"A Testbed of Civil War-Era Newspapers."

Abstract

Newspapers contain a significant proportion of the 19th and 20th-century American cultural heritage material important to persons interested in the social and cultural development of the United States. Their content is heterogeneous and flexible, simultaneously reflecting national politics, local events and concerns, and a range of minutiae significant to research in demographics, institutional history, and consumerism, among other topics. Few primary historical sources serve an audience so diverse, ranging from K-12 to professional historians to college and Elderhostel students. However, until recently digital libraries of historical newspapers have been hindered by prohibitively high keyboarding costs, technical limitations on OCR of microfilm sources, and problems with storing and transmitting high-quality images of sufficient size to cover a newsprint page.

The University of Richmond proposes a two-year project to produce a digital testbed of newspapers relating to the Civil War, in collaboration with the Perseus Project at Tufts University and the Virginia Center for Digital History. The project has four primary goals. First, to accumulate newspaper images from both Union and Confederate perspectives centering on the vital years 1857-1865, with metadata available for all articles and full-text for 5% of the corpus. Second, to produce a body of best-practice recommendations for digitizing newspapers that analyzes the costs and benefits of a variety of acquisition methods, from uncorrected OCR of physical page images to keyboarding from microfilm, and to disseminate these throughout the digital library community. Third, to strengthen cross-institutional sharing of content with a particular focus on the American South in the 19th century, by cross-linking newspaper materials across existing corpora such as Perseus' 40,000,000-word Civil War collection and the Virginia Center for Digital History's Valley of the Shadow project. And fourth, to develop workable models for digital library technology transfer, by exploring how the existing expertise of the Perseus Project and the Virginia Center for Digital History can be used to seed digital library activity at the University of Richmond, using the development of this collection as a training set.

Outcomes of the project will be: a website hosted at Richmond providing metadata search for articles and serving page images of newspapers; a publicly-available testbed of digital source materials for further research into searching and improving the quality of OCR on newspapers; a set of best practice guidelines for the acquisition of historical newspapers, with a particular emphasis on the cost/benefit balance of various forms of text content acquisition and intended for widespread use by later newspaper projects; and a sustainable infrastructure at the University of Richmond appropriate to the further pursuit of digital library activities.
National Impact.

Newspapers contain a significant proportion of the 19th and 20th-century American cultural heritage material important to persons interested in the social and cultural development of the United States. Their content is heterogeneous and flexible, simultaneously reflecting national politics, local events and concerns, and a range of minutiae significant to research in demographics, institutional history, and consumerism, among other topics. Few primary historical sources serve an audience so diverse, ranging from K-12 to professional historians to college and Elderhostel students. However, until recently digital libraries of historical newspapers have been hindered by prohibitively high keyboarding costs, technical limitations on OCR of microfilm sources, and problems with storing and transmitting high-quality images of sufficient size to cover a newsprint page.[1] This project proposes to build a digital testbed of three key newspapers from a pivotal moment in United States history, the Civil War, to investigate historical uses for a synchronic body of digitized newsprint and the cost/benefit tradeoffs of several methods for acquiring digital copies of newsprint, ranging from unadorned OCR to perfect double-keyboarding.

The proposed work consists of both a demonstration and a research component. While each of these represents a viable self-standing project, the principal researchers felt that a joint project would strengthen each side. Hence, the demonstration component at Richmond will participate in a research agenda directed by the Perseus Project at Tufts and supported by the Virginia Center for Digital History at UVA. Conversely, Perseus and VCDH have substantial experience in digitizing collections and will be able to help Richmond design and implement its digitization efforts.

The demonstration component will center on three newspapers illustrating the debate over slavery and the Civil War period. We will digitize a full run of William Lloyd Garrison’s weekly, the Liberator (1831-1865), providing a thirty-five year view of the anti-slavery debate. Additionally, we will acquire runs of two daily papers, one from Richmond and one from Philadelphia, for the period 1857-1865; these will provide a range of perspectives on national and local events from a Confederate and a Union vantage point, respectively. These newspapers will be linked to supplementary materials from the Virginia Center for Digital History’s well-known Valley of the Shadow Project: we chose Philadelphia and Richmond daily papers in part because these newspapers were particularly important to the Pennsylvania and Virginia communities which the Valley of the Shadow already covers. While Perseus is best known for its classical collections,[2-5] more recent work has focused on topics such as Shakespeare and the History of Mechanics,[6] a paper describing Perseus’ work on a new collection about London (which subsequently received IMLS support) won the Vannevar Bush award for best paper at the 2001 Joint Conference on Digital Libraries.[7] Over the past two years, Perseus has assembled a 40 million-word testbed on slavery and the Civil War. This collection already spurred collaboration between Perseus and Civil War historians at UVA. This project will establish a forum within which Perseus and VCDH can extend the interconnections between their Civil War materials, with the new newspaper collections at Richmond, and with other collections on the Web.

