"A Testbed of Civil War-Era Newspapers"

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

Semi-annual report, March 2005-September 2005

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. Work on outcomes assessment using the project's Logic Model
3. Copyright permission secured
4. Progress on developing best practice guidelines, including cost/benefit
5. Progress on implementation of FEDORA/DAMS at Richmond
   - evolution of use of DLXS for delivery
   - Fedora Users conference
   - AGU scripts from Carter Glass
   - License XPAT; XPAT workshop
   - Installation of XPAT/DLXS
6. Workflows and production
   a. Production:
   b. Workflows:
7. Issues in automatic tagging of newspapers

1. No-cost extension request granted

On August 11, 2005 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: 09/30/2005
> Modified award dates: From: 10/01/2003; To: 09/30/2006
> Revised reporting schedule is included below.

Type	Due Date	Date Received	Delinquent
Final Financial	12/29/2006		No
Final Narrative	12/29/2006		No
Interim Narrative	05/01/2006		No
Interim Financial	10/31/2005		No
Interim Narrative	10/31/2005		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. Work on outcomes assessment using the project's Logic Model

As explained towards the end of section #5 below, delays in release of the new version of DLXS software will delay the public debut of the project's Web site offering access to digitized content. We anticipate being able to go live with the site in early January. This forces a revision to the schedule for some assessment activities. Appendix #1 to this report is a revised version of our assessment logic model. Dates highlighted in light green are the revised dates.

3. Copyright Permission Secured

After discussions with ProQuest, producer of the microfilm version of the Richmond *Dispatch*, we secured permission to use scan their microfilm:

From: Lee-Perriard, Marta [mailto:Marta.Lee-Perriard@il.proquest.com]
Sent: Monday, August 15, 2005 10:38 AM
To: Rettig, Jim
Cc: Norris, Kevin
Subject: Thank you

Dear Mr. Rettig,

We are eager to support digitization projects and appreciate your willingness to share the details of your project with us. Thank you, in particular, for letting us know about your plans to digitize using our microfilm. We are glad to make these years available for such a worthwhile project.

I would like to thank you for offering to include an acknowledgement to ProQuest on the webpage. We would very much appreciate it.

I'm looking forward to seeing the results. Please feel free to contact me with any questions.

All the best wishes,

Marta Lee-Perriard

Publisher, Historical Newspapers
ProQuest Information & Learning
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, Michigan 48106-1346
Tel: 734 761 4700 Ext. 3160
Fax: 734 975 6271
Email: marta.lee-perriard@il.proquest.com

4. Best Practices Guidelines

As we finish with the process of outsourcing digital data creation from our vendors, we will begin analyzing cost benefit strategies in regards to digitizing 19th-century papers. Working with two other IMLS funded Newspaper projects, Colorado and Utah, we will be able to compare and analyze the costs between projects – the report on this data will be included in our final IMLS report. This matrix was developed by Brenda Bailey Hainer of the Colorado project and was approved by John Herbert, Utah, and Rachel Frick, Richmond. A sample of the various data collection points we will be using is included in Appendix #2 of this report.

Another information sharing point with the Utah and Colorado group is user survey data. This survey is also included in the appendix. The members from the three projects met at the symposium, *Digitizing Historic Newspapers: A Practical Approach* on July 18, 2005 in Denver, Colorado, and discussed using the same exact user survey in order to compare use data across projects. We also decided to try and steer users using the same language and visual cues from the projects' front splash pages. For example, Utah's user survey can be accessed at http://www.lib.utah.edu/digital/unews/. (This user survey is reproduced in Appendix #3 of this report.) Utah and Colorado will be giving a preliminary report on user response in their October 2005 final reports. University of Richmond will follow up on this data in Spring 2006, in our final report, analyzing user data collected from our project, and data collected from the two other projects since October.

Rachel Frick and Andrew Rouner have been presenting at regional, national and local conferences, discussing the IMLS project's process and the tools and workflows we have developed.

In June 2005, Rachel presented a poster session at the Northeast Document Conservation Center sponsored, School for Scanning. The poster compared and contrasted the IMLS newspaper project with the University of Richmond's digitization of 90 years of the student newspaper. On June 20, 2005, the *Richmond Times- Dispatch* published an article on the IMLS project - http://www.timesdispatch.com/servlet/Satellite?pagename=RTD/MGArticle/RTD_BasicArticle&c=MGArticle&cid=1031783385202 This article was picked up by the Associated Press and was republished in several regional papers, including the Virginia Pilot, and linked to by various Civil War history web sites.

In October 2005, Andrew Rouner and Rachel Frick will present at the EDUCAUSE annual conference in Orlando, FL. The focus of this presentation will be the workflows and tools developed during the course of this project and how they could help small to medium sized organizations in starting or to further develop their digital library activities. Ms. Frick is also presenting at the Virginia Library Association's Annual conference in Williamsburg, VA about outsourcing digitization services and will discuss helpful workflows, tools, and communication guidelines the University of Richmond has created in the course of our IMLS project. In December of 2005, Andrew Rouner and Rachel Frick are scheduled to speak at the symposium on Digital Asset Management Systems sponsored by the Associated Colleges of the South.

5. Progress on Digital Asset Management Systems

There has been a shift in the strategy for the *delivery* of project content in the Richmond digital library, as a result of further information and exploration in the process of designing a digital library asset management system.

Early on in the project, when various content-delivery systems were being explored, Richmond was in particular considering DSpace and Fedora. Fedora was chosen because of its potential to do much more in terms of file formats accepted, versioning and (potentially) delivery. For those and other reasons, the decision was made to base the project on Fedora. One feature advertised by the developers was searching. When asked at lunch following his presentation at the University of Richmond in July, 2004, one of the co-developers of Fedora, Thornton Staples, noted that Fedora's internal searching capabilities were weak. Fedora only searched the metadata (the Dublin Core information in a Fedora digital object). Staples mentioned that the development team was looking into the possibility of incorporating *full-text* searching capabilities into future versions of Fedora (he mentioned Lucene and eXist) but that neither seemed ready at this point.

