"A Testbed of Civil War-Era Newspapers"

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

Semi-annual report, April 2004-September 2004

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 have made notable progress on this project. This report addresses progress towards attaining the following planned outcomes:

1. Project Website
2. Publicly available testbed of digital source materials for further research into searching and improving the quality of OCR on newspapers
    a. Tagging of the source materials' files
    b. Factors affecting scope of project
3. Set of best practice guidelines for the acquisition of historical newspapers, with particular emphasis on the cost/benefit balance of various forms of text content acquisition and intended for widespread use by later newspaper projects
4. A sustainable infrastructure of the University of Richmond appropriate to the further pursuit of digital library activities
5. Supplementary materials to provide context for content of the newspapers
6. Scholarly conference on October 25, 2004
7. Outcomes-based assessment measures

This report also addresses several operational issues:

1. Project personnel
2. Project communication
    a. Conference calls
    b. Face-to-face meetings
3. Project budget issues

PROGRESS TOWARDS INTENDED PROJECT OUTCOMES

1. Project Website

The project's Website is at http://oncampus.richmond.edu/is/library/digital/IMLSpd.htm. The site describes the project and its goals. Interested parties can also find there a project timeline, past reports, and the metadata tagging specifications to be used by the vendor digitizing the *Richmond Dispatch* (http://oncampus.richmond.edu/is/library/digital/IMLSdata/url_vendor_specs.htm).

An image of the site's home page appears as Appendix #1 to this report.

2. Publicly available testbed of digital source materials

The goal for this aspect of the project is to create a repository of digital newspapers that is freely available to the public primarily as a resource for the civil war newspaper content, and secondly, as a resource for research about digitization al 19th century newspapers, OCR, and tagging. University of Richmond Libraries (URL) chose three Civil War Era newspapers to make up this repository:

- *The Richmond Dispatch*
- *The Public Ledger* from Philadelphia
- *The Liberator*, an abolitionist paper published in Boston by William Lloyd Garrison

The first two were selected, in part, to complement to the Virginia Center for Digital History's Civil War digital project *Valley of the Shadow* (http://valley.vcdh.virginia.edu/).

We selected the Richmond Dispatch because on the eve of the Civil War this publication had the largest circulation of any Richmond newspaper. Indeed, its circulation equaled that of all of the city's other papers combined. The reason for this is that it was a penny paper, which made it affordable even for the working-class residents. Further, it was the only daily non-partisan paper in the city and therefore featured relatively unbiased news. Finally, the University of Richmond Special Collections has a fairly complete run of the original paper that has aided in analyzing a number of important concerns. For example, we had to determine how often the identical advertisements were repeated in the newspaper over days and weeks since we did not want to digitize the same ad over and over and thereby add considerable cost to the project. By having the hard-copy of the ads at the University of Richmond it was easier to establish a sampling procedure which would maximize getting as many different ads as possible and minimize this expensive repetition.

At the end of spring 2004, the primary task at hand was selecting a vendor to image the Civil war newspapers from film and provide base level TEI tagging of the content.

The University of Richmond Library accepted bids from two digitization vendors:

Digital Divide Data (http://www.digitaldividedata.com/index.asp)
Byte Managers (http://www.bytemanagers.com/)

In selecting a vendor, we gave priority to the least cost per character, because we were especially concerned to produce the most content within our budget.

Another factor in selecting a keyboarding vendor was the previous experience our partner, The Perseus Project of Tufts University, had had with Digital Divide Data.

While Digital Divide Data (hereafter, DDD) was significantly below competitor in their cost-per-character bid, we have found subsequently that the keyboarding error rates of the files we received from DDD did not achieve the industry-standard rate of no more than 1 in 10, 000 characters. Furthermore, while DDD was directed to tag the page-images of the newspapers in the TEI (Text Encoding Initiative, http://www.tei-c.org/) XML (extensible Markup Language) application, their knowledge of XML was negligible, while their knowledge of TEI was virtually nil.

This lack of knowledge was reflected in the beginning of the second-round spec file sent to us by DDD:

```
<?xml version="1.0" encoding="UTF-8" ?>
<?xml-stylesheet type="text/xsl" href="Richmond1.xsl"?>
<!DOCTYPE rootElement PUBLIC "" "entity.dtd">
<text>
<place>RICHMOND</place>
<div1><vol>XVIII</vol><issue>105</issue><date>01/11/1860</date><page>03
</page><head>TELEGRAPHIC NEWS.</head>
```

The improvement in this markup was the inclusion of an XML declaration. However, certain aspects of the markup indicated an awareness of some XML elements, without a real understanding of them, most obviously the reference to "rootElement" after the "doctype" element. The "root element" in an XML document is the parent element of all child elements in that document. In a TEI document, for example, the root element, naturally enough, is "TEI." The appearance of "rootElement" in an XML document is analogous to a non-native speaker giving "Noun verb direct object," as an example of a sentence in English: it reflects some acquaintance with sentence structure, but a clear lack of practical application, as well as a lack of a foundational understanding of the language.

