"A Testbed of Civil War-Era Newspapers"

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

Semi-annual report, October 2004-March 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 have made notable progress on this project. This report addresses the following highlights of the past half year:

1. October 2004 conference
2. No-cost extension request
3. Minor revision to outcomes assessment Logic Model
4. Scope of work on the Philadelphia *Public Ledger* and the *Liberator*
5. Progress on developing best practice guidelines, including cost/benefit
6. Progress on implementation of FEDORA at Richmond
7. Development of TEI mark-up specifications
8. Workflows and production
9. Summary of two white papers
10. Current tagging work at the Perseus Project
11. Authority list creation and refinement

1. October 2004 Conference at the University of Richmond

On October 25, 2004, we held a conference for Civil War historians on the University of Richmond campus. The individuals selected were representative of the people we believe will be the primary users of our site, so we wanted to a) inform them of the project, and b) discuss with them aspects of overall site usability. We held this conference very early in the process because we wanted to be sure that our users' needs were designed into the project from the very beginning. Thirty-nine individuals were invited to attend the conference and 37 people were able to attend. The attendees included college and high school teachers, university and state librarians, museum curators, state historical site directors, and independent scholars. The morning sessions were devoted to explaining this specific project to the attendees, as well as placing this project in a broader national context. The afternoon was spent in three discussion groups where the attendees discussed what types of information that will be on the site and the functionality that they would like the site to have.

A survey was conducted at the end of the conference to formally collect their feedback on the project and our process. All attendees (n = 26) believed that newspapers are valuable teaching and research tools and that online repositories of primary historical resources are and will continue to be valuable research and teaching tools. Most (79%) personally use online resources frequently in their work, with 17% using them occasionally, and 4% seldom using online resources. The majority of attendees believe that the content (88%) and the functionality (83%) currently proposed for the site will be useful for their research and teaching. All of the attendees wanted to continue to be involved in the project, with 26% having resources to contribute, and 17% having resources, but no staff time to offer. Most (77%) are confident or highly confident that the online repository being developed will be a valuable research and teaching tool, with 15% moderately confident, and 8% with low confidence.

We will be contacting the conferences attendees again in May, 2005, to provide them with a progress report on the project. In addition, we will ask them to answer a very brief survey to determine if attendance at the conference has altered their use of online resources or their opinion of the validity of online resources. As the project progresses, this group will continue to serve as an important conduit for us into the user community.

The conference agenda and the list of conference participants are appended to this report.

2. No-cost extension request

In March 2005 the University of Richmond submitted to IMLS a no-cost extension request until September 30, 2006.

The principal reason for requesting this extension is that it was not possible for the University of Richmond to fill the grant-funded position immediately upon the project start date—i.e., the date that IMLS funds became available to the University of Richmond. That date was October 1, 2003, the first day of the federal government's 2003-04 fiscal year. In the months following October 1, 2003, we conducted a search to fill that position and were pleased to hire Dr. Andrew Rouner who has brought to the project vital knowledge of meta data tagging schema and conventions as well as significant experience in their application. Because of his commitment to another project on which he was working, Dr. Rouner was not able to start his employment at the University of Richmond until June 1, 2004.

Another reason for the requested extension is that Digital Divide Data, the vendor selected to digitize and tag the Richmond Dispatch for our project, was not able to meet our standards in early test work. Rectifying this problem required several months of negotiations, clarification of expectations, and new tests of Digital Divide's ability to tag files thoroughly enough to meet the project's needs. It proved propitious that these problems and Dr. Rouner's start date nearly coincided. We received the first digitized files that met our standards shortly after Dr. Rouner joined the project team. He was, therefore, able to give them his immediate attention.

3. Minor revision to outcomes assessment Logic Model

The reasons cited for requesting a no-cost extension for the project are also the reasons that minor revisions have been necessary in the timeline for the project's outcomes assessment processes. The revised logic model is appended o this report.

4. Scope of work on the Philadelphia *Public Ledger* and the *Liberator*

Currently the Philadelphia *Public Ledger* and the *Liberator* are being imaged as 8 bit 400 dpi greyscale TIFF images. Flat text files are being generated for the 1,000 pages of the *Liberator*, via a combination of OCR and double keystroke, as microfilm image quality dictates. When we complete the text files for the *Liberator*, we will then evaluate how much, and what issues we will provide text files for the *Public Ledger*. For both papers we will provide page level metadata according to NEH guidelines outlined in their Newspaper Digitization project. The planned completion date for data output is September 2005.