Since the project had opted for a relatively narrow range of newspaper publication, instead of a large range of images and uncorrected OCR text (which produces at best unreliable search results) and furthermore (for the Richmond *Daily Dispatch*) chose keyboarding of highly structured text in TEI XML, it was imperative that very robust text-indexing software be incorporated. This requirement is all the more essential, given the work in automated tagging of name authorities being done by the Perseus project in the grant. Again, without robust text-indexing software that allows for excellent delivery of results as well, it was agreed that these efforts at higher-level tagging would be wasted.

Having looked into Lucene, eXist, Zebra and others, it was finally agreed that Richmond would license XPAT from the University of Michigan to fill this need. XPAT is the XML- and Unicode-aware text indexing software known previously at PAT 5 when commercially marketed by the Open Text Corporation. When Open Text Corp. ceased development on PAT, the University of Michigan came to an agreement

with them to continue PAT's development at the Digital Library Extension Service (DLXS) and sub-license it to other institutions, primarily academic. While not an open source application, per se, XPAT is nevertheless dramatically less expensive than commercially marketed comparable solutions, such as Tamino and Xyleme, which run upwards of $80,000. Richmond purchased a license and support agreement late in the 2004-05 fiscal year with the intention of integrating XPAT and Fedora, using the former along with a series of PERL scripts to deliver search results, and Fedora to supply browsing functionality.

Shortly before licensing XPAT, Andrew Rouner attended the inaugural Fedora Users' Conference, sponsored by Rutgers University, in May. The most useful result of the conference was seeing the possibilities of development with Fedora, on the one hand, and its practical limitations. However, these applications were the result of very sophisticated and specialized programming efforts. Thornton Staples, co-developer of Fedora, has always insisted that Fedora is an architecture, not a "solution," despite some of its immediate capabilities. It became clear that for Richmond to achieve even basic content-delivery capabilities through Fedora would require a significant programming effort to develop Fedora behaviors and disseminators.

One presentation at the Fedora Users' Conference, however, was compelling in a very different way. As opposed to developing content-delivery on top of Fedora, the presentation by Carter Glass of the American Geophysical Union underscored Fedora's underlying and current strength, which is not as a content delivery system, but as an archival and versioning system. Glass was also building on Fedora, but primarily to ingest content into the Fedora system. Andrew Rouner approached Glass at the conference to discuss his Fedora ingest scripts, and Glass indicated he was willing to share them.

On June 23, Rick Neal and Andrew Rouner visited Carter Glass at the AGU in Washington. In that meeting Glass explained in greater detail how the PERL scripts functioned. The scripts were coded with a config file, and therefore were intentionally adaptable by other institutions. Glass later sent an archive of the scripts, and these scripts will be modified at Richmond to automate the creation of METS files and auto-ingestion of Fedora digital objects generated from the TEI XML files of the Richmond *Daily Dispatch*. Programming assistance from other departments at Richmond was not available until the fall, so efforts were focused instead on completion of XML file checking, deferring delivery issues.

The primary support opportunity provided by DLXS for XPAT is an annual workshop run by the developers in Ann Arbor, Michigan. Andrew Rouner went to Michigan for this workshop. It is of course also an opportunity to learn more about the product, and there was in fact a great deal more to the product than had been anticipated. The traditional strength of PAT 5 had been its abilities to index highly-structured texts, as well as very fast retrieval of search results and the ability to create a link to the result in context, (being a text-oriented rather than documented-oriented search tool, which most others are). In addition to adding XML and Unicode support to XPAT itself, the DLXS has developed extensive middleware (also called DLXS) which itself *is* open source, as an interface for both search results *and browsing*. Having been developed for several years now, it is a mature delivery system. While Richmond had licensed XPAT strictly on the basis of its search capabilities, Rouner learned that the DLXS middleware also had extensive browse capabilities, including a page-turner, where a user can switch from page-image to text and back, while the DLXS middle ware tracks the page, and this middleware deserved consideration as a delivery solution in Richmond's digital library.

It became apparent that the DLXS middleware would allow for sophisticated content delivery for project material, without the extensive additional programming required by Fedora. At the same time, while the Fedora user community will undoubtedly begin to share some of the behaviors and disseminators that have been developed, and the base architecture of Fedora itself will continue to improve, at present Fedora requires considerable additional programming. (One alternative is the Elated application, developed by ACS/NITLE, and built on top of the Fedora architecture. For various reasons, including the problem of the integration of robust text indexing with Fedora and Elated, Elated was not finally seen as a workable solution for Richmond's purposes.) At the same time, as underscored in Glass's presentation, Fedora's strength in its current version, and with the current state of user community contributions, is not in its content delivery, but in its archival capabilities. Viewed from a *curator's*, rather than an end-user's perspective, the modest search capabilities extending only to the digital objects' metadata, make sense and are sufficient. One of the primary needs Fedora addresses is simply the problem of *keeping track* of ever-expanding digital libraries in diverse formats. It is important to recognize, however, that this *is a separate*

problem from that of the delivery of digital resources, even though it ultimately makes sense to deliver digital content through the same architecture that tracks, manages and assures the integrity of that content. But for institutions lacking extensive programming resources to develop delivery solutions for Fedora, that is not yet feasible.

The modified plan for Richmond, then, is to employ the current strengths of these two important pieces of digital library infrastructure—the full-text searching and content-delivery capabilities of XPAT/DLXS on the one hand, and the archival, storage and versioning capabilities of Fedora, on the other—side by side. As content is added and improved and more basic programming requirements are addressed, Richmond will begin to look into the possibility of integrating these two pieces. At that time, behaviors and disseminators for Fedora will likely have been contributed by the user community, and Richmond hopes to be in the position of being able to make similar contributions as well.