Acknowledging from the outset that DDD's knowledge of TEI and XML was limited, the primary question was whether or not DDD was capable of learning enough about XML generally, and TEI in particular, to be able to produce work that was sound in its tagging structure, and accurate in its keyboarding, to meet our needs in a timely fashion. DDD was also eager to be given the chance to prove they were capable of producing this work, and the decision was ultimately made to give half of the reels of microfilm for DDD to process, and the remaining half to Byte Managers.

The task that emerged with respect to DDD was two-fold: on the one hand, it fell to us to educate DDD on TEI and XML, focused in such a way as to communicate to them everything they needed to know to adequately tag the files, short of a full course on XML. On the other hand, and at the same time, we had to develop specifications for their markup that took absolutely all aspects of production-level data into account.

Development of TEI Markup Specifications:

While this course was less than ideal and frustrating for us in many respects, than the alternative of using experienced, for-profit keyboarding vendors that meet industry standards for double-blind keyboarding, XML specifications and keyboarding error rates, we believe that this circumstance of having to instruct DDD in the basics of TEI and XML has already proven beneficial to the University of Richmond Library (URL) in terms of the "technology transfer" aspect of the IMLS grant. One of the great paradoxes of education is that the best way to learn is to teach, and this has been true for us with respect to DDD.

The fact that DDD was essentially a *tabula rasa* with respect to TEI and XML meant we could not assume anything, and forced us to specify absolutely everything we needed in the files. With another keyboarding company, we would have been able to simply indicate we wanted files tagged in the TEI application of XML, and we would have been able to count on receiving production-level data. Because we had to specify absolutely everything we wanted in dealing with DDD, we were forced to consider and make judgments on every aspect of production-level XML data, from file-naming conventions, to line endings of text files, to the best use of the tags available in the TEI DTDs (document type definitions), as well as to consider whether or not to modify the teixlite DTD.

The primary, tangible result of our efforts to instruct DDD in exactly what we wanted in the markup of our TEI files was an html document that outlined the basic structure of TEI documents, what we required in our naming conventions, file-types, line length limitations, specifications for what should be keyboarded from an image, as well as numerous examples of the elements more likely to be employed in the newspaper mark-up, with examples of what DDD had done, and our preferred markup.

This file is currently at the University of Richmond Library site at:

http://oncampus.richmond.edu/is/library/digital/IMLSdata/url_vendor_specs.htm

It will be permanently available on the Linux server from which the IMLS project will be served as soon as the web server is enabled.

We used "Text Encoding Guidelines for Keyboarding Vendors," with permission, from the Digital Library Production Services (DLPS) unit at the University of Virginia Library, as the basis for these specifications, though it reflects industry standard practices, and the document we created reflects different choices, including the choice to remain inside the teixlite DTD, and organizes the subjects differently. The DLPS document is available here:

http://text.lib.virginia.edu/bin/cgi-dl/dlps/doco/text/kb/markup_guide/

In addition to this document, we also provided DDD a corrected file which included "working" examples, relative to other elements, of a file that parsed, as a spec. We also provided DDD with a template for the TEI Header and the front section that should apply, with minor modifications, to every file. Some highlights of the specifications document that represent decisions we had to make in accommodating the codex-oriented TEI to newspaper markup include:

- making the article the fundamental organizational unit of the newspapers
- the consequent decision to mark page- and column-breaks with the milestone tag
- the decision to use Unicode entities, and the related decision

- to encode double-quotation marks as Unicode entities, which will allow us to distinguish between quotations by actual persons in the content of the newspapers, and quotation marks which appear in the XML code, *i.e.*, in attributes
- a specification to identify the source of informally quoted articles while remaining inside the TEI lite DTD, using the citation element and the rend attribute (subject to an altered implementation with the same rationale)
- the decision *not* to keyboard all advertisements, but instead do full text markup of only 2 issues per month, while including structural markup placeholders for where individual advertisements appear, and a path to insert full-text advertisements later

Again, while working with a vendor lacking diverse experience with markup languages has proven frustrating in many respects, the decision at this point is proving to have been economically sound, on the one hand, allowing UR to gain greater full text content encoded markup, while on the other hand forcing us to consider in a deeper way exactly what we wanted from our files, and how certain tag structures could maximize the value of our data relative to the cost of mark-up.

All the above applies to markup for the primary newspaper in the project, the Richmond *Daily Dispatch*. Standards for the other two papers will be somewhat different, and will follow as a minimum the standards set recently by the NEH for its national newspaper project (available from http://www.neh.gov/grants/guidelines/ndnp.html).

