"A Testbed of Civil War-Era Newspapers"

IMLS grant #LG-02-03-0082-03

Semi-annual report, March 2006-September 2006

Submitted by: James Rettig, PI
University Librarian
Boatwright Memorial Library
University of Richmond

Introduction

The University of Richmond and its partner the Perseus Project of Tufts University continue to make progress on this project. This report addresses the following highlights of the past half year:

1. No-cost extension request granted
2. Ingestion of entity-tagged files into DLXS/XPAT
3. Site usage and feedback
4. Metadata for the Richmond *Daily Dispatch* Collection
5. Influence of the "A Testbed of Civil War-Era Newspapers" project
6. Project Perseus Accomplishments
    a. JCDL 2006 presentations
    b. Sending a programmer to Richmond
    c. GAMERA, machine learning, and research with newspapers
    d. The potential of newspaper digitization: an update
    e. Historical newspapers and metadata needs: recent developments
    f. Newspaper digitization: projects report continuing challenges
    g. Semantic searching and historic newspapers
    h. Conclusion

1. No-cost extension request granted

On October 3, 2006 Laura Mahoney of IMLS sent notice that our request for a no-cost extension was granted as follows:

> Please be advised that Award Start and / or Award End date for the following award has changed.
> Award Log Number: LG-02-03-0082-03
> Organization Name: University of Richmond, Boatwright Memorial Library
> Original award dates: From: 10/01/2003; To: 09/30/2006
> Modified award dates: From: 10/01/2003; To: 09/30/2007
> Revised reporting schedule is included below.

Type	Due Date	Date Received	Delinquent
Final Financial	12/29/2007		No
Final Narrative	12/29/2007		No
Interim Narrative	05/01/2007		No
Interim Narrative	10/31/2006		No
Interim Narrative	05/01/2006	04/12/2006	No
Interim Financial	10/31/2005	04/11/2006	No
Interim Narrative	10/31/2005	10/18/2005	No

Interim Narrative	05/01/2005	03/31/2005	No
Interim Narrative	10/31/2004	10/01/2004	No
Interim Financial	10/31/2004	12/22/2004	No
Interim Narrative	05/01/2004	03/31/2004	No

2. Ingestion of Entity-Tagged files into DLXS/XPAT

Significant progress was made during this reporting period in the areas of data conversion and data ingestion into DLXS/XPAT. As noted in a previous report, the process can be fairly involved. Yet by building on Andrew Rouner's work, a team consisting of Nancy Woodall, Rick Neal, and Chris Kemp was able to successfully load all issues of the *Daily Dispatch* in mid-June. The collection provided a good beginning for the collection building process, and allowed full-text searching as well as browsing the collection by date. This success ultimately proved to be a test-run, however, since the ingested files were not the entity-encoded xml files created through the transducer process at the Perseus Project.

The University of Richmond received these files from Perseus at the beginning of July, 2006, and started work immediately to convert the data into the format required by XPAT for the indexing process. Since the files delivered to Perseus for named entity work were the original files Richmond received from Digital Divide Data (the vendor responsible for the rekeying of the *Daily Dispatch*), the same procedures used to build the first collection would be used again in the processing of the entity-encoded files. The level of encoding performed by Perseus, however, required Richmond to determine what DLXS would be capable of doing with the data. Several considerations were taken into account during this process:

- Will DLXS be able to handle the level of encoding present?
- Is DLXS capable of referencing external authority sources?
- What entities will be of most use to our collection's users?
- How should our workflow be adjusted to handle the encoded data?

These questions were discussed in consultation with DLXS support and meetings with the University of Michigan's Digital Library Production Service personnel when a Richmond contingent attended the annual DLXS workshop in July. It was determined that while DLXS and XPAT would be able to process the level of encoding present in the entity-tagged files, the repetition of search terms within the tags themselves would present an interesting problem when performing keyword searches within the full text. Since XPAT indexes each word in a collection, both within the text and within the markup itself, full text searching returns hits on normalized values of text that reside inside the entity tags. It was decided that in the case of encoded personal name entities, since much of the markup creates authority at issue- rather than collection-level, that much of it would be stripped in order to facilitate clean search results. However, normalization tagging was retained within place name entities since they reference the Getty number for a geographic location, thus providing external authority at the collection level. At present, Richmond's proficiency with DLXS is not at the level where use can be made of external authority records, but a tool within DLXS called a "word wheel," which creates an SGML index of each word in a collection, shows some promise for use with this collection and is being explored. Richmond has retained the encoded files as delivered by Perseus and is currently working on potential uses in DLXS for such deep encoding within place name tagging.