At present, XPAT and DLXS Release 11 has been installed on the IMLS server. Fedora 1.2.1 is also installed on the IMLS server, and the two architectures are co-existing peacefully. The demo objects have also been ingested on the server installation of Fedora. Richmond is remaining with Fedora version 1.2.1 for the time being, in part because the Fedora ingest scripts from AGU were created to work with this version. Release 12 of DLXS has been pushed back until October 31, so this will not be able to be installed until then. However, Release 12 adds support for non-continuous tone images (i.e., jpegs) used in the page-turner application, which will be essential for our project. We will load data and modify content delivery in DLXS, and anticipate user access of the XML and images for Richmond *Daily Dispatch* by November 30.

6. Workflows and Production

6a. Workflows

In our previous report, we noted the recent installation of Subversion (SVN) on the IMLS server, to facilitate correction and/or checking workflows done primarily by student workers. At that time, SVN had just been installed. Prior to this students only had access to XML files and images through an external hard drive. It was difficult to pass this between students, much less to the Digital Resources Librarian for retrieval of corrected files. Workflow has been dramatically improved through a workflow based on SVN.

While in early 2005 we were able to reduce the correction time on a given file by the Student Metadata Editors to around three hours, we found that too many files needed to be returned to the keyboarding vendor for correction. It would not be possible to determine all the files that needed to be returned through correction of files, and have them returned to the vendor, in a timely fashion. It became clear that 1) a regular process needed to be established for returning files to the vendor, and receiving them back, and 2) it was necessary to shift from a workflow of file *correction* to one of file *checking*.

With SVN installed on a server, users are able to *access* (not *download*) XML files through a select number of XML editors through the WebDAV protocol. Fortunately, that included both oXgen, and XMLSpy. Because files are not downloaded, but edited on the server, there is no possibility of local files being lost, or of generating multiple versions of the same file. On the other hand, because it is a versioning system, there is no danger of files being overwritten, and users may actually work on the same file simultaneously. The use of SVN on a server accessed through an XML editor is a relatively new application of versioning software. The more traditional application of a CVS is one where all files are downloaded to a given user. SVN also fills this need, and therefore has server and client software. The use of this technology potentially interfered with the Digital Resources Librarian's need to make global changes on all files (and therefore to have access to all files at once). But this turned out not to be a problem, because the client version could be installed on desktop that had a UNIX operating system (in this case, Mac OS X). Not only did this solve the problem of global access to the files, but it also turned out to be an effective serves as a means to monitor students' progress on the files. When the svn update command is run in the terminal, all new and modified files are downloaded to the client SVN program (on the desktop), and the modified files are listed in the process. SVN also tracks which users are modifying which files, and so forth. All this information is especially useful in a context in which the supervisor is not able to work physically in the same room as student workers. By the same token, global changes made to local files on the client version can be uploaded with the svn commit command, which puts the files into the SVN repository on the server, and without fear of over-writing changes made by student editors. SVN detects where there are conflicts in files, and files can then be reconciled as needed. Through a combination of compressing tiff files into archives and wringing out every last bit of server space

(including serving from the Digital Resources Librarian's desktop) we were able to make all project images available to the student editors via download, and were therefore able to move to an entirely server-based workflow, and dispense with the HDD.

Another open source UNIX program, xmllint, fits particularly well with the workflow of a supervisor using the SVN client. A few years back, the installation of an SGML parser had to be done on a server, and was a complicated and expensive endeavor. Now, xmllint is included with the libxml2 package, and can be used to parse all the XML files in a directory against a specified DTD. While the parsing of all files for any given occasion by opening them into an application like oXygen would simply not be feasible, the work that would involve makes the fact that xmllint can parse 1300 long and complex XML files in roughly two minutes all the more impressive. This is another small, but critical piece in the development of a viable digital library infrastructure at the University of Richmond, which includes a content workflow.

To track which files have been processed under this workflow, a simple text file with dates for a given week was put in the SVN repository on the server. Students add to to the list as they complete checking a given file. At the end of the week, entries from this file are recorded in the "imls_xml_filetracker" spreadsheet by the supervisor (used to track files originally received from vendor, as described in the last report). Those files being returned to the vendor for correction are also recorded in the "ddd_returned_files" spreadsheet. However, these transactions are not simple transfers of information from one list to another. Because all XML files (and the text file listing completed files) are accessible to the Digital Resources Librarian via the SVN client, the files on the list can be opened, and entries made by the file-checkers in the Headers of the XML files are checked to verify that the file needs to be returned, and/or that the problem indicated does indeed exist, before submitted for return.

We not only had to adjust our workflow because we realized too many files failed to meet specifications and would have to be returned to the vendor, we also had difficulty with the vendor not making corrections *on the version of the file we had returned to them.* (To this point, the origin of the files they did return to us remains a mystery.) Thus, even though a file corrected by the vendor and returned to Richmond might be "improved," because it was not corrected from the version sent to them from Richmond, any changes or corrections previously added were lost. This was another issue that was addressed by creating a more formal system for the return of files to the vendor. A password-protected "vendor" web page was created which had several project spreadsheets and documents (including, for example, the specifications document). Typically on a semi-weekly basis, information from the text file on the server listing processed XML files would be transferred to the "imls_xml_filetracker" spreadsheet as just described, and the XML files to be returned to the vendor would be prepared. The spreadsheet also indicates the date on which a given file was returned to the vendor. The XML files to be returned on a given week are put into a folder and zipped, so the vendor can download the archive.

At this point, each one of these pieces of the workflow fits with another in a significant way, and the workflow we've established with these technologies could be applied to any number of digital projects in the future, including collaboration with external partners.

6b. Production

With the new workflow of checking XML files, we are approximately 90% finished. Of the Richmond *Daily Dispatch* XML files, about 35% have needed to be returned to the vendor. About 5% of the files returned to the vendor have been corrected and delivered to Richmond. We anticipate complete shipments of all files returned to DDD by then end of November, 2005. The original target of completion of all re-keying of Richmond *Daily Dispatch* would have been met, except for the high rate of Q-A failure by vendor.