The other major instrument to arise from the articulation of markup specifications, as well as negotiations with DDD, was a revised statement of work (SOW) which has been signed by both parties. The SOW binds DDD to delivery of TEI-encoded XML files every other week, delivered by FTP, and 8-bit bi-tonal tiff images delivered on DVD by the same date. The schedule calls for the delivery of 60 issues per month by November, 2004, resulting in "a minimum 700 issues by August 2005." We anticipate that DDD will provide more files than the 700, as production in the last month was double than what they had committed to in the SOW. While the *Richmond Dispatch* is being treated by DDD, the PA Public Ledger and the Liberator has been sent to another digitization company (Bytemanagers) to be imaged and OCR'd. URL is scheduled to have this data from the Bytemanagers in 2 segments – the first by January 15th 2005; the second will be received and paid for by July 1st 2005. To date we have received over 83 issues. With this received data in place, URL is on track to be able to ready files for higher-level tagging by Perseus-Tufts, and complete final editing of all files, all papers by June, 2006.

Scope of Project:

Revision of the Statement of Work and discovery that each page of the *Richmond Dispatch* hold more characters than anticipated have resulted in a higher cost than anticipated to digitize this one paper. The reasons cited above for selecting the *Richmond Dispatch* are also compelling reasons to make this newspaper the centerpiece of the project. In order to treat its contents in sufficient depth to carry out the research agenda of this project and in order to provide a singular resource for the study of the Civil War in Richmond, plans to digitize significant runs of the *Liberator* and the *Public Ledger* have been reduced. The scope of the project is now as follows:

Newspaper	Dates	Format
The Daily Dispatch (Richmond, VA)	November, 1860-May, 1865	article-level metatdata, full-text markup in TEI XML from page images
The Liberator (Boston, MA)	January 1, 1861-May, 1865	2 years clean OCR and page images; page-level metadata 2 years "dirty" OCR and page images; page-level metadata
The Pennsylvania Public Ledger	January 1, 1861-May, 1865	4 years "dirty" OCR and page images; page-level metadata

The major responsibility of URL is the imaging and OCR/keyboarding of the three Civil War-era newspapers. Of these newspapers, the focus will be on the full-text markup in TEI XML of *The Richmond Dispatch*. Scope of imaging and treatment of the other 2 papers is outlined above. This variety of newspaper data will provide a rich research opportunity in regard to examining accuracy rates of tagging, searching and retrieval. URL will be responsible for the initial quality-assurance of the files. The major responsibility of Perseus-Tufts will be the above-mentioned higher-level tagging of *The Daily Dispatch*, and the markup of several supplementary sources, both textual and geospatial, to create tools that can leverage the sources in unique ways through web delivery. Through the creation of authority files for personal and place-names, the markup of Civil War era, Richmond area directories, and markup of Richmond area maps, these data-sets will be able to interoperate.

The other major responsibility of URL will be to host the newspaper text files and page-images on a dedicated server. The implementation of software to host and track and deliver digital library objects constitutes the other major aspect of the "technology transfer" goal of the grant.

WORKFLOWS AND PRODUCTION

At the beginning of September, two undergraduate students were hired as "Metadata Editors" to assist in the quality-assurance processing of files received from DDD. The student workers have been trained by the Digital Resources Librarian, who supervises them, and are now beginning the first pass over the files received from DDD, which focuses on parsing, spell-checking, and the integrity of the tagging structure.

On September 15, 2004, project members from UR met with project members from Perseus-Tufts in Boston, primarily to discuss metadata standards and work-flow.

It was agreed that UR would be responsible for the first pass of quality assurance for the files received from DDD. Perseus-Tufts would then be responsible for the automated implementation of higher-level tagging to capture information from the files including:

- abbreviations
- personal names
- place names
- addresses
- military units
- dates (in ISO format)

Perseus-Tufts estimates that the automated tagging using their transducers (built from a series of Perl scripts) can capture 90-95% of the targeted information; URL will then correct these files, concentrating on tagging the remaining 5-10% targeted information.

The transducers employed by Perseus-Tufts represent several years' work, and are maintained primarily due to work-flow. However, they have identified new software capable of learning to identify recurring patterns of text, called GATE (General Architecture for Text Engineering) available at: http://www.gate.ac.uk/, produced by the University of Sheffield.

As part of the "technology transfer" rubric of the IMLS grant, a selection of the files will be treated by UR using GATE to implement the higher-level tagging, with assistance from Perseus-Tufts, which will contribute to an infrastructure and work-flow for future digital library projects.

A CVS will be set up to then allow members of both teams to "check out" and process files as needed.