Research themes include the following:

Cost/benefit studies for digitizing newspaper from various sources and to various levels of accuracy. Creating digital versions of historic newspapers constitutes a major challenge, given the quality of newsprint, the complexity of the page layout and sheer size of the pages: the most advanced digital newspaper technology available is tailored to support information seeking via fulltext search (http://www.olivesoftware.com), but does not produce a clean text transcription of newspaper text. The costs of various input methods can vary radically: At the high end, double keyboarding a page of newsprint can cost $36 or more while OCR of microfilm source lends itself to automation (www.nytulib.org; publications in New York Times project) and could be accomplished at 1/100th and probably 1/1000th the cost—the per page costs decline precipitously as a stable system ingests more pages. Perseus is already engaged in a four year NSF ITR Project led by Sayeed Choudhury, Milton S. Eisenhower Library Hodson Director of the Digital Knowledge Center at the Sheridan Libraries. This project builds on previous IMLS research and is investigating new technologies for OCR. The newspaper collection at the core of this project will provide us with an important testbed for study. We are particularly interested in the relative merits of microfilm (of which vast quantities of varying quality already exist) and directly digitized page images (such as we will create in this project for the three selected newspaper runs). This study includes not only comparisons between different OCR systems (e.g., the commercial Primerecognition software available
at Perseus and the open source Gamera system developed at JHU) but also a survey of the readability of microfilm vs. page image: in many cases, we find the microfilmed sources are damaged or shadows from the curvature of the bound source make the inner sections of a page illegible: 100% accurate OCR from such materials will still not provide 100% accurate transcripts of the original. Similarly, we will study the impact of varying error rates from uncorrected OCR on full text searches of various kinds (e.g., [8]). Our goal will be to assess not only numbers but also the impact that these numerical rates have upon different audiences as they perform different tasks. We are particularly interested in the impact of error rates on potentially new forms of search: error rates may have less impact on full text retrieval to which our users are now accustomed than on more sophisticated searches (e.g., the ability to ask a system not only to find instances of Springfield but to identify those passages that refer to Springfield, Missouri). Present cost/benefit analyses must not only regard present practice but also look towards future uses as well.

Collection development strategies for thematically coherent digital libraries: In its earlier classical efforts and more recent DL/2 work, Perseus has stressed the creation of digital reference works that can be mined for significant phrases, glosses, and authority lists for people, places, organizations and other named entities.[9-13] The goal is to develop systems that can automatically assemble contextual materials relevant to a given object by creating a network of interlocking electronic reference materials.[14] This work has led to a National Science Digital Library Services Project for automatic linking and authority list management - services which we will be able to exploit in the work proposed here. The relevant reference works vary from domain to domain, but basic patterns are beginning to emerge and the newspapers will provide us with another important case study. An ongoing IMS project at the Tufts library is entering historical nineteenth century city directories for Boston and will provide us with authority lists of names, addresses, and other statistics. Documents already digitized in Perseus and Valley of the Shadow include resources such as Dyrr’s Compendium of the Civil War (which has comprehensive lists of battles, regiments, and divisional commands). We will work to identify additional reference sources (e.g., city directories for Philadelphia and Richmond, the first (1869) Rowell and Company American Newspaper Directory, which provides vital statistics for many of the newspapers cited in the three newspapers which we will digitize). This element of our research addresses the problem of creating highly-integrated scalable electronic collections.