<u>Vendors – Digital Divide Data</u>
1355 issues of Richmond *Daily Dispatch* (Nov 1860-Apr; Dec 1865)
(6 days a week; approx 4 pages per issue)

1355 fully re-keyed and encoded in TEI XML text files
4092 tif images
4055 jpg images

<u>Vendors – Byte Managers</u>

approx. 367 issues Boston *Liberator* (Jan 1859-Dec1865)
(weekly; approx 4 pages per issue)

1,474 txt files corrected OCR and tagged text files
1,479 tif images

1,355 issues of Philadelphia *Ledger* (Dec 1860-Oct 1865)

uncorrected OCR and two months (for Battle of Gettysburg, and two months: April& May 1865, to cover time when Richmond was evacuated and there was not any *Richmond Dispatch* issues produced) corrected OCR and tagged text files to be delivered November, 2005

5,016 tif images

7. Issues in automatic tagging of newspapers

Evaluation Process

While automated systems may never perfectly grasp the nuances and complexities of the language in historical newspapers, the amount of labor they save in manual identification and tagging of various entity types is impressive. Human error checking of automated output may still be a necessary task, but the combination of automated systems and human effort can lead to a workable and efficient system.

Much of the current work with newspaper tagging has involved checking the output of the automated system. This has involved the manual creation of a number of limited authority lists including commodities, company and organization names. This work has also included the identification of terms/entities that needed to be deleted from computer generated authority lists or put in as stop-words or negative terms that should never be tagged as a particular entity. These authority lists at the moment are simple text files that are edited in a spreadsheet. Future work will involve exploration of how to best encode these growing authority lists (such as in the MADS format proposed by the Library of Congress), so that they can potentially be utilized with other digital historical Civil War materials to support named entity tagging.

Another important outcome of manual evaluation of the newspapers is the identification of expressions or words that need to be added as stop-words or negative terms that should never be tagged as organization names. Examples include “unlawful assembly” for the “assembly” organization type, “affairs” and “fair daughters of Richmond” (of which there were apparently many, due to the frequent use of this expression) for the “fair” organization type and “congress gaiters” which were a popular type of shoe for the organization type of “congress.”

Manual evaluation of newspaper tagging has also involved the identification of regular expressions or syntactical patterns that may assist the system in “learning” to better identify entities. Since the named entity system currently uses both rule based learning and statistical analysis as major parts of its semantic tagging, the identification of problematic patterns and incorrect statistical classifications is very important. For example, one problematic grammatical pattern that was identified was that a string of terms ending with an organization name type word such as “bank” or “bureau” of “court” would tag as an organization. Some examples include “tables, washstands, bureau” tagging as a bureau, "office, custom house, bank" tagging as the "Office Customhouse Bank" and "Manners, Morals, Court” tagging as the “Manners Morals Court.”

Evaluation of the automated tagging in newspapers has involved several steps. One step began this spring was the “perfection” of several newspapers, where several entire issues of the *Richmond Daily Dispatch* were examined and all entity tagging errors fixed. Another type of evaluation involved the creation of computer generated authority lists of organization names and commodities that were then manually checked by project staff. Incorrectly identified entities were systematically tagged to be fed back into the system. This section will include a discussion of what we have learned so far regarding particular types of entities and how they appear in newspaper text and some future efforts to improve the tagging system.

Tagging of Personal Names

Personal name tagging within newspapers is quite challenging, since hundreds of distinct individuals can typically be identified in just one issue alone. Many references to individuals include a surname only, an abbreviated name, or refer to an individual that cannot be found within currently existing authority lists such as the Library of Congress Name Authority File. The names that are automatically extracted are currently being stored in a relational database, and we are currently exploring how metadata formats that are native XML such as MADS records might be used to store and structure this data.

A variety of errors were identified in the tagging of personal names. The system currently has some trouble identifying honorifics and other titles. Of particular issue is when a name has two honorifics, like the "Rev. Dr." The system sometimes fails to tag these titles or tags each of these titles as a forename. Suffixes like Jr. or the IV or III are also currently not tagging. A related problem is that "Esq" often tags as a surname. Nationalities such as "French" "African" and "English" also frequently improperly tagged as surnames or forenames. A similar matter is that many individual names in newspaper text include abbreviated first names, such as "wm" for William, "jno" for "Jonathan", "Chas" for "Charles", "geo" for "George" to name only a few. When these abbreviations do tag, they tend to tag as a single forename and are not attached to the surnames names that typically follow. This leads some names to be missed and many partial names only to be tagged.

Occasionally certain grammatical patterns also led to tagging of phrases or expressions as personal names. Whenever two capitalized proper nouns occurred in a sentence, they were typically tagged as a proper name. This same error could also happen whenever a string of nouns appeared separated by columns. Newspapers made frequent use of all capital letters as well as strange capitalization of proper nouns, depending on the context. This makes the systems use of certain grammatical or syntactical patterns which might accurately tag personal names in other types of texts problematic for newspapers. A number of phrases that were incorrectly tagged by the system as proper names in its first pass include "Anglo Saxon" "Double Refined" "Medical Students" "Good Harness" "Bill Head" and most entertainingly, "SECRET DISEASES" which had a tendency to appear in patent medicine advertisements in all capitals.

Another difficulty is that many common personal surnames and forenames can also have a place name or proper noun context, such as Banks, Black, Cash, Church, Day, Price, Rice, White, Winter, and Young for surnames and Bill, for forenames. Tagging of the term White proved particularly frustrating, as sometimes it would tag for a color and other times its use would lead to the tagging of expressions such as "White Persons" as a person's name.