In addition to stripping tagging from personal names, Richmond removed tagging from entities identifying foreign languages, distances and numbers, among others. This was deemed appropriate since the collection will be searched for largely for the personal name, place name and organizational name entities. Customizing searching based on these entities involves coding of the map file and search file for the collection as a whole. The search file creates customized searches in the XPAT language, and the map file relates these searches to terms that appear in the user interface.

The major workflow adjustment that needed to be made involved the collection's DTD. Perseus uses a catalog file to validate their xml files, a practice which is not functionally supported by DLXS. Adrian Packel, a programmer from the Perseus Project, traveled to Richmond in early August in an effort to assist us in implementing this, but UR was ultimately unable to do so. To solve the validation problem, a DTD

that would allow validation of the nearly 1,400 files was hand-coded. The software package Oxygen was used to create a baseline DTD from ten concatenated files, which was then edited until all files validated against it.

A collection of ten files was built on the entity-encoded data by early September, and using this as a model Richmond successfully ingested all 1384 files into DLXS on 16 September. Several data irregularities were discovered that appeared to be the result of the tagging process performed by Perseus. These are being compiled and UR will submit them to Perseus shortly. Another collection addressing these anomalies was built and released in early October. Currently, work is continuing at Richmond in an effort to optimize search capabilities using named entities in DLXS.

3. Site Usage and Feedback

A "soft release" of Richmond's beta collection at http://dlxs.richmond.edu/d/ddr was performed on 20 October. This involved posting notifications and links to the collection on five Civil War message boards and sending email messages to five Civil War scholars. Scripting to track web site usage via Google Analytics (http://www.google.com/analytics) was added to eleven static web pages on 23 October. Preliminary statistics for an approximately 24-hour period show 410 site visits from 353 unique visitors, with a total of 1189 page views. Viewers have seen the collection from the United States, United Kingdom, Germany and Denmark. Unfortunately, Google Analytics cannot compile statistics on the number of queries to the newspaper collection itself.

In addition, Richmond has received four responses to its site survey and sixteen responses to message board posts as of 24 October. The site has also been linked to from the University of Tennessee at Knoxville's American Civil War Homepage (http://sunsite.utk.edu/civil-war/warweb.html#general).

4. Metadata for the Richmond *Daily Dispatch* Collection

Richmond is working on ways to leverage the advanced metadata encoding of the RDD collection to its best advantage in the DLXS system. Currently, we can provide precision search options based on the named entity tags for person name, place name, company names, organization names, newspapers, military units and railroads. With the next version of DLXS, we hope to incorporate the "word wheel" feature, which would allow users to view a virtual "authority" list of related terms based on the name they are searching.

We are exploring our options for exposing OAI compliant metadata of our collection. We plan to create a collection record available for harvest as well as marc record for the collection and the individual newspapers in our local catalog and OCLC. We are investigating and planning the work for submitting metadata to the AmericanSouth.org collection. All metadata work will be completed by December 31st 2006.

5. Influence of the "A Testbed of Civil War-Era Newspapers" Project

The project has advanced far enough and has attained enough visibility that it is beginning to have influence on other digitization projects:

- We are working with Will Thomas at University of Nebraska-Lincoln to harvest railroad data from the Civil War newspaper collection.
- We are collaborating with the Virginia Center for Digital History, providing testbed newspaper data for pre-Civil War language analysis.
- Rachel Frick is providing consultation services to Rhodes College in relation to the newspaper component of their IMLS-funded Cross Roads digital project
- We have provided extensive documentation detailing the conversion and loading of XML documents into DLXS to Dr. Andrew Rouner, head of digital libraries at Washington University in St. Louis.

- We have provided information and recommendations to the Library of Virginia in relation to microfilm newspaper digitization for use in developing workflows for LVA's participation in the NEH funded newspaper digitization project.

6. Perseus Project Accomplishments

Since the last report, the Perseus Project has engaged in a number of tasks 1) presenting a paper regarding the IMLS project at the JCDL 2006 2) sending programmer Adrian Packel to Richmond to assist with the DXLS implementation 3) continuing research into GAMERA, machine learning and other article segmentation and named entity techniques on historical newspapers 4) some final research into recent developments with historical newspapers. This report will deal with each of these issues in turn.