5. Progress on developing best practice guidelines, including cost/benefit

Currently the University of Richmond has developed a specification document to guide organizations in communication with digitization vendors. This is freely available at http://oncampus.richmond.edu/academics/library/digital/IMLSdata/url_vendor_specs.htm

As we gather cost data this summer from both of our digitization vendors, we will begin analyzing cost benefit strategies in regards to digitizing 19th-century papers. The goal for the University of Richmond is to have this analysis complete and reported in our final IMLS report.

6. Progress on implementation of FEDORA at Richmond

Work continues on implementing FEDORA at the University of Richmond:

- IMLS Server:
  - Oracle installed
  - Fedora installed and configured
  - SVN installed

- Development Servers:

  Two machines have been used as development servers:

  - A dual-processor G5 (Mac OS X)
  - A single-processor X86 (Linux Red Hat)

  On both machines, the following have been installed:

  - MySQL 3.x / 4.x
  - PhPMyAdmin
  - Fedora 1.2

Fedora has been configured on both machines to ingest the demo objects, and experimentation on Fedora object modeling has been initiated; further development will proceed when the digital object models for objects containing TEI XML files, and/or specifically files derived from newspapers, are shared by the University of Virginia Library. Andrew Rouner is scheduled to attend the Fedora Users Conference, hosted by Rutgers University, in May 2005, and we expect this conference will be especially useful as a resource for the modeling of Fedora digital objects, and for delivering content through Fedora, possibly through software developed to work on top of Fedora, *i.e*. eLated.

Toward that end, we have received a promise of assistance from the developers of eLated, in a conference call we had with Eric Jansson recently on March 16th. This software was developed by the ACS Technology Center. Its associate director, Eric Jansson contacted UR to inquire about our use of Fedora in our digital library infrastructure. Since we are one of the few, perhaps the only, institutions of this size or smaller implementing Fedora, we are in a unique position to provide leadership for institutions of a similar size in integrating Fedora in our digital library architecture, and ACS is particularly interested the adoption of this technology by similar institutions, and is eager for feedback on our successes and obstacles to using Fedora, and may be in a position to offer add-on open source software solutions to address the obstacles, and to promote the successes as solutions where we achieve them.

The earlier installation on the G5 development server included installation of the open source Image Magick, which provides the ability to batch-process images. Fedora's SaxonServlet was used independently from its other functions to provide an early model of the delivery of images and text via the web, where images processed though Image Magick were delivered as thumbnails in the XML documents, and allowed users access from them to the larger jpg image. The XML files were processed through the SaxonServlet via an XSL style sheet, which will be used as a behavior in the construction of TEI-based Fedora objects. Thus functionality was demonstrated at the October 25, 2004 Conference on Civil War

In contrast to the initial installation on the Mac OS X development server, the more recent installations on the Linux server were performed with all the personnel of the Library Systems area (as well as Andrew Rouner) and all have been participating in Linux training in the process. Rick Neal has been designated sysadmin of the development server. With this installation/configuration, the process is being documented, both for purposes of simply being able to quickly reproduce these installations (both for UR, and to share with other institutions) but also as an exercise in best practices (i.e., the use of PhpMyAdmin as useful tool for viewing Fedora database tables through a GUI, Image Magick for processing images, etc.).

7. Development of TEI mark-up specifications

While an initial "specifications for keyboard vendors" document was developed almost immediately upon entering into a contract with Digital Divide Data (referenced in the previous report), that document has undergone several revisions as needed since. Some standards, of course, do not apply to keyboarding vendors, but to post-processing of files by UR and the Persseus Project. We have had three opportunities to meet face-to-face with members of the Perseus Project—most recently Andrew Rouner's visit to Boston

from March 8-11—and through those meetings, have established a number of specifications for the TEI XML files beyond the parameters of keyboard vendor-provided markup. Much of this will be accomplished through the automated markup transducers of the Perseus Project, but some will be hand-corrected. A separate document is in development for these specifications. Amongst the additions to this area of TEI specifications are the addition of the **`bibl`** element to all newspapers cited within the Richmond *Daily Dispatch*, and the use of OCLC numbers in the **`ref`** attribute to the bibl element. We have also discovered that the *Daily Dispatch* used a form of metadata to control publication of the ads it received, specifying in a shorthand when an ad should begin running and for how long, which we will tag with the syntax of **`<term type="printrun"></term>`**. For geographic locations we will be using Getty Thesaurus numbers as unique identifiers, and similarly will use LOC name authority file numbers for the top 1000 most frequently-occurring names in the XML files as unique identifiers for content identified as persons.