Role names, particularly those of royalty, also led to some interesting tagging results. A number of newspaper articles frequently reference European royalty, so the tagging of role names using preexisting TEI tags such as Lady, Baron, Lord, etc. has been attempted. The role name "Lady" proved most problematic and often the generic use of the word lady, or ladies, such as the expression "Ladies of Richmond" would lead to the tagging of a personal name such as Lady Richmond. In one case where an article detailed that a young lady had been murdered, her identity was tagged as "Lady Shot" since the headline for the article read "YOUNG LADY SHOT" in all capital letters. Another issue was the tagging of names such as "Duke of Newcastle" which would typically tag the term duke as a role name but then Newcastle as a placename, rather than tagging the whole expression together as one name. Another challenging role name was "Miss" where expressions such as "Miss May" or "Miss Semon" had "Miss" tagged as a forename rather than as a role name.

Tagging of Commodities

The automated tagging of commodities proved to be very difficult. Not only were there commodities in the nineteenth century that are no longer sold or consumed today such as "burning fluid", but the same commodity may go by a number of names or be represented several ways in the text such as "windowglass" "window-glass" or "window pane glass." Refining the results of this system has at times proved both entertaining and quite challenging. The identification of nineteenth century commodities that no longer exist often involved the use of the Oxford English Dictionary Online to determine what a particular word meant and if it was being used as a commodity. One popular sales term for advertisements in the nineteenth century was "furnishing goods" yet the system routinely tagged this as a generic commodity labeled "goods."

Although the accurate tagging of commodities is quite complex, previous research conducted for this project illustrated that newspaper advertisements were one of the most heavily used sections of newspapers by historical researchers. Their tagging, however, presents a number of problems. To begin with, commodities with multiple terms in their names such as "anthracite coal" and "cider vinegar" are frequently missed because the system tags the first as "coal" and the second tags only as "vinegar." This brings up the question of how specific a level of tagging we hope to support, would someone want to search for cider vinegar or would vinegar be granular enough? In addition, words such as "goods" and "stock" are currently tagging as commodities but these may be too general to support useful searching.

One major issue yet to be resolved is to try and figure out a way to tag multiple commodities such as "gold and silver watches" with the single commodity of watches, rather than as three commodities "gold" "silver" and "watches." Another question is how to either capture or filter out the tagging of brand names as surnames. For example, the "Willcox Gibbs Sewing Machine" was a popular model for sale, but Wilcox and Gibbs tag as individual surnames.

Tagging of Place Names

The tagging of place names in newspapers presents a number of unique problems. The system currently in place uses the TGN to identify possible places, but this gazetteer does not include local place names such as "Broad Street Hotel" "Schad's Hall" "Jefferson Ward" or "Shockoe Slip." Thus part of our work entails identifying those local place names that appear prominently in the newspapers to add to the place name authority lists that we are in the process of automatically creating. Another issue is correctly identifying local place names that could also be person names. Phrases such as "King William" and "Princess Anne" consistently tag as people though almost all references are to the counties in Virginia.

One of the most difficult types of place names to tag is street names, particularly street intersections such as "Marshall and Clay" or "7th and Franklin." Frequently, these terms will tag as stand alone surnames rather than place names. We are currently examining adding in heuristics to the system that will not only use a "gazetteer" of street names but will also take into account the context or words which appear in the text around those street names. For example, we found that if a term such as "Cary" or "Marshall" appeared before word such as "corner" or "between" it was almost always a street name rather than a surname. A number of street names such as Cary typically tagged as surnames, even though they typically referred to streets. Fixing this error will most likely involve altering the statistical model for these terms.

A number of abbreviations also caused some place name tagging errors. N.B which was often used for "nota bene" in advertisements kept tagging as New Brunswick, Col. typically tagged for Colorado, when it almost always stood for Colonel, and Geo. kept tagging for Georgia when it was used for the name George over 90% of the time. Another frequent error encountered involved the use of the word new. Whenever the word "New" appeared in all capital letters in front of another proper noun all in capitals, it tagged the phrase as a place name, for example "NEW BOOKS" or "NEW OFFER" tagged as places.

Another issue was the assigning of incorrect place names to terms that were correctly identified as places. The most frequent example was Broad Street, a major thoroughfare in Richmond that persistently tags as Broad Street in Kent, England. Similarly, the use of the local place name "Club House" an actual location in Richmond continues to tag as Clubhouse Crossroads, South Carolina. We are exploring how to correct these errors.

Tagging Issues With Organization Names

Currently we are experimenting with tagging a number of different organization types, including companies, ships, military companies, associations, ethnic groups, courts of laws, and schools to name a few. This section will examine the issues inherent in the automated tagging of each of these types of entities. A significant part of this work has involved the manual creation or correction of authority lists for each type of entity encountered in order to improve the next round of automated tagging.

Company Names

Perhaps the most frequently found organizations within historical newspapers small businesses or companies owned by local individuals. Almost half the text of nineteenth century newspapers such as The Richmond Times Dispatch consisted of advertisements. Many of these were purchased by individuals such as lawyers, druggists, storeowners or by small companies. One issue in tagging of these organizations is that many of these company names consist of a list of surnames such a "Harris, Spicer & Harris." These

strings of text currently tend to tag as a list of individual surnames when in truth it is an organization. The variety of naming patterns for small businesses in newspapers was extensive and the same company name could include abbreviations one time while including the full name another such as for the "Chas. T. Wortham & Co" which also contained listings as "Charles T. Wortham & Co". While the use of certain grammatical patterns such as "& Co" or more specifically the XML string, "& Co" may help to identify text strings as organizations, not all companies contain this expression.

While some more major companies such as the "Powhatan Steamboat Company" and the "Richmond Trunk Factory" are relatively easy to identify, the rules for tagging these names, such as "tag all expressions that occur with the term "company" and "factory"" can have unforeseen problems. The use of terms that are too common or generic as patterns in identifying organization names can lead to the tagging of sentences where factory or company are used as nouns and not as part of name expressions.