3. Set of best practice guidelines for the acquisition of historical newspapers

G. CRANE CONTRIBUTION HERE

4. A sustainable infrastructure of the University of Richmond appropriate to the further pursuit of digital library activities

One of the outcomes of this grant is to create a "sustainable infrastructure at the University of Richmond appropriate to further pursuit of digital library activities." Part of that process involves exploring "digital repositories of as internal library systems and instruments of interoperability. Tufts and Virginia are both implementing the Mellon funded Fedora digital repository, while Richmond will select a repository strategy of its own. By including the repository question from the beginning, this project dramatizes one promising strategy of digital preservation (e.g., the library maintains complex digital objects over time)." In order to chose a digital repository architecture that would be both suitable for this grant and serve as the foundation for future digital projects, the University of Richmond team embarked on a selection process that involved the review of currently available open source digital repository software.

The first step in the selection process was educating the team on the institutional repositories in general and on specific open source digital repository software. The team gathered information on digital libraries and institutional repositories from the web, journal articles and books. One of the most useful web sites was at Budapest Open Access Initiatives, http://www.soros.org/openaccess/software/, because it compared several open source products.

The team also developed a list of features that would be essential for building the UR repository. The software should:

- Function as both a digital library and an institutional repository
- Be capable of handling a wide variety of digital objects
- Meet the OAI standards for harvesting digital information
- Operate on a UNIX-based system, preferably Linux or Solaris
- Work with MySQL or Oracle
- Have strong indications of support for future development
- Provide a web-based front end for searching

Because the team also determined that preparation and ingestion of materials was to be mediated by librarians and other IS professionals, it was not essential initially for faculty or other members of the UR community to directly submit digital objects into the system. Comparisons were made between the list of desired features and the actual capabilities of several open source digital repositories.

In investigating the available repositories, one thing was very clear: digital repositories are in their infancy. None of the open source software available met all the criteria that the team had specified. The two products that came closest were Fedora (from the University of Virginia Library and Cornell University, available from http://www.fedora.info/) and DSpace (from the Massachusetts Institute of Technology Libraries and and Hewlett-Packard, available from http://www.dspace.org/).

At this point, it became important to communicate with those who had some experience with one and/or both of these products. Selected members of the UR team visited the Digital Knowledge Center at Johns Hopkins University in March, 2004. Sayeed Choudury and his digital team were in the process of reviewing both DSpace and Fedora. After discussing both digital repositories with the Digital Knowledge Center staff, the UR team concluded that DSpace was the more developed package, but Fedora came closer to meeting the goals of the IMLS project.

While the two architectures are suitable for creating institutional repositories, they are designed to serve different purposes. DSpace, as indicated on their web site, is software that "captures, stores, indexes, preserves, and redistributes the intellectual output of a university's research faculty in digital formats." DSpace seems primarily aimed at preserving intermediate and especially "unfinished" material, such as notes, draft material, data-sets and so on: material that might otherwise be lost simply because digital material must be maintained in such a way that analogue materials do not require.

Fedora, by contrast, represents an evolution is what has been called in recent years, digital library architecture. According to its web site, "Fedora is a general-purpose digital object repository system that can be used in whole or part to support a variety of use cases including: institutional repositories, digital libraries, content management, digital asset management, scholarly publishing, and digital preservation." The broader scope of the Fedora architecture appeared to meet both the immediate needs of the IMLS project and provide a foundation for future projects. Fedora was also better suited to handle the overall goals of the UR digital repository, which is to organize, track and make available "finished" output, such as published books and other media. Finally, the nature of the IMLS project, consisting of definite deliverables and unique, discrete digital objects, much more conforms to the digital library model of information organization than that of DSpace. Fedora appeared to be the better choice.

The remaining question was the ability of UR to actually implement Fedora, which has a much more complex architecture. Thornton Staples, Director of Digital Library Research and Development at the University of Virginia and a developer of Fedora, agreed to meet with the UR team and provide a detailed explanation of Fedora. This visit confirmed that Fedora was more flexible and powerful in terms of its ability to organize and maintain the types of digital library objects planned for the UR repository. It was also clear that more work would be needed on the part of the UR team to implement Fedora. The fact that Tufts was also planning to use Fedora was seen as further opportunity to share and develop technical skills and expertise. The decision was made to select Fedora, with the understanding that it would require more investment of time and resources but would prove to be the better product in the long run for building an institutional repository that could handle a wide variety of complex digital objects.

To give others involved in the IMLS project an opportunity to better understand Fedora, the University of Richmond arranged a presentation by Thornton Staples on July 16, 2004. In order to education others in the community about Fedora, invitations were sent out to the major library institutions in the Richmond area and members of the Virginia Library Association. In addition, the presentation was webcast live.

Fedora is widely known in the digital library community, but has been implemented, for the most part, only on an experimental basis by most institutions exploring possible use of Fedora. The ability to have one of the developers of Fedora come and speak significantly improved the team's grasp of what Fedora is and is not capable of, and how it could be most successfully deployed. One theme underscored by Mr. Staples is that Fedora is digital library architecture, not a "solution." That is, it requires significant software and programming on top of its architecture to effectively deliver digital library objects. He also clarified Fedora's search capabilities. Fedora does advertise this feature, but inquiries through the developer site and at the presentation of July 16th confirmed that these capabilities are extremely limited, and insufficient to meet the desired search functions for a digital library.