Automatic tagging and metadata generation: A third strand of research will examine the challenges of providing metadata: both traditional cataloging records and broader new categories (e.g., lists of people, places and things in a given document). The collection that we will assemble will include more than one hundred thousand individual news stories and is simply too big for manual cataloguing. Automatic document analysis allows us to generate far denser markup than in comparable collections that focus primarily on structural markup, with more than five times as much markup in the Perseus Civil War collections than in comparable documents from the American Memory or Documenting the American South: most of these tags encode personal names, dates, places, military units and other named entities. The Perseus Project members and I have already published about the need for corpus editors, new specialists who combine domain knowledge with an ability to adapt information extraction systems to their particular domains,[15, 16], while information extraction systems may emerge as standard digital library services.[17-22] We know from these and other studies how effective different language technologies can be, but we often do not understand the significance of these figures. How useful is a system that correctly guesses that a given “Springfield” in a given text is Springfield, MO (vs. Springfield, MA, NH, etc.) 80% of the time? How much labor does it take to bring the precision and recall up to 90% 95% 99%? To what extent do these figures vary with genre and source quality? We will draw upon our colleagues at the Simmons Graduate School for Library and Information Sciences to help us study the costs and benefits of the metadata that these emerging systems provide.

Information retrieval models for disambiguating OCR: Research indicates that information retrieval systems’ performance improves when algorithms incorporate authority list materials.[23] such as can be generated from reference markup produced by humans. By integrating results from several approaches to markup (uncorrected OCR, cleaned OCR, keyboarded text) for the same sources, we will be able to evaluate the efficacy of using human-produced markup to train disambiguation algorithms working over output from standard OCR engines. In many cases we will be able to exploit genre consistencies particular to newspapers: articles often begin or end with a newswire source, date, and city of reference, and use a closely-related set of vocabulary.[24] When evaluation of a manually-keyed article from a given date and location indicates a close correlation between two named entities, the probability that articles from uncorrected OCR output containing one of these items will contain the other is high. Hence, approaches common to the information retrieval community for retrieval of “dirty” OCR [25] can be applied to historical newspapers to provide full-text search and aid in disambiguation.
XML markup and structured information retrieval: While metadata-centric approaches such as the Open Archives Initiative and Encoded Archival Description have proven the viability of digital libraries as tools for locating historical sources, much work remains to be done on the retrieval and analysis of actual document content, particularly when it is in highly-annotated forms such as the Text Encoding Initiative recommendation. Such markup can be extremely useful to searchers, however, for it encourages semantic as well as syntactic searching strategies. For example, using a markup-aware retrieval system users can formulate a query that exploiting existing markup, whether automatically- or hand-generated, to search only for references to Washington the person (<persname>Washington</persname>), and excluding references to Washington, D.C. (<placeName>Washington</placeName>). While querying structured data is a topic of growing concern in the database community, (cf. the W3C XQuery effort, http://www.w3.org/TR/xquery, [26, 27]) digital library applications require additional functionality, such as relevance weighting, fulltext integration, and support for structural indeterminacy.[28] We will work with experts in markup-aware information retrieval to investigate the applicability and uses of such searching for our corpus of newspaper data.

Adaptability.

Documented style sheets for evolving best practices: While this project will explore encoding practices that draw on emerging information extraction technologies differing substantially from that of those collections which we have been able to study (especially Michigan and Cornell’s Making of America, the Library of Congress’s American Memory, UNC’s Documenting the American South), our evolving practice draws more heavily upon elements from the established and well documented Text Encoding Initiative. On the one hand, Perseus and VCDH will use this project as an opportunity to make their collections more closely interoperable. At the same time, building on their experiences with the established VCDH Valley of the Shadow Digital Library and a new 40 million word Perseus Civil War collection, they will work with Richmond, the three allied projects, other collections and especially the academic advisors to establish a common style-sheet and best practices for new materials and old ones. This project not only seeks to create useful new materials but to extend existing best practices. The DocSouth documentation for text encoding (http://docsouth.unc.edu/) has proven extremely helpful to us in our own work and we will produce similar descriptions of our own shared practices. Our guidelines will extend beyond text markup, documenting the costs and benefits of new digitization of newspapers with varying levels of refinement.

Newspaper digitization: While much of our work will have broad application, this project places particular stress upon the prospects and challenges associated with newspapers. We will study the comparative advantages of micro-film vs. original scanning, uncorrected OCR output from both sources, clean-up of OCR of varying quality vs. original double keyboarding, etc. Of particular interest are the many newspaper advertisements that provide crucial social and economic data but are often left out of digitization projects and are, indeed, often quite difficult to encode, with idiosyncratic content and formatting.

Technology transfer: The collaboration between the University of Richmond on the one hand and Perseus and VCDH on the other constitutes an interesting case study of technology transfer. By providing joint support to Richmond, Perseus and VCDH seek both to expand their own collaborations and to help another institution use a major digitization project to drive the evolution of its library infrastructure.