Another difficulty in identifying company names is that both businesses and military organization can contain the key term "company" such as "Company A." This is an example of where an authority list generated from a 19th century city directory could be invaluable in providing the system with a list of local business names to tag that can be added too as more names are discovered. A preliminary list of company names has been created while checking the results of automated tagging. Preliminary examination has illustrated that a small number of companies accounted for a significant number of advertisements which may make creation of a manual list feasible.

One frequent error in the tagging of company names was that occurrences of the term "co" or "co." led to a tag for a company name. Yet this word was usually a reference to a county, particularly "Chesterfield County" and "Kanawha County." A typical problem pattern was "city name, county" such as "Clever Depot, Halifax Co" which then tags as the "Clever Depot Halifax Company."

Newspapers

Tagging of newspapers continues and reference works such as Rowell's Directory of American Newspapers have provided much needed assistance in successfully identifying newspaper titles. A large number of the stories in nineteenth century newspapers are quoted or borrowed wholesale from other newspapers.

Banks

The identification of banks can be challenging and river banks have on more than one occasion been tagged as financial institutions. Examples include the frequently tagged "Bank of the Mississippi" and "Bank of the Rio Grande" when all references were to the rivers. We are examining adding in a heuristic that will prevent any banks with a lower case b being tagged as organization names, because this use typically indicates the bank is a natural feature

Another issue that came up was the tagging of expressions such as "shares city bank" "shares continental bank" as full bank names. Financial sections of the newspaper typically listed all share prices in this manner, leading to a number of incorrect tags. In the newspapers these terms are all typically run together without punctuation to indicate share prices for a particular bank

Ships

Perhaps one of the most difficult entities to consistently identify have been ships. Certain contextual clues are helpful, such as that newspaper sections entitled "SAILED" and "ARRIVED" all in capital letters or "Marine Intelligence" tend to indicate a section that will list recent departures and arrivals of ships. A list of ship "term types" has been created such as "Canal boat" and "packet schooner" to help the system identify ship names. In newspapers, ship names typically appeared after a listing of the ships type, most frequently in all capital letters. Manual evaluation has also proved very important in this process. Checking of automated tagging led to the identification of "schr", an abbreviation for schooner, as an important "term type" to add into the list of term types for ships.

Military Organizations

One of the most difficult types of entities to identify in newspapers is military organizations. The Perseus Project has digitized a number of military reference works that include lists of military organizations that may assist in the disambiguation of these organizations in newspaper text. One major problem is the variations in how military organization names can appear such as "1st Reg Volunteers" and "1st Reg Va Volunteers" both of which upon examination are the same regiments. Many local companies

are also referred to simply as “Armory Co. A” or “Armory Co. B” without any more contextual information that helps identify them. Part of our current task is to identify those sections of the newspaper that typically provide a listing of military units, in order to examine patterns of how military units are typically labeled or named in newspaper texts. One frequent section title identified was “Orders.” We found that many military organizations were listed in combined expressions such as “Frontier guard, lane’s company” which would then lead the system to tag it as the “Frontier Guard Lane’s Company.” We are examining how to get these items to tag separately. A related issue was that many military company names are simply listed by their captain’s names such as “Lovell’s Company” and the system thus tags it as a company without any reference to the formal company name. We hope to use some of our historical reference works to link these entities to their official military units.

A related issue is what level of specificity we want to support in the tagging of military organizations. Items such as “French fleet” and “French riflemen” are currently tagging, but often lead to tagging of generic expressions rather than actually military organizations. Evaluation of several newspaper issues also led to the identification of a number of military terms that were not tagging as organization names and needed to be added to the list of military term types such as “regiment” “battalion” “greys/grays” “cadets” and “troops”, for we were missing expressions such as “New York 69th regiment” and “Carolina grays”. It also led to the identification of several negative phrases to be added to a list of stop-words, including, "national guard hat" "such rifles" AND "colt's rifles."

Problematic Organization Types To Eliminate

Currently Perseus is experimenting with automated tagging of a broad variety of organization types. Several organization types are being considered for deletion due to problematic tagging

Establishment

We have experimented tagging the organization type “establishment” but the system simply tagged uses of this word in noun phrases or as a verb. None of the expressions tagged were actually organizations. Examples include the tagging of “bargain establishment of Alfred Moses” or the “northern establishment.”

Station

While “station” is currently being tagged as an organization type, it is possible that this tag might be more appropriate as an attribute type to be used with place name tags. While most of the entities tagged were actual stations, they were typically city names or railroad stations. Examples include “Union station” and “Liverpool station.” One issue was that this organization type was also tagging expressions such as “station of vice president” and occasionally tagged for the use of the word “devastation” such as in the “devastation of Ship Island.”

Organization Types that Need Modification

Manual evaluation also led to the identification of several types of organization names that while useful may need some modification for greater tagging accuracy.

Commission

While the term “commission” occasionally tagged for an actual military or political commission, these phrases typically involved a commission for a specific person, such as "Willey's commission in the navy" or “Holt’s commission as captain.” Other incorrect tags include tagging of “commission house.” We are currently exploring the creation of better rules to tag only actual organizations.

District

Tagging for “districts” as organizations also proved to be problematic. While most of the terms tagged were districts, the districts referred to were places rather than formal organizations. Examples include "Wheeling District" and "Chaptico District." Often a phrase including district that tagged would be for a “district commissioner” or “district attorney.”

Office

Attempts were also made at tagging “office” in order to tag expressions such as “Office of the President.” The most frequent use of the term office in newspapers was to offices that were being referred to as a means of giving directions in an advertisement or as a placename, such as the location of a meeting. Examples includes apply at the “Manager’s office,” across the street from “Cloptin’s Office.” Another problem was that many people announced their intention to run for office in newspapers, so most of the

tags for items such as "Office of the city sergeant,", "Office of postmaster," or “office of city grain measurer,” weren’t for activities or directives of that actual office.