Those involved in the IMLS project realize it will be important to determine a way to provide additional searching capabilities for the digital repository. This will mean exploring options such as creating an XML-based database outside of Fedora and using open source search engines, such as Lucene. It is also essential for the team to keep abreast of current developments.

For example, two recent postings on the fedora-users listserv announced projects that could prove useful. One announcement was from Andrew Treolar in Australia. "The Australian Research Repositories Online to the World (ARROW) project has selected FEDORA as its underlying repository software. ARROW is partnering with VTLS to develop software that will work with FEDORA and which will be released as Open-Source." Further clarification indicated that only parts of VTLS contribution would be open-source. VTLS "will extend the functionality of FEDORA either by contributing back to the core FEDORA code or by writing a series of ARROW-commissioned modules. This will all be open-sourced using the same license as the FEDORA code." The second announcement was from Eric Jannson, who supervised a group of software engineering students from members of the Associated Colleges of the South. This was a summer ACS technology project, and the group released a "beta version of a Java-based Fedora client" called ELATED on August 23, 2004. Jannson states that "ELATED is a general-purpose application for managing digital media files that uses Fedora as its back-end. Our goal was to create a software product that provides a web-based front-end to Fedora that would make it possible for institutions and organizations with few development resources to begin using Fedora." The UR team will have to test a variety of products to find the best solution for accessing the digital repository.

In selecting Fedora, the UR team believes it has both chosen the best option for the IMLS project and for the UR institutional repository of the future. The process has also involved a significant amount of "transfer of technology", another goal of the IMLS grant. This "transfer of technology" has come not only from Tufts, but from many other members of the digital community.

Fedora and the Richmond IMLS Server:

Following the decision to use Fedora, a server was chosen and ordered: a Linux server running Red Hat, with a 64-bit capable Opteron processor. Listed below are the server specifications as ordered:

ProLiant DL580 G2 Intel® Xeon™ Processor MP at 2.70GHz/2MB (2P Model)
Two Intel® Xeon™ Processors MP 2.70GHz/2MB
Intel® Xeon™ Processor MP 2.70GHz/2MB - Option Kit

8GB Base Memory 4x1024,4x1024
Standard One Ultra3 SCSI Drive Cage (2x2 Duplex std or 4x1 Simplex)
Integrated Smart Array 5i Plus Controller (Dual Channel, Ultra3)
Standard Battery Backed Write Cache Enabler (up to 64MB Write Cache)
RAID 0 setting (Requires minimum of 2 matching drives)
72.8 GB Pluggable Ultra320 SCSI 15,000 rpm Universal Hard Drive (1")
72.8 GB Pluggable Ultra320 SCSI 15,000 rpm Universal Hard Drive (1")
72.8 GB Pluggable Ultra320 SCSI 15,000 rpm Universal Hard Drive (1")
72.8 GB Pluggable Ultra320 SCSI 15,000 rpm Universal Hard Drive (1")
1.44MB Floppy Disk Drive
Slim Line CD-RW/DVD-ROM 24X Combo Drive
HP NC7170 Dual Port PCI-X 1000T Gigabit Server Adapter
Two (2) 800W Hot Plug Redundant Power Supplies
Redundant Hot Plug Fans

The server arrived in late August and was installed in September. The next step is the installation of Oracle. Oracle was selected as the back-end relational database for Fedora because the University of Richmond already has a campus-wide Oracle license and in-house expertise. The team has been experimenting with the installation and configuration of Fedora. Expected date for completion of the Fedora installation is early October.

5. Supplementary materials to provide context for content of the newspapers

G. CRANE CONTRIBUTION HERE

6. Scholarly conference on October 25, 2004

The purpose of the conference is to tap into the expertise of a wide variety of scholars, librarians, archivists, digital-specialists, and experts in Civil War studies in order to share their ideas with us about how we should structure our website. Because many of these individuals and institutions—such as the Library of Virginia-- already have devoted considerable resources to digitizing newspapers, we believe they can give invaluable advise. Others, such as the Digital Library and Archives at Virginia Tech, have already established model websites for their Civil War resources. In total, not including the University of Richmond and Tufts University (the sponsor institutions), fourteen different institutions (see list below) will be represented. A number of these institutions will benefit considerably by the availability of the website—the Museum of the Confederacy, the Virginia Historical Society, the Richmond National Battlefield Park, and the Tredegar National Civil War Center. Hence, we are particularly seeking input from these institutions at this point so we can be especially sensitive to their needs. Further, in addition to collegiate-level Civil War scholars (seven will be present), we are inviting two different high school teachers to consider how the website might be used in the secondary educational setting.

