynar's Introduction to Cataloging and Classification*. Libraries Unlimited, Westport, Connecticut, 2004.

[31] B. Tillett. Authority control: State of the art and new perspectives. In *Authority Control: Definition and International Experiences*, International Conference, Florence Italy, February 10-12, 2003, February 2003.

[32] VIAF: The virtual international authority file. Available online at: `http://www.oclc.org/research/projects/viaf.`

[33] Y. Wang, F. Makedon, J. Ford, L. Shen, and D. Goldin. Generating fuzzy semantic metadata describing spatial relations from images using the r-histogram. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 202–211. ACM Press, 2004.

[34] O. Yilmazel, C. M. Finneran, and E. D. Liddy. Metaextract: an NLP system to automatically assign metadata. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 241–242. ACM Press, 2004.