Courts of Law

Another organization type that was difficult to tag accurately were courts of law. The system often tagged individual judges’ courts such as “Judge Meredith’s Court” but would miss references to Henrico County Court. A related issue was that the most frequent tagging of courts was the tagging of “court-house” as a court when the reference is actually to a physical court house. Examples include “Henrico court-house” “Goochland court-house” and “Grayson court-house” all tagging as courts of law, when the references were to the physical place.

Band

This organization name type sometimes tagged expressions with the word “husband” such as in the tagging of “Band of Madame Jessie White Mario” as a band when the reference was to the “husband of Madame Jessie.” It also frequently tagged general expressions such as "band of minstrels" as organizations.

Board and Bureau

While terms such as "board of directors" and “board of managers” tagged correctly, the references were frequently to a generic use of the phrase in an article without any specific link to a particular organization or company. This led to a lot of superfluous organization names. Stop-words also need to be added into this organization type such as “aboard” and “sideboard.” A similar problem with furniture occurs in the tagging of Bureaus, since advertisements for bureaus frequently tagged as government agencies.

Church

The tagging of churches as organizations led to a number of unexpected issues. The most entertaining mistag was an article entitled “sudden death in church" tagging as the “Church of Sudden Death.” Another frequent error was the tagging of phrases such as "attends church" as a church name. In addition, phrases such "Jim's church" also tagged as organization names, when the expression simply referred to “Pastor Jim’s Church.”

Exposition

This organization name type most frequently tagged the use of “exposition” as a noun. Examples include "Exposition of Divine Truth" tagged as the "Divine Exposition" and the expression "Exposition of the President's Views" tagged as the “President’s Exposition”

Party

The tagging of political parties frequently led to false tags such as tagging expressions such as “old party feeling” as the “Old Party” or a “large party of women” as the “Women’s Large Party.” Many tags were also generated for expressions such as the “Breckinridge Party” and the “Douglas Party” and we are examining how to link these entities to their official parties.

Railroads

In general, railroad names tagged accurately, and a number of railroad names that were not tagging were subsequently identified and added to the gazetteer list. Occasionally strange expressions tagged such as “shocking railroad accident” tagging as the “Shocking Railroad.”

Schools

The tagging of schools frequently led to the tagging of expressions such as the “school of David Parke” and “school-house” and “school books” as actual organizations. We are examining heuristics to enable only higher level tagging of schools.

Union

This organization name type is challenging for a number of reasons. Many of the items the system tagged as a union were place names such as "Union Hill." It also tagged any general use of the word which including expressions such as "numerous union", "Strong union speeches" "prayers for the union" and “unconditional union” as actual organizations.

Works

This organization name type was added in to allow expressions such as the “Richmond Iron Works” to tag. It led, however, to the generation of many false tags for organizations since it tagged “fire works” as a type of works. The most frequent use of this term turns out to be for literary works so expressions such as "poetical works" "standard works" "new works" "milton's complete works" and the "works of Washington Irving" tagged as factories.

Conclusions

As this overview illustrates, there are a variety of problems that need to be resolved in the automated tagging of newspapers. Many of the issues we have identified frequently occurred due to the heterogeneous and complicated nature of nineteenth century newspaper text. Work will continue on refining these automated procedures as we come to better understand the unique problems of working with historical newspapers. The current workflow of combining automated methods with human evaluation works far better than automated systems or human tagging alone, and makes possible both extensive and complicated types of named entity identification.

Appendix #1: Revised Outcomes Assessment Logic Model

Organization Name:	University of Richmond		
Project Name:	A Test bed of Civil War Era Newspapers		
Date Created		Date Reviewed	

Program Influencers (*Key entities that help define the program or to whom the program will report results)*
Digital library community, U of Richmond Administration, Tufts University and Greg Crane, Historians and teachers, IMLS

Organizational Mission *(Organization's mission statement or key action words)*

Program Purpose	
We do what? *(Summary of key proposed services)*	*Digitizing Civil War-era newspapers from North and South using cutting edge processes to generate clear, useful images accompanied by consistent, easily searchable metadata and to transfer complementary knowledge between partner institutions*
For whom? *Target population(s)*	*The library digitization community so it can adopt new best practices and improve upon those practices.* *For scholars, students and teachers to have free access to newspapers*
For what outcome(s)? *(Benefits/changes in skills, knowledge, attitude or life condition.)*	*Other newspaper projects will adopt and improve our best practices* *We will establish a repository for 19th century newspapers and Newspapers will be used in university and high school curricula* *Knowledge (knowledge of what?) will be enhanced between project partner institutions.*

Inputs *(List items dedicated to or consumed by the program)*	Outputs *(Program products)*
New position *Equipment* *Newspapers* *Web site* *Outsource vendors* *Training consultants* *Database admin.* *% of various staff* *historian* *tufts staff* *space*	# of newspapers digitized Authority file Website DTD's Raw data sets Repository # of images metadata

Program Activities *(List key activities needed to provide or manage services.)*	Program Services *(List services to be delivered directly to participants.)*
Digitalization DCR Metadata tagging Authority work Iterative testing Reports – IMLS and more Web design Confer with others Hire for position Purchase computers Establish DTD's	Website best practices Workshop for academics and teachers Access to papers Knowledge exchange

Target Population *(List specific characteristics of primary intended participants)*
Historians, library digitization community, teachers, students

Intended Outcomes *(Changes in skill, knowledge, attitude, behavior, life condition or status)*	Indicators (Measures) *(Concrete evidence, occurrence, or characteristic that will show the desired change occurred)*
Immediate:	
Intermediate:	
Long-term:	

Outcome #1 Digital library technologies peer group will demonstrate knowledge of The Civil War era Newspaper project

Indicator(s)	Data Source (Where data will be found)	To Whom (Segment of population to which this indicator is applied)	Data Intervals (Points at which information is collected)	Target (the number , percent, variation or other measure of change)
The # and % of those who attend conference presentation that articulate 2 project purposes and know one element they can apply to their projects	Presentation evaluation	Conference presentation attendees	Immediate—at conclusion of presentation	50%
The # of sites that link to our repository	WWW	Digital Technologist with repository projects	Every 3 months	5
The # of hits on web site after an announcement of project via a listserv	Web log	Members of listserv	Week after broadcast emails	20