The agenda of the conference is divided into two parts. The first part will consist of three short talks: a description of the present status of the grant project, an overview of other Civil War and newspaper-centered digital projects, and a presentation of availability of sources dealing with Richmond's African Americans during the Civil War. The second part of the conference will be centered on discussion groups focusing on three topics: the content of the website, functionality (how the website might be organized), and other available resources (which might be provided by some of the institutions represented). The one-day conference will conclude with an evaluation that will be filled out by participants.

The 14 institutions sending participants include:

1. Library of Virginia
2. Virginia Historical Society
3. Valentine Museum
4. Museum of the Confederacy
5. Richmond National Battlefield Park
6. Tredegar National Civil War Center
7. Historical Highway Marker Program of the Virginia Department of Historical Resources

8. University of Virginia
9. Virginia State University
10. Mary Washington University
11. Virginia Tech
12. Trinity University of Texas
13. Governor's School of Richmond
14. St. Anne's Belfield School of Charlottesville

7. Outcomes-based assessment measures

In keeping with best practice for IMLS-funded projects, the project team developed a .set of intended outcomes and devised appropriate measures for those outcomes. These appear as Appendix #3 to this report. The # and % targets still need to be defined for some of the measures.

OPERATIONAL ISSUES

1. Project personnel

On June 1, 2004, Dr. Andrew Rouner joined the Richmond team, filling the grant-funded position. Before joining this project Dr. Rouner served as Project Manager for the Center on Religion and Democracy at the University of Virginia's Electronic Text Center. In August he completed all requirements for his doctorate in religion.

2. Project communication

Each month the University of Richmond and Tufts teams hold a conference call to review project progress and discus issues requiring attention.

Members of the Richmond team visited Tufts on September 15, 2004, to discuss metadata tagging and production issues. Appendix #2 to this report is an in-depth report on this meeting .

Another face-to-face meeting is scheduled at the University of Richmond on October 26, 2004. The purpose of that meeting is to review the project's first year.