Open source services: Perseus, Tufts’ Digital Collections and Archives and the Digital Knowledge Center at Johns Hopkins are collaborating on a new National Science Digital Library Services grant to create automatic linking and authority list management services. This NSDL project builds on the partners’ previous IMLS and D1.1-2 funding, allowing the collaborators to extend and generalize their own internal services, making these accessible to a wider audience. While we already have the capability to meet the technical needs of the project proposed here, the NSDL support will allow us to support a much broader audience of users and collaborating projects.

Article-level image segmentation and web access: Part of Perseus’ research program will address image segmentation and caching for dynamic publication from digital repositories for very large images. Leveraging our existing technology for disseminating raster GIS data at several levels, we will investigate possibilities for addressing transmission- and storage-level issues for newspaper archives by segmenting and delivering individual articles derived from full-size TIFFs.

Open Archive Initiative metadata exchange: Perseus is a registered OAI 2.0 data provider. The IMLS has commissioned a team from Graduate School of Library and Information Services to help develop OAI metadata for the IMLS.
community. This project will play an active role in supporting the broader effort. Much of our research will work with widely available tools and services, and will concentrate on the costs/benefits of automatic and semi-automatic (i.e., scalable) metadata generation for large collections.

Digital repositories 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).

Design.

The following describes the production of the core newspaper runs.

Image acquisition: 22,500 page images for newspapers; c. 30,000 pages of supporting materials: Imaging will be performed using paper sources and a Minolta PS7000 planetary scanner operating at 400dpi in 8-bit greyscale. OCR input and presentational versions will be derived from these masters, thus producing user-legible images even for pages that produce OCR of low quality. Recently, the VCDH captured images of the Franklin Repository, a Civil-War era newspaper, from microfilm. Interpolating from storage requirements for this project (4.6mb/b&w; 37/8-bit greyscale), we calculate our corpus of 22,500 physical page images at 400dpi will require approximately 1.4T of space in 8-bit greyscale (66mb/piece). Note that the supporting materials reflect (1) much smaller pages (typically 2K of text vs. 50-70K of text on a newspaper page). Many supporting materials have been recently reprinted and these modern reprints can be disbound and scanned quite rapidly (c. 1,000 pages/hour).

Archiving of core data in a digital repository: Richmond will provide the ultimate home for the data collected in this project. Both Virginia and Tufts are, however, collaborating in the development of FEDORA, a Mellon sponsored digital repository system. While Richmond determines its repository infrastructure (e.g., FEDORA, MIT’s Dspace, etc.) Virginia and Tufts can manage Richmond’s repository services. A reliable file server (Proliant ML530 dual-processor @2.4Ghz w/1T SCSI drives) will fill needs of the project through the first year of work (through fall 2004), and Richmond will have time to put its own system in place as part of its general library/IT development.

Image processing: We found that specialized image processing software for page images (Book Restorer) substantially improved subsequent processing. Scanned page images from print will be processed before being OCR’d. Original greyscale images will, however, be stored archivally and used to produce presentational versions; these will constitute part of the publicly available testbed of digital sources.

Page segmentation and article metadata—c. 112,000 columns of newspaper text: Various algorithms exist for segmenting multi-column page images but none are foolproof. The Universal Library project at CMU has developed very reliable page segmentation software for the New York Times, but Robert Thibadeaux reports [e-mail Tue. 18 Sep 2001] that building a page segmented is a difficult task and generally requires about a year of labor. Persius therefore outsources page segmentation to a data entry contractor and we anticipate continuing this practice in this grant. Page images are shipped to the contractor who then returns coordinates for named regions on each page (e.g., “coll start-x, start-y, end-x, end-y”). We will thus have presentational images and metadata for articles, together with an archival data set useful not only for our own project needs, but which can then be used as a training set/evaluation instrument for subsequent page segmentation systems.

OCR (Perseus)—c. 1.3 gigabytes of newspaper textual data: Once individual pages are digitally restored and article regions have been identified, OCR software will be applied. The Perseus Project owns a high-end OCR system (Prime Recognition) in which six different OCR engines vote on each character. The OCR output produces ASCII files that include both the confidence of the OCR system on each scanned character and the coordinates of each character as well. These files allow us to link text to original source images. We will archive the raw OCR output files in this ASCII format, preserving page layout, location of characters on the page and the OCR confidence levels for each character. The confidence and location data expands the size of these files but increasing storage requirements by an order of magnitude (from 1 to 10 gigabytes) would have little practical impact; given our storage capacity.