Outcome #2 Digital library Technologists will adopt best practices in future newspaper digitization projects

Indicator(s)	Data Source (Where data will be found)	To Whom (Segment of population to which this indicator is applied)	Data Intervals (Points at which information is collected)	Target (the number , percent, variation or other measure of change)
The # of projects that reference any of the project's best practices OR	Survey project managers; Examination of project documentation	Known newspaper digitization projects	~~May 2005~~ October 2005, then every 6 months	3
The # and % of staff from other projects who report they were influenced directly by the Civil War Newspaper project	Survey of project managers/staff	– staff involved	~~May 2005~~ October 2005, then every 6 months	5

Outcome #3 Historians know about the Civil War Newspaper Repository

Indicator(s)	Data Source (Where data will be found)	To Whom (Segment of population to which this indicator is applied)	Data Intervals (Points at which information is collected)	Target (the number , percent, variation or other measure of change)
The # and % of historians who attended the workshops who can name the purpose of the project AND	Workshop evaluation	Those who attend workshop	At end of workshop	100%
The # and % of historians who attended the workshop who revisit the project Web site	Interviews and/or survey	Those who attend workshop	~~May 2005~~ October 2005, then every 6 months	80%

Outcome #4 Historians use the Civil War Newspaper Repository

Indicator(s)	Data Source (Where data will be found)	To Whom (Segment of population to which this indicator is applied)	Data Intervals (Points at which information is collected)	Target (the number , percent, variation or other measure of change)
The # and % of historians who do at least 1 of the following: • Incorporate database in a class they teach • incorporate in their research	Interviews and/or survey	Those who attend workshop	~~June 2005~~ January 2006, then every 6 months	50%
The # and % of historians who attended the workshop who report one way in which they have used the repository in their work or research.	Interviews and/or survey	Those who attend workshop	~~June 2005~~ January 2006, then every 6 months	80%

Outcome #5 Project partner Institutions' contributors know new skills and technologies

Indicator(s)	Data Source (Where data will be found)	To Whom (Segment of population to which this indicator is applied)	Data Intervals (Points at which information is collected)	Target (the number , percent, variation or other measure of change)
The # and % of contributors at each partner institution can name 2 new ways the technology can be used or 2 new skills they learned	Interview	Grant participants at all organizations	March 2006	100%
The # and % of partner institution contributors use new skills in other projects	Interview	Grant participants at all organizations	March 2006	50%
The # or % of contributors that build on skills acquired during project	Interview	Grant participants at all organizations	March 2006	25%

Appendix #2: Best Practices Data Points

	CHNC	Utah	Univ Richmond
Technical Costs			
Scanning			
OCR/distillation			
Microfilm pulling & reshelving			
Microfilm duplication			
Newspaper (print) scanning			
Newspaper (print) rekeying			
Newspaper (print) indexing/OCR			
Preservation (print)			
Newspaper (film) scanning			
Newspaper (film) rekeying			
Newspaper (film) indexing/OCR			
Preservation (film)			
Hardware			
Hardware - initial cost			
Hardware - annual maintenance			
Cost for data storage (add'l purchased)			
Back-up tapes			
Offsite data storage for back-ups			
Software			
Software - initial license			
Software - annual maintenance			
Interface software (if separate from database software)			
Miscellaneous			
Internet connectivity (T1, router, firewall, etc.)			
Shipping (to/from providers)			
Personnel			
Project management			
System administration			
Schlepping print or microfilm			
PR/marketing			
Fund raising			
Educational activities			
Fiscal administration			
Programming			
Website/interface design			
Marketing/Promotion			
Print materials - marketing			
Print materials - educational			
Paid advertising			

Appendix #3: Utah Newspapers Project Online User Survey

Utah Digital Newspapers

1

How did you first hear about Utah Digital Newspapers?

- [ ] Internet search engine (Google, Yahoo, etc.)
- [ ] Word of mouth
- [ ] Referral or link from another website
- [ ] News media
- [ ] Other, please specify

2

How frequently do you visit?

First Time	Once a Year	Few Times a Year	Monthly	Weekly	Daily
1	2	3	4	5	6

3

What is the purpose of your visit today?

- [ ] Genealogy/family history
- [ ] General and/or historical research
- [ ] School (teachers and students)
- [ ] Just curious
- [ ] Other, please specify

4

How relevant and accurate are the results returned from your searches? Do your searches provide the information you expect?

Poor	Fair	Average	Good	Excellent
1	2	3	4	5

5

How would you rate the response time of the website?

Poor	Fair	Average	Good	Excellent
1	2	3	4	5

6

What type of internet connection do you have?

Modem/Dial-up	DSL	Cable	T1	Other high speed
1	2	3	4	5

7

What overall rating would you give the website?

Poor	Fair	Average	Good	Excellent
1	2	3	4	5

8

How does the website compare with other sources on Utah history?

Not useful	It's just okay	Among the better ones	My first choice
1	2	3	4

9

How can we get better?

10

As a result of using this website (select all that apply):

- I am more knowledgeable about U.S./world history
- I am more knowledgeable about Utah history
- I found material more quickly and efficiently
- I found/used new sources for my research
- I completed my research/project in less time
- I will return to this website soon
- I will tell others about this website
- GENEALOGISTS: I am more knowledgeable about my family's history
- STUDENTS: I completed/researched a school assignment
- TEACHERS: I compiled teaching material for my class

11

Where do you live?

In Utah	Outside Utah
1	2

12

What is your age?

Less than 20	20 - 40	40 - 60	Over 60
1	2	3	4

13

May we contact you for additional comments?

YES NO

If "Yes", please enter your first name and contact information

14

Would you be willing to donate to the program?

YES NO

If "Yes", please enter your first name and contact information