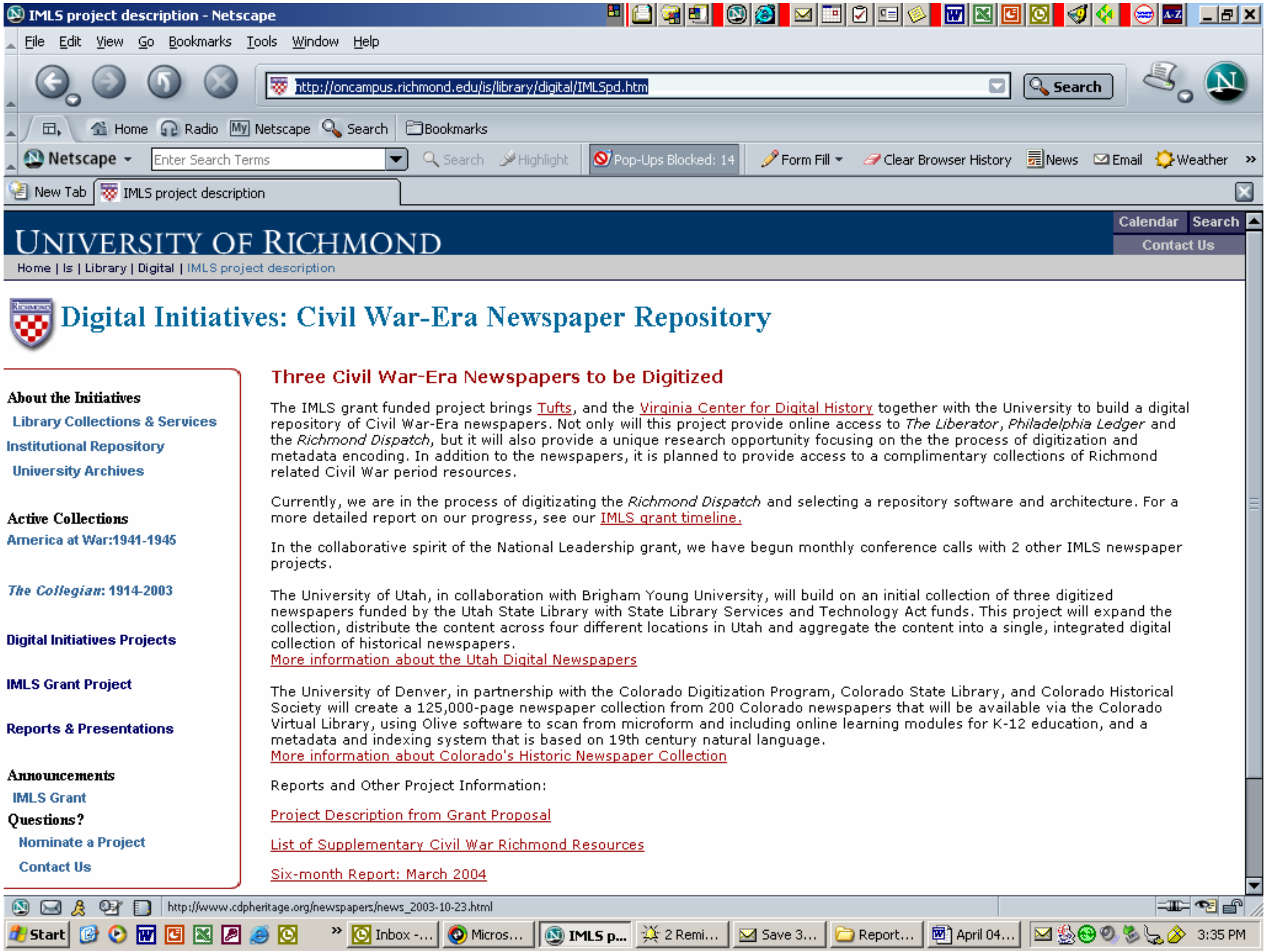
3. Project budget issues

As soon as the end-of-month budget report for September become available, Tina Snellings of the Foundation, Corporate and Government Relations office at the University of Richmond and Jim Rettig will revise the University of Richmond project budget to cover three fiscal "years:"

- October 1, 2003-September 30, 2004
- October 1, 2004-September 30, 2005
- October 1, 2005-May 31, 2006

This revised budget will become a model for Tufts so it can revise its budget for the same periods. This budget revision is necessary because the grant-funded position was not filled until June 1, 2004. Once the budgets at both institutions are revised, the PI will submit a request for an extension of the project through May 31, 2006.

Appendix #1: Image of Project Web Site Homepage:


IMLS project description - Netscape
File Edit View Go Bookmarks Tools Window Help
http://oncampus.richmond.edu/is/library/digital/IMLSpd.htm
Search
Home Radio Netscape Search Bookmarks
Netscape Enter Search Terms Search Highlight Pop-Ups Blocked: 14 Form Fill Clear Browser History News Email Weather
New Tab IMLS project description
UNIVERSITY OF RICHMOND
Calendar Search
Contact Us
Home | Is | Library | Digital | IMLS project description
About the Initiatives
Library Collections & Services
Institutional Repository
University Archives
Active Collections
America at War:1941-1945
The Collegian: 1914-2003
Digital Initiatives Projects
IMLS Grant Project
Reports & Presentations
Announcements
IMLS Grant
Questions?
Nominate a Project
Contact Us
http://www.cdpheritage.org/newspapers/news_2003-10-23.html
Start Inbox -... Micros... IMLS p... 2 Remi... Save 3... Report... April 04... 3:35 PM

Digital Initiatives: Civil War-Era Newspaper Repository

Three Civil War-Era Newspapers to be Digitized

The IMLS grant funded project brings Tufts, and the Virginia Center for Digital History together with the University to build a digital repository of Civil War-Era newspapers. Not only will this project provide online access to *The Liberator*, *Philadelphia Ledger* and the *Richmond Dispatch*, but it will also provide a unique research opportunity focusing on the the process of digitization and metadata encoding. In addition to the newspapers, it is planned to provide access to a complimentary collections of Richmond related Civil War period resources.

Currently, we are in the process of digitizating the *Richmond Dispatch* and selecting a repository software and architecture. For a more detailed report on our progress, see our IMLS grant timeline.

In the collaborative spirit of the National Leadership grant, we have begun monthly conference calls with 2 other IMLS newspaper projects.

The University of Utah, in collaboration with Brigham Young University, will build on an initial collection of three digitized newspapers funded by the Utah State Library with State Library Services and Technology Act funds. This project will expand the collection, distribute the content across four different locations in Utah and aggregate the content into a single, integrated digital collection of historical newspapers.
More information about the Utah Digital Newspapers

The University of Denver, in partnership with the Colorado Digitization Program, Colorado State Library, and Colorado Historical Society will create a 125,000-page newspaper collection from 200 Colorado newspapers that will be available via the Colorado Virtual Library, using Olive software to scan from microform and including online learning modules for K-12 education, and a metadata and indexing system that is based on 19th century natural language.
More information about Colorado's Historic Newspaper Collection

Reports and Other Project Information:

Project Description from Grant Proposal

List of Supplementary Civil War Richmond Resources

Six-month Report: March 2004

Appendix #2: Summary of the Tufts—Richmond Meeting:

Boston, September 15, 2004

Agenda:

The Perseus-Tufts team set the agenda for the meeting, which took place on Wednesday, September 15, 2004.

The meeting began with discussion of the white paper by Perseus staff members Alison Jones and Gwynne Langley, which surveyed current newspaper digitization projects.

This was followed by a presentation by Kim Tryka of VCDH on their projects which have delivered analogous content, demonstrating the possibilities of inter-operable content and linking.