Article metadata OCR Cleanup/Contractor—c. 3% of running text (65 mbbytes): OCR output is transmitted back to the data conversion contractor for subsequent cleanup. In this project, headers of all newspaper articles will be corrected and
marked with Text Encoding Initiative (TEI) tags. In practice, we have found the data conversion contractors proficient at adding fairly complex structural markup (e.g., the fields within openers and closers of letters). The structural markup in the newspapers is less challenging than that in some of the reference works on which we have worked and very similar to that in many of the documents that we have already entered. These corrected and tagged headers will provide the basis for metadata about the documents within the newspapers (including important information, such as the sources from which the various newspapers draw their excerpted). The added structural markup will allow us to determine the beginning and end of individual articles, thus facilitating more sophisticated textual analysis than is possible if each page or column is an undifferentiated “bag of words.” Furthermore, since we preserve the original OCR output files (with the coordinates linking each character to a location on the page image), we will be able to generate the coordinates not only for columns but also for items within and across columns.

Clean (99.95% accurate) reference sample (Contractor/Perseus/VCDH/Richmond)—c. 65 mbytes of tagged newspaper text (5% of the whole) and c. 50 mbytes of supporting materials: we will select a subset of the overall collection for careful data entry and more detailed work. Even this subset will be substantial (65 mbytes). The magnitude of this subset will vary according to the actual quality of the original print (which varies widely). Although we will solicit new bids for this project, we have been able to commission reliable, 99.95% accurate data entry with basic structural markup $600/mbyte. Where the OCR output is sufficiently accurate, however, we can achieve our goal of 99.95% by having the contractor clean-up the OCR output. The cost of clean-up varies with the quality of the output and complexity of the document, but the cost is generally about 50% that of manual data-entry. The figure of 65 mbytes for newspaper data assumes that much of the reference sample will have to be entered manually but that the more inexpensive OCR clean-up approach will be appropriate to some. The figure of 50 mbytes for supporting material is an estimate and reflects the heterogeneous nature of such supporting materials. NB: while we have allocated enough resources to create a rich set of supporting materials, it is possible (though not likely) that we will choose to reallocate some of the supporting materials budget to extend the core newspaper coverage. At present, 23% ($30K out of a project total $130K) is allocated to supporting materials.

We will use this subset of newspaper materials as a training set to enhance automatic analysis for the rest of the corpus. We will also release this subset as a testbed for language technology researchers, inviting them to apply to these historical newspapers the techniques that they have developed for modern testbeds (often themselves collections of newspaper stories) for such evaluation forums as TREC, DUC, and ACE.

Review and refinement of structural markup (Richmond): Richmond will review the structural markup for the overall corpus and be responsible for converting this into a format compatible with the planned IMLS OAI Metadata standard.

Named Entity and Relation Detection and Tracking (Perseus): The scale of this project will force us to rely on automated and semi-automated methods, to identify the people, places, organizations and other entities in the article metadata. This is where the research component of the project, led by the Perseus Digital Library Project at Tufts, begins. Perseus has already implemented named entity identification routines that track people, places, and objects such as military organizations (e.g., automatic recognizing that “10th Mass.” “Tenth Massachusetts Infantry,” “Mass tenth” etc. all refer to the same entity), ships (e.g. the “Merrimack” as ship rather than as place) newspapers, railroads, etc.

Relation detection describes the task if identifying connections between named entities (e.g., “10th Mass. at Gettysburg,” states that the “10th Mass Regiment” was at Gettysburg. We track named entities and relations when we try to resolve references (e.g., “it” “10th Mass. Regiment”) and changes (e.g., infer that the 10th Mass. had moved from Gettysburg to Harper’s Ferry).

Authority list integration (Perseus): Named entity identification has two fundamental stages. Semantic classification determines whether Springfield in a given context describes a rifle or a place. Once we know that Springfield is a place, we still need to determine which of the dozens of Springfields in a major gazetteer the particular passage describes. Likewise, we may be able to distinguish whether a given “General Lee” refers to Robert E. Lee, Fitzhugh Lee or Sidney Lee. In a given document, the fullest citation for this name may be “Robert E. Lee,” “R. E. Lee,” “Robert Edward Lee,” etc. We need to link all of these descriptions to a canonical listing (e.g., “Lee, Robert E. (Robert Edward), 1807-1870”).