6a. JCDL 2006 Presentation

At JCDL 2006, in Chapel Hill, North Carolina, staff member Alison Jones presented the paper "The Challenge of Virginia Banks: An Evaluation of Named Entity Analysis in a 19th-Century Newspaper Collection." The paper was nominated for the Vannevar Bush award at the conference and the presentation seemed well received, with a number of questions asked regarding the potential of named entity systems for other types of historical materials. The JCDL meeting also presented an opportunity for Perseus staff members to meet with Chris Kemp, who replaced Andrew Rouner at the University of Richmond and discuss a variety of issues.

6b. Sending a Programmer to Richmond

As part of the technology transfer aspect of the grant, we sent Perseus Project staff programmer Adrian Packel to Richmond in August to assist in configuring the DLXS system so that the XML files with the named entity tags could be successfully loaded, an obstacle that turned out to be insurmountable. To load the files, the various XML files had to be converted to the format required by DLXS, but project staff were unable to get the files to work with XPAT, DLXS's search engine, and the collection would also not display. A larger discussion of this issue, as well as the work around to get the named entities to display can be found in the University of Richmond section.

6c. GAMERA, Machine Learning, and Research with Newspapers

As indicated in the original grant, there was an initial plan to explore the potential of GAMERA for article segmentation on the historical newspapers. A computer scientist research student spent this summer at Perseus exploring how GAMERA could be used on a variety of materials including some of the Liberator files. It was determined that GAMERA did not produce competitive results in terms of article segmentation for 19th century newspapers when compared with either commercial alternatives or double key entry. The research report attached will discuss some of the current solutions to article segmentation that are being implemented at other projects.

Additionally, one comment raised by a number of individuals at the JCDL was that the Perseus named entity system, which is currently based on a combination of hand written and automatically learned rules as well as gazetteers and other knowledge sources, could be improved through the use of a machine learning algorithm such as HMM (Hidden Markov Models) or CRF (conditional random fields). Former Perseus project member David Mimno along with Alison Jones is currently exploring the use of CRF on the Richmond *Daily Dispatch*, to see if a superior named entity system can be built.

There are several different approaches to named entity recognition. The most common methods (such as the Perseus System and GATE) is to make lists of strings that could be people, places, railroads, etc, and then define patterns like "[Place] University" -> EducationalInstitution. While this works fairly well, as indicated by our evaluations this requires a large number of lists and requires the implementer to create large numbers of patterns and discover special cases, which can take considerable time and expertise.

Another approach is to take texts with named entity annotations as training data (such as an XML file of the Richmond *Daily Dispatch*), and attempt to make a computer learn the patterns for you. This usually involves transforming input XML like Figure One into a sequence like Table One.

Figure One: XML Named Entity Strings

<orgName>Tufts University</orgName> is in <placeName>Medford</placeName>

Table One: Sample Pattern Learned By Computer

Observed Word	State
Tufts	ORGNAME
University	ORGNAME
Is	WORD
In	WORD
Medford	PLACENAME

One tool for this sort of task is a Hidden Markov Model, which is made up of two tables: one is the probability of going from one state to another, and the second is the probability of emitting a particular word if you are in a particular state. If you make a few assumptions (each state only depends on the previous state), it's relatively simple to learn the two tables from training data. HMMs are "generative", in the sense that you could start in a random state and "hallucinate" a sentence that would look sort of like English, just by picking a random word and a random new state based on the probabilities you learned from the training data.

A Conditional Random Field is slightly more flexible, because it only learns the probability of the states given the words. In other words, the model doesn't say anything about how to generate new sentences, it just says something about sentences you provide it. This means that it doesn't have to make as many assumptions, and you can have richer sets of observed features. So in the previous example, you could use training data like this:

"Tufts" CAPITALIZED IN-LIST-OF-UNIV [ORGNAME]
"University" CAPITALIZED [ORGNAME]
"is" [WORD]
"in" [WORD]
"Medford" CAPITALIZED IN-GAZETTEER [PLACENAME]

In other words, you can combine information from several different sources and still be able to train efficiently.