These presentations were interspersed with discussions about the possibilities and limitations of mapping and/or transforming metadata, the balance of markup cost against complete data (i.e., limitation of ad data in the Richmond *Daily Dispatch*), and the use of local DTDs (document type definitions) and metadata schemes.

The remainder of the day was primarily filled with demonstrations of content, delivery models or digital library technology being developed for the Civil War project, or which could be conceptually applicable to the Civil War project.

Content:

URL will be responsible for the primary content (newspaper images and xml files, see below) in their primary form; Perseus-Tufts will be responsible for higher-level tagging, and for supplementary materials. Supplementary materials consist primarily of:

- Authority files
- City directories (names and addresses)
- Maps and GIS data

Perseus-Tufts will be scanning Civil War-era maps of Richmond, transposing current topographical maps over the historical maps, and by tagging address information on them, create user tools which allow for the interaction between geographical and other data-sets, for example the ability to search for individuals by name from a city directory, and locate his or her address on a map.

The interim director of the Tufts Digital Collections and Archives (DCA), Anne Sauer, demonstrated the Boston Streets project, which gives an idea of what Tufts wants to do with Richmond data.

Tagging:

URL will only be immediately responsible for the receiving and correcting the xml files from DDD (and other vendors for the *Liberator* and *Ledger*), general QA.

Tufts will use a series of Perl scripts run as a transducer for the value-added tagging of the files (once they are initially spell-checked and proofed by URL), to include tags as:

Tufts will send a complete listing of the information they want, are able to, and will tag. Their transducer will capture 90%-95% of text that it aims to capture.

While Tufts invested several years in developing the Perl script-based transducer for automated tagging, they look to the future in a Java-based, open-source application developed by Sheffield University, called GATE (General Architecture for Text Engineering) available at: http://www.gate.ac.uk/.

GATE describes itself as analogous to a software development environment—it is not an out-of-the-box solution, and appears to have several different modules that work together in a complex fashion: it has a significant learning curve.

What GATE promises is the ability to recognize grammatical parts of speech, as well as typically recurring texts patterns which can be identified by the program, for example, as an address, a personal or place-name, an institutional name, or other standardized information. GATE could conceivably be used to automate tagging of electronic text.

As helpful as it will be to have Tufts take responsibility for most of the higher-level tagging, for the purposes of the "technology transfer" element of the grant, as well as for development of future digital projects at URL, it will be important for URL to take responsibility for the higher-tagging of a small selection of texts, and to learn to use GATE for this purpose.

Delivery:

Greg Crane wants to set up a CVS (concurrent versioning system) that will allow project-members from Tufts and URL to "check out" files for processing. We have not established where the CVS will live. Our preference is that the CVS live on the IMLS server at Richmond.

Procedures for how and on what schedule data will be shared through a CVS will be determined at the upcoming meeting in October, at which point we may also begin discussion of how data will be shared through FEDORA.

FEDORA:

Tufts maintains a separate database outside FEDORA to assign PID numbers (this may have been due to the fact that Tufts began ingestion into FEDORA with an earlier version which did not automatically assign PID numbers).

Tufts does not have its files "live" in FEDORA, but instead points FEDORA to the URL (directory) of the relevant files.

We will need to create an object model for the digital object of a given newspaper, including the xml file and related images for a given issue, and the issue's membership within the larger digital object called The Richmond Daily Dispatch. We will be in contact with Erin Stahlberg of UVa, who was responsible for the object model of The Cavalier Daily, to discuss newspaper digital object modeling.

Tufts DCA originally tried to contain original tiff files within FEDORA, and dynamically generate jpgs and gifs from the tiffs, but found this was too processor-intensive; they now create static jpgs and gifs that are identified as associated with a given digital object.

Tufts DCA has not found an automated way within FEDORA to generate the Dublin Core metadata, or the TEI Header file (in the FEDORA, "METS-like" format) from the original TEI file. However, this process can be automated to a significant extent outside FEDORA. We will need to create scripts, if possible, to generate these files, and implement this as part of the work-flow. Tufts DCA staff indicated there was a way to batch-ingest xml files, but we did not discuss their success with the batch-ingest tool.

Indexing:

Tufts Digital Collections and Archives currently uses Oracle to index files outside FEDORA.

They acknowledge problems of translating native xml into relational database format, and are looking at alternatives, especially eXist (as a "native xml database").

Kim Tryka at the University of Virginia is using eXist on a VCDH project currently, which works well.

Perseus is using Lucene on its new site, but it is not scalable.

There is strong consensus that eXist is the best route for the short-term; at the same time, it is not a long-term solution because of its inability to give KWIC (keyword in context) results and to generate a link to the relevant section; only XPAT has this feature.

Appendix #3: Outcomes 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, 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, 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	June 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, 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, 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%