Evaluation and Refinement of Results from Automated Methods (Perseus and Richmond): In preliminary analysis of the errors that automatic systems make, researchers at Perseus have found that a small number of errors often account for a large percentage of the errors. We should, for example, assume that all references to “West Point” refer to the location of
the US Military Academy and not any of the 69 other West Points listed in the Getty Thesaurus of Geographic Names. Thus, even though the narrative may describe events in Virginia, assume West Point, NY unless the text includes a marker such as “West Point, Virginia.” Such assumptions can be represented as heuristics (e.g., simple lists of defaults such as “if West Point appears, assume West Point, NY”) or statistical measures from training sets (e.g., human editors edit machine generated tags for a corpus; the system then observes that all 38 instances of West Point describe “West Point, NY”) or a combination of the two (e.g., create heuristics to bootstrap automatic analysis, correct the results and then use the resulting statistics to refine subsequent automatic analysis). Substantial work has gone into studying how most efficiently to create such systems in the intelligence community (e.g., [29]). We will study the applicability of this work to those who create and use historical collections such as those at the center of this project.

Publication of metadata via the OAI: The IMLS has begun developing its own community OAI metadata standards. We will generate metadata that conforms as closely as possible to these standards. Much of our research will focus on determining the most useful categories of metadata that we can manually create and scalably generate from this corpus. OAI metadata should allow users to pose queries such as “how frequently do the three newspapers quote stories from the Charleston Mercury?” [tracking the source contributions in excerpted stories], “what news stories are most relevant to events near Lynchburg, VA in August 1863?” [tracking places and dates in full text of individual articles], “which stories are most relevant to Sidney Lee? The 54th Massachusetts?” [named entity identification and tracking to keep R. E. Lee from swamping references to Sidney Lee and to map the multiple permutations of 54th Mass. Infantry].

Management Plan and Personnel

The partnerships involved in this project are in some ways easy to understand. Boston and Richmond are logical partners for a project on slavery and the civil war. The University of Virginia and the University of Richmond are fairly close and well-suited to collaborate. Perseus and the Valley of the Shadows are two of the largest and longest running digital humanities projects in the United States, with growing ties between project staff. Perseus’ DLI-2 project has allowed it to branch out beyond classics, with most recent work focusing on the Civil War and 19th century America.

The partnership between Tufts and the University of Richmond reflects an established working relationship dating from when June Aprille, Richmond Provost, was Associate Provost at Tufts. Tufts PI Crane is a member of the Information Technology Council, on which Dr. Aprille sat. The two of them worked together to upgrade the CIO at Tufts to a Vice-President, establish a permanent Geographic Information System center in the main Tufts library, found the Web services department for the university and, most importantly for this project, to create the Digital Collections and Archives group at Tufts. Founded only in fall 2000, the DCA, led by Gregory Colati, has already received two IMLS grants and has established itself as an emerging force in digital library work. Crane and Aprille worked on planning each of these. Associate Provost Aprille played the leading role in making these ideas reality at Tufts. The same partnership between Crane and Aprille is responsible for this proposal. In the large picture, Crane can provide the technical vision. Aprille can provide the leadership needed so that this particular project at Richmond becomes the catalyst for systemic development at Richmond, rather than simply a one-time effort.

This project could be split into separate proposals for digitization (Richmond) and research (Tufts), but we felt that the two projects were best if integrated. As a result, each major partner will participate in the work of the other: Richmond will contribute to the research at Tufts, while the extensive experience in digital collection development by Perseus and VCDH provides the foundation for Richmond’s digitization efforts. Each partner could accomplish its basic goals in isolation, but working together the two can accomplish far more than either could in isolation.

The University of Richmond will facilitate an initial project kick-off meeting and regular communication throughout the duration of the project. Richmond plans to hold a monthly conference call to discuss progress and issues. Towards the end of the first year, Professor Kenzer (below) will host a seminar for Civil War scholars to inform them about the Civil War Era Newspaper repository and to get their advice and input.