In practice, training a CRF extractor takes a long time (hours to weeks), but it runs very fast. It requires labeled training data. Nonetheless it can find more complicated, sensitive patterns than humans can because it is better at finding and taking into account all the exceptions to rules, unintended consequences, etc, that make writing patterns so time-consuming. We are currently examining how the use of a CRF extractor might provider better named entity results with the Richmond *Daily Dispatch*.

6d. The Potential of Newspaper Digitization: An Update

In the last two years, the landscape of newspaper digitization has changed dramatically. The announcement of the National Digital Newspaper Project (NDNP),1 now soliciting grant proposals for its second major stage,2 and various smaller projects have illustrated the growing importance of historical digital newspaper projects. For example, in the spring of 2005, OCLC held a conference entitled "Digitizing Historic Newspapers: A Practical Approach" dedicated to historical newspaper digitization. Similarly, in May of 2006, IFLA held a conference in Salt Lake City entitled, "Newspapers of the World Online: U.S. and International Perspectives" that covered commercial vendors in the field, the status of national digitization project, technical issues, and the NDNP. This report will provide a brief overview of recent research in newspaper digitization conducted by the Perseus Project.

[1] http://www.loc.gov/ndnp/
[2] http://www.neh.gov/projects/ndnp.html

6e. Historical Newspapers and Metadata Needs: Recent Developments

The creation of the NDNP has led to the development of a number of metadata standards and schemas designed specifically for the encoding of historical newspapers that were not previously available. For the time being, the NDNP has chosen not to support individual article segmentation but rather is focusing on creating high quality image files, OCR and structural metadata that will be able to be more easily used by later changes in technology.

A recent presentation by Helen Aguera and Mark Sweeney at the IFLA conference described what the NDNP had accomplished in the last two years.3 Whenever possible the project plans to use open source software and standards, and all materials produced will be available to be reused, which they hope will allow deep linking and persistent identification to support citation. The NDNP is using the OAIS model where the information object is the newspaper, and the data objects are the related TIFFS, JP2, PDF, OCR text, structural metadata, and preservation metadata. The OCR text conforms with the ALTO model (which maps OCRd text to image coordinates), although in general it will be uncorrected. The structural metadata used is METS, and includes title, issue and reel objects.4 They have also chosen Fedora as their repository software and Lucene as their indexing system.

In addition, they make the important point that technology is going to change, for they believe that the accuracy of OCR, automated article segmentation and open source tools will all improve, so they have focused on this first stage at providing preservable content. Additionally, they argue that user expectations will change and that scholars will someday want tools such as text mining, and time and place analysis. They thus conclude that "content is more important than today's system." Time may well prove them right in terms of their belief that many out of the box solutions will prove to have long-term preservation problems.

A recent article by Justin Littman in D-Lib Magazine further described the NDNP's metadata standards and their validation and ingest processes.5 They are currently performing both quantitative and qualitative assurance tests, and have developed a validation scheme for their newspaper digital objects. Each newspaper title is represented by a METS record, which includes both a MARC XML and a MODS record. Every newspaper issue is also described by METS record, with each section and page described by MODS records. In order to validate the newspaper digital objects, they have created annotated templates for each of the METS record types and are using the JHOVE tool. The use of this tool required the addition of special validation rules specific to newspaper data, and the team thus created a software package called the NDNP Validation Library.

There are a number of different established vendors involved in newspaper digitization such as Olive's ActivePaper Archive and CONTENTdm that do support article level segmentation. OCLC recently purchased CONTENTdm, and newspaper digitization services are offered through the OCLC Preservation Center. They use docWorks newspaper edition software developed by CCS (Content Conversion Specialists) a company in Germany.6 This application was based on NDNP standards and can scan from both film and hard copy originals while exporting both METS/ALTO and PDF formats for ingestion into a variety of content management systems. This system also includes support of the NDNP data format, full text searching, highlighted search word hits, and full article segmentation. In fact the California Newspaper project, which is participating in the NDNP, has chosen the OCLC Preservation Services center to provide article level access at the University of California Riverside.7

Some projects have reported scalability issues after choosing to provide article level access. Recently the University of Utah, and Brigham Young University sought to aggregate their newspaper collections stored on disparate CONTENTdm servers. They found that the zoning of individual articles meant there were a much larger number of files than pages. At the University of Utah, 200,000 newspaper files

---

[3] Aguera, Helen, et. al. "The U.S. National Digital Newspaper Program: Thinking Ahead, Designing Now." Presentation at the IFLA Newspaper Conference, http://www.loc.gov/ndnp/pdf/IFLA_NEHLC.pdf

[4] For a fuller discussion of their metadata strategy, please see Murray, Ray L. (2005)."Toward a Metadata Standard for Digitized Historical Newspapers." *Proceedings of JCDL 2005*, pp. 330-1.