Another significant development in this area was the hand-tagging of a single issue of the Richmond Daily Dispatch, primarily during November, 2004, not only with the primary level of TEI tagging (as iterated in the "specifications for keyboard vendors" document) but also including all the tagging of named entities, including persons, geographical locations, military units, railroads, ships, organizations, and more, which will be automatically tagged via the transducers developed at the Perseus Project. This file was used as the basis for Persesus' transducers to "learn" what to tag as named entities. It also serves to show the prohibitive cost of attempting to hand-tag such information, and serves a research aim of the grant in this respect.

8. Workflows and production

Local production: While initial purchases of XML Spy software for purposes of XML editing and other higher functions for some staff members, both UR and Perseus have since standardized on oXygen as a relatively low-cost XML editor for student metadata editors and for additional work station installations.

As an upgrade from the concept of the CVS for workflow, SVN has been installed on the IMLS server. Perseus configured SVN on their servers so that correctors can open files from a server and edit them directly through oXygen on the server. We have recently installed SVN on the Richmond IMLS server, and this will facilitate not only the local workflow with our student correctors, but also with Perseus Project team members.

Having examined the time it would take in person-hours to give very thorough examinations of both content and tag structure to each file received from keyboarding vendors, we have determined that this would be prohibitive, both in terms of the time this would take, and in terms of cost. We have revised our strategy to focus individual examination of files on correction of content, on the one hand, in order to process them faster, and to improve batch-correction of files on the other, through the implementation of PERL scripts and other utilities, which can correct mistakes we have found typical in files received from keyboarding vendors. We are now close to a 3 hour turnover per file in correcting XML files received from vendors.

We have also improved tracking of files received from vendors, and in implementing simple tools (i.e., server-based spreadsheets) to track information and share it readily. These spreadsheets furthermore do not simply reflect data entry, but are themselves quality-control tools. The spreadsheets are generated through the use of search utilities such as GREP, to generate information about file names, issue numbers and dates, issue day-of-week, numbers of pages per issue and so on. With this information arrayed, errors in filename duplication, missing issues and other quality control issues can be readily identified and addressed. A similar spreadsheet separately records files in the order in which they were received from the keyboarding vendor, and keep a continuous count of costs-to-date, while linking to vendor invoices.

Over the summer, we will lose most or all of our student workers, while the work load for other areas of the library lightens somewhat. In light of this, we will train a few FTEs on the library staff in correction of the TEI XML files. In order to make this a genuine training opportunity, training will take the form of an introductory short course on XML and related technologies, engaging issues beyond the absolute needs of simple content-correction of files, but which will bring greater knowledge of this crucial component of

digital libraries to more staff members. Andrew Rouner is currently developing a syllabus for this short course, which will also be offered as a continuing education course at Richmond in fall 2005.

Keyboarding vendors: We will be revising the contract (SOW) yet again in the beginning of April in terms of issues processed by DDD, to account for the fact that so few issues are now being keyboarded in full (most ads are being skipped in most issues) and for the fact that, as the war progressed, fewer and fewer issues were published in the original four-page format, and were instead typically reduced to two pages per issue. We project having the entirety of the Richmond Daily Dispatch keyboarded, from the original run of November, 1860 through May, 1865, completed by the end of August, 2005.

9. Summary of two white papers

The first white paper, *The importance of yesterday's news: challenges and opportunities in newspaper digitization* (http://oncampus.richmond.edu/academics/library/digital/Documents/White%20Paper%20on%20Newspaper%20Project.doc or http://tinyurl.com/54rtm), completed in the fall of 2004 surveyed the currently available digital newspapers projects and compared their major features. It also provided an overview of available software options and recommended best practices for an ideal digital newspaper collection. An updated version will be mounted on the project website. The second white paper, *The many uses of newspapers* (http://oncampus.richmond.edu/academics/library/digital/Documents/The%20Uses%20of%20Newspapers.doc or http://tinyurl.com/5b7hf), examines the different research and personal uses of newspapers by various communities including historians, sociologists, linguists, scientists, teachers, genealogists and public library patrons. It explores the uses of historic newspapers in a variety of formats, including hard copy, microfilm and digital, since little research has currently been done on the use of historic digital newspaper collections. This white paper also explores how different parts of the newspaper such as advertisements, editorials, obituaries and political cartoons have been used for different research purposes. It has been mounted on the project website. The research findings of these two papers will be used in helping determine how best to structure the newspaper collection and what types of searching and browsing to support.

10. Current tagging work at the Perseus Project

Work continues on refining the automated tagging of the newspaper XML files. The Perseus Project is currently trying to determine what types of additional tagging to support such as the tagging of commodities and organizations. Currently the newspaper files have been tagged for persons and places. Research continues into determining what searching functions will be supported with this collection, and learning what the maximum amount of information is that can be obtained from these newspapers. One current question we are exploring is how to best tag newspaper advertisements so we can identify products and commodities in a useful manner. Newspaper XML files are being examined and corrected to determine tagging accuracy levels for named entity recognition and to create training sets to refine the automatic tagging.