Principal investigators Jim Pettig (Richmond) and Greg Crane (Tufts) will collaborate in overseeing the project. The Richmond, Tufts, and UVA teams will convene as a whole three times during the course of the project, at UVA and Richmond during the planning and acquisition phases of the first year and at Tufts the final, research-oriented year; they will also hold informal meetings at the IMLS Web-Wise conferences each year. Other administration will be conducted by Lisa Carrato (Managing Editor at Tufts, Rachel Friek (Head, Bibliographic Access Services) at Richmond, and Kim Tryka (Associate Director) at VCDH.
RICHMOND

The Richmond principal investigator will be Jim Rettig, University Librarian (vita attached.) Mr. Rettig will devote 10% of his time to supervising and coordinating the project. He will manage the budget, coordinate and submit the 6-month performance reports, and the financial and project-end performance reports. An administrative assistant has been assigned at 5% to provide clerical support, coordinate meetings, and distribute materials as required.

Robert C. Kenzer, William Binford Vest Professor of History (vita attached) specializes in the Civil War era, particularly the southern side of the conflict. He is author and editor of several books, ranging from impact of the Civil War on family and community life, to black economic opportunity in the postwar years and Southern unionism. Professor Kenzer will devote 20% of his time to this project, providing research support and serving as our primary scholarly resource on the Civil War.

A full-time digital materials staff-person at Richmond will work with Tufts and UVA staff to coordinate scanning and markup of the source materials. He or she will perform the majority of the tagging, clean-up, and authority list work; will work with Tufts and UVA to design a web interface for the newspaper repository; and will be responsible for testing and quality-assurance.

Jim Gwin, Richmond’s Head Collections Development Librarian and library liaison to the History Department, coordinates collection-building with curricula and research needs, evaluates collections and assesses the needs of academic programs, and manages collections-related spending. Mr. Gwin will devote 10% of his time to research support for the project, focusing on issues of accessibility and applicability to undergraduate research and serving as liaison between project staff and faculty.

Nancy Woodall, Head of Library Systems, oversees all aspects of technology in the libraries, integrating technology into library workflow and overseeing the design of user-oriented access to electronic library resources via the Voyager integrated library system and the ENCompass software suite. Nancy will spend 10% of her time on the project helping to evaluate, design and test technical infrastructure, and will evaluate ENCompass FEDORA and Dspace repositories.

Leigh McDonald, Head of Cataloging, is responsible for cataloging policy, creates library project metadata schemas, directs special processing projects, and acts as liaison on cataloging and technical processing issues. Leigh is assigned to this project 10% of her time to advise and oversee metadata creation.

Rachel L. Frick, Head, Bibliographic Access Services, leads staff providing the libraries with acquisition, cataloging, and serials control services for print, media, and electronic resources. She will devote 10% of her time to increasing the level of bibliographic support for digital materials and to development of finding tools for them.

An Academic Technology Services liaison (at 10% commitment) will implement the web interface and do usability testing.

A Database Administrator will evaluate, select, and design a database for the digital repository, and install, test, and maintain database services.

An Information Services System Administrator will provide the project with hardware and technology infrastructure services.

TUFTS

Professor Greg Crane, co-principal investigator for this grant, holds the Winnick Family Chair in Technology and Entrepreneurship in Classics at Tufts University and is Editor-In-Chief for the Perseus Project. He will devote 10% of his time to the project, coordinating the technology transfer to Richmond and overseeing the research effort in natural language technologies for markup of newspaper content.

Lisa Carrato, Managing Editor at the Perseus Project, will devote 10% of her time to managing the data acquisition process, supervising the scanning staff-person and coordinating with the data entry contractor.
Anne Mahoney, Perseus Project Programmer, will serve as engineer for the natural language research project, working closely with the postdoctoral researcher in information science.

A postdoctoral researcher in information science will devote 100% of his or her time to Tufts’ information extraction and markup-generation research program.

A graduate student worker will do scanning, and schedule the OCR, and image clean-up processes.

VIRGINIA CENTER FOR DIGITAL HISTORY

Kim Tryka, Associate Director of the VCDH, will work as a consultant for the project, developing markup and metadata standards to be implemented by the Richmond markup team. Because of her background in Library and Information Science (MLIS, University of Pittsburgh), Kim will play an integral role in bridging the traditional and digital library practices during the technology transfer project to the University of Richmond.

Will Thomas, Director Virginia Center for Digital History, will provide additional consulting on digital history and assist in the technology transfer effort.

A graduate student worker will do scanning and image clean-up.

Budget and Contributions

The budget and contributions are detailed on the accompanying sheets.

Project Evaluation.