[5] Littman, Justin. (2006). "A Technical Approach and Distributed Model for Validation of Digital Objects." *D-Lib Magazine*, http://www.dlib.org/dlib/may06/littman/05littman.html

[6] www.oclc.org/services/brochures/digitizingnewspapercollection.pdf

[7] "California Newspaper Project Update." (2006). IFLA News, No. 14 , May 2006.

translated into 2,000,000 individual article files. They found that they couldn't simply aggregate the newspaper files into their existing multi-site server because they didn't want to overwhelm all the search results in the Mountain West Digital Library with newspaper results. Arlitsch and Jonsson report that: "The aggregation plan calls for the installation of a separate MSS, which will aggregate only newspaper collections. Users of the MWDL will be able to select whether to include newspapers in the general search, or whether to search them separately."8

Nonetheless, both vendors and open source projects often argue that without article level segmentation, full access to newspapers cannot truly be provided. Proquest's Lynda James Gilboe argues that "the article focused approach is essential for capturing the essence of the newspaper experience as it harnesses the power of digital technology."9

6f. Newspaper Digitization: Projects Report Continuing Challenges

A number of newspaper digitization projects have reported on continuing challenges they have faced. Within Canada, Sandra Burrows reported that most newspaper digitization projects within Canada are still thematic rather than systematic in nature such as on the "opening of the West."10

A number of national digitization projects are also gaining steam, particularly in the United Kingdom. A recent article by Edmund King of the British Library argued for the importance of automated indexing to provide searchable text, as well as an ability to search by topic, date, range of dates, personal names and keyword search terms.11 His points illustrate the importance of named entity searching within historical newspapers, although it is rarely supported.

Jane Shaw, also of the British Library, has also recently detailed the findings of the British Newspaper Project.12 One of the biggest issues they faced was the poor quality of their source material. Their project began in 2004 and was to include scanning of entire microfilmed content, as well as article zoning and page extraction. The main objective was to provide up to two million pages of British national, regional and local newspapers from 1800-1900, and to offer a sophisticated searching and browsing interfaces that will support names and dates, as well as obituaries and advertisements. Again, the need for named entity searching was evident through their digitization project.

In terms of article segmentation, they found that "The density of nineteenth century texts makes machine identification of breaks between articles a more difficult task thus requiring a balance between automatic metadata generation and human intervention. Working with a supplier with many years experience, the BL took the view that human intelligence would give the best quality result and therefore shaped the project around computer assisted/human intelligence throughout the whole cycle." Their results reflect our own findings of the importance of determining the right combination of human and machine intelligence for automated processes. Their project is supporting four levels of metadata: title, issue, page and article. Since many of their articles do not have titles, they have constructed them from the first few lines of text. Page images will consist of both articles with search terms highlighted and full pages. In addition, they are also providing article level categorization, including: "news (domestic), news (foreign), advertisements & notices, arts & popular culture, births, deaths, and marriages, obituaries, court & society, crime & punishment, commerce, letters, sports, editorial comment, miscellaneous, none, and illustrations."

---

[8] Arlitsch, Kenning and Jeff Jonsson. (2005)."Aggregating Distributed Digital Collection in the Mountain West Digital Library with the CONTENTdm multi-site server." *Library Hi-Tech*, 23 (2), pp. 220-232.
[9] Gilboe, Lynda James. (2005). "The Challenge of Digitization: Libraries are Finding That Newspaper Projects are mot for the Faint of Heart." *Serials Librarian*, 49 (1/2).
[10] Burrows, S. (2005). "Canadian Developments for the Digitization of Newspapers." http://www.nla.gov.au/initiatives/meetings/newspapers/program.html International Newspapers Conference
[11] King, Edmund. (2005). "Digitisation of Newspapers at the British Library." *Serials Librarian*, 49 (1/2).
[12] Shaw, Jane. (2005) "10 Billion Words: The British Library British Newspapers 1800-1900 Project: Some Guidelines for Large-Scale Newspaper Digitisation." World Library and Information Congress: 71th IFLA General Conference and Council "Libraries-A Voyage of Discovery", April 14-18th 2005, Oslo, Norway. http://www.ifla.org/IV/ifla71/papers/154e-Shaw.pdf