11. Authority list creation and refinement

Research continues into determining how best to structure and integrate a number of authority lists that will be used to help refine named entity recognition within the newspaper collection. Harper's Gazetteer of the World has been digitized and tagging work continues on this reference source that will ultimately be used as one possible authority list. Other major nineteenth century reference sources will soon be digitized to serve as additional authority lists.

Appendix #1: October 2004 Conference Agenda

Civil War Historians Conference
October 25, 2004
Brown-Alley Room, Weinstein Hall

Agenda

9-10am	Registration, continental breakfast
10am	Welcome June Aprille, Provost
10-10:15am	Introductory Remarks Bob Kenzer, UR
10:15-10:40am	Present status of grant project Rachel Frick, UR
10:40-11:05am	Review of national digital newspaper projects Alison Jones and Gwynne Langley, Tufts
11:05-11:15am	Break
11:15-11:45am	Overview of other civil war/digital projects Elizabeth Roderick, UVA
11:45-12:15pm	African-Americans during the Civil War Carey Latimore, Trinity University, Texas
12:15-1:15pm	Lunch, Jepson Faculty Lounge Participants will be seated by group number
12:45-1:15pm	Desired content on the site Group discussion/brainstorming session
1:15-1:25pm	Return to Weinstein Hall
1:25-1:50pm	Desired functionality on the site Group discussion/brainstorming session
1:50-2:15pm	Resources that individuals may be able to contribute Everyone together
2:15-2:45pm	Where digital libraries are headed and how this project fits in Greg Crane, Tufts
2:45-3pm	Next steps for the group; evaluation Bob Kenzer, UR

Appendix #2: Participants in the October 2004 Conference

Civil War Newspaper Project
University of Richmond
October 25, 2004

Participants

Scott Arnold
Dept of Historical Resources
scott.arnold@dhr.virginia.gov

John Barden
University of Richmond
jbarden@richmond.edu

Michael Bell
Independent Scholar
mebell2@cox.net

Greg Colati
George Washington University
gcolati@gwu.edu

John Coski
Museum of the Confederacy
JCoski@moc.org

Greg Crane
Tufts University
gregory.crane@tufts.edu

Martha Crawley
IMLS Project Director
mcrawley@imls.gov

John Deal
Library of Virginia
Jdeal@lva.lib.va.us

Rachel Frick
University of Richmond
rfrick@richmond.edu

Meghan Glass
Valentine Museum
archives@richmondhistorycenter.com

Mike Gorman
Richmond National Battlefield Park
Mike_Gorman@nps.gov

Jim Gwin
University of Richmond
jgwin@richmond.edu

Doug Harvey
Tredegar National Civil War Center
dharvey@tredegar.org

Tom Illmensee
Virginia Historical Society
tillmensee@vahistorical.org

Alison Jones
Tufts University
alison.jones@tufts.edu

JP Jones
University of Richmond
jjones@richmond.edu

Bob Kenzer
University of Richmond
rkenzer@richmond.edu

Gregg Kimball
Library of Virginia
gkimball@lva.lib.va.us

John Kneebone
Virginia Commonwealth University
jtkneebone@mail1.vcu.edu

Nelson Lankford
Virginia Historical Society
nelson@vahistorical.org

Gwynne Langley
Tufts University
gwynne.langley@tufts.edu

Carey Latimore
Trinity University
carey.latimore@trinity.edu

Kevin Levine
St. Anne's Belfield School
kevlvn@aol.com

Jeffrey McClurken
University of Mary Washington
jmcclurk@umw.edu

Leigh McDonald
University of Richmond
lmcdonal@richmond.edu

Ann McMillan
Novelist
rahhallman@aol.com

Gail McMillan
Virginia Tech
gailmac@vt.edu

Steve Ramold
Virginia State University
sramold@vsu.edu

Jim Rettig
University of Richmond
jrettig@richmond.edu

Elizabeth Roderick
University of Virginia
eroderick@earthlink.net

Andrew Rouner
University of Richmond
arouner@richmond.edu

Suzanne Savery
Valentine Museum
ssavery@richmondhistorycenter.com

Errol Somay
Library of Virginia
esomay@lva.lib.va.us

Brent Tarter
Library of Virginia
btarter@lva.lib.va.us

Will Thomas
University of Virginia
wgt9m@virginia.edu

John Wilkes
Governor's School
jwilkes@gsgis.k12.va.us

Nancy Woodall
University of Richmond
nwoodall@richmond.edu

Appendix #3: 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%