Project evaluation will proceed on four fronts and combine traditional and outcome-based evaluation methodologies. Traditional evaluation will include:

Collections availability: The project will build a collection of page-front images and searchable article-level metadata for 22,200 newspaper pages. Success of the collection-building effort may be evaluated by measuring the number of newspaper pages available for public access and the proportion of their content available as corrected running text. The extent to which this number exceeds the baseline of corrected text for article headings only will provide a measure of the project’s success in the arena of natural language technology-assisted OCR.

Technology transfer: The strength of Richmond’s digital library infrastructure and activities will provide a measure of the project’s goals for inter-institution technology transfer. Metrics for this include user statistics on Richmond’s project-specific web and OAI servers, the continued presence of personnel trained in digital methodologies, and the extent computational and physical infrastructure available for digital library work at Richmond. A final indicator of successful technology transfer will be Richmond’s success in garnering follow-up funding for digital library activities.

Outcome-based evaluation will include:

Best practices dissemination: Development of best practices style sheets, public availability of testbed newspaper materials, and refinement of open strategies for handling the complexities of newspapers OCR are central components of this project. The number and size of other digital newspaper projects that publicly adopt our best practices recommendations and the number of publications based on research using our collections material as a testbed for work in natural-language, OCR, and information retrieval research, will measure the project’s outcomes in the digital library and information technology communities.

User-study outcomes: In coordination with the Perseus Project’s other user-study activities, the project will gather results from web questionnaires associated with the newspaper content. Deploying a several questionnaires will track different user profiles. An educational questionnaire will gauge the extent to which the project increases the use of primary source newspaper material among the K-12 audience. A research-oriented questionnaire will gather testimonial material on how digital newspaper research altered professional and collegiate research in history.
Dissemination.

**Web publication**: Most users will interact with the materials produced under this grant via the web as part of an integrated digital library. Tufts (Perseus and DCA), UVA (VCDH) and Richmond will exchange metadata with each other as well as with the rest of the OAI community, creating highly interoperable collections of their own and opening up their collections to third parties.

**OAI Metadata exchange**: As mentioned above, we will use the OAI to disseminate metadata about this collection.

**Publications via conferences and other forums**: Perseus project members regularly publish the results of their work in publications relevant to digital libraries, information retrieval and other technical areas. Collaboration with VCDH and its research into new forms of historical publication will help us reach historians of 19th century America.

**Workshops and outreach**: Boston and Richmond are centers for institutions that support the study of slavery and the Civil War. These include but are not limited to colleges and universities. We will hold workshops for postsecondary faculty, k-12 history teachers, historical societies and museums (e.g., the Bostonian society, Mass Historical Society, American Antiquarian Society), local historical societies. Civil War roundtables, etc.

**Collaborations**: The Tufts research component will provide opportunities for students from the Simmons Graduate School for Library and Information Science to participate. Such ties are already in place, since Simmons students have done much of the work on the two IMLS grants already awarded to the Tufts Library. This project will, however, provide more direct contact between the DLH2 research under Crane and the LIS program at Simmons. We see this project as a way of expanding connections between Tufts and Simmons, with Simmons faculty and students increasing Perseus' connections to the formal LIS community while Perseus offers new research opportunities for LIS students from Simmons. While our emphasis will be on scalable services and library collections, collaborations with the Boston University Editorial Institute (at which two Perseus researchers have already taught courses) allow us to exchange ideas with more traditional humanist editors.

**Advisory Committee**: Representatives from three complementary projects have agreed to serve as advisors to this effort.

- **Nuala Bennett**, the Coordinator of the Digital Imaging and Media Technology Initiatives at UIUC and was the project coordinator for Teaching with Digital Content, an IMLS funded project. She is a co-PI in UIUC's IMLS Funded Metadata Project: In a three-year IMLS research project, the Library of the University of Illinois at Urbana-Champaign will create a collection-level registry of digital collections created with IMLS funding from 1998 to 2005 and will research, design and implement a prototype item-level metadata repository service based on the Open Archives Initiative Metadata Harvesting Protocol. Her participation will help us coordinate our work, since evaluation of metadata practices will be at the heart of our work.

- **Sayeed Choudhury**, Hodson Director of the Digital Knowledge Center at the Sheridan Libraries of Johns Hopkins University: the DKC and Perseus are formally collaborating on two projects that augment our ability to conduct the work proposed here: the GAMERA ITR project (research on OCR and text acquisition) and NSDL SCALES Project (automatic linking and authority list management.)

- **Maria