A recent newspaper digitization project has also been undertaken by the Tuzzy Consortium in Barrow, Alaska.13 They differed from many other projects by using contractor supplied software to enable article segmentation and metadata markup locally and by using Greenstone as their repository software. Over 35 years of the Tundra Times were digitized, and they used ABBYY FineReader for their OCR tool. Most of the work was entirely outsourced and consequently produced by IArchive, including scanning, generation of image files, OCR, generation of XML, OCR text and metadata. A Greenstone consultant helped them develop a XML plugin to enable search term highlighting within PDF Searchable Image files, although this turned out to be a complex procedure. All of their metadata is created by human beings, and a human technician was responsible for assigning page and article level metadata, the articles were also tagged with byline, headline, classification, and if it was a lead story.

6g. Semantic Searching and Historic Newspapers

Perhaps the greatest area of growth in the last two years is the development of advanced semantic and other language technologies and their use in historical newspapers, particularly in Europe.

A system developed for the Vikelea Municipal Library of Heraklion includes OCR based page analysis, article level metadata generation, and semantic indexing and classification of articles using a specially developed thesaurus. Doerr, et. al. believe that such a precise level of metadata is necessary because, "An important part of the study of historical newspapers consists of classifying the material and annotating it such that its future retrieval is made easier."14 Their system focuses on the notion of a "segment" as the basic conceptual unit, which can consist of one or more parts of the newspaper document that are relevant. After full text has been generated by OCR, expert users are helping them generate metadata for each annotated segment based on the CIDOC CRM ontology. Their system includes three basic parts 1) a digital library system that deals with archival preservation, and thematic indexing, the core of which is a Fedora repository with its functionality enhanced by using a thesaurus management system that helps users classify and retrieve articles, 2) a documentation tool that provides a web interface that allows users to isolate specific entities/segments within documents and create metadata on the fly 3) administrator tools that allow for mass storage and transforms images into materials that can be indexed by users using the documentation tool.

Similar work has been conducted by the Neptuno project in Spain which is attempting to apply Semantic web technologies to improve access to newspaper archives, albeit more modern ones.15 They developed a platform that included an ontology defined in RDF for new archives based on journalist and archivists expertise. Their work also included developing a knowledge base where materials in the archive are described using the ontology. In the process, they discovered that editors searched for information in very different ways than it was annotated by archivists. By conducting a DB to ontology conversion, they were able to automatically integrate existing legacy archive materials and support a semantic search across the entire archive. They make several points that are relevant for all historical newspaper archives, however, by critiquing the fact that there is no support for conceptual searching, it is difficult to integrate various archives, platforms cannot be easily extended, browsing is not supported, and there is no semantic searching. Their final contribution they believe is the development of a semantic search module, which takes advantage of semantic information in the knowledge base and thus allows users to formulate more precise and expressive user queries. "A semantic search engine has knowledge of the domain at hand." Castellis, et. al. explain, "The availability of a domain ontology that structures and related the information according to its meaning allows the implementation of a search system where users can specify search criteria in terms of modeled concepts and attributes."

---

[13] Terpstra, Judith A. K, et. al. "The Tundra Times Newspaper Digitization Project." *RLG Diginews*, Feb 15, 2005. http://www.rlg.org/en/page.php?Page_ID=20522&Printable=1&Article_ID=1706

[14] Doerr, Martin, et. al. (2006). "Digital Library of Historical Newspapers." http://eprints.sics.se/299/01/index.html, July 2006

[15] Castellis, P. et. al. (2005). "Neptuno: Semantic Web Technologies for a Digital Newspaper Archive." *Proceedings of ESWS 2004*, pp. 445-458.

6h. Conclusion

As this brief research review illustrates, there have been a number of recent developments in the field of historic newspaper digitization, as well as recognition of the importance of more sophisticated means of access. The Perseus Project is continuing to study the findings of these other groups and explore how named entity recognition and other advanced language technologies might be further used to enhance access to the newspapers that have been digitized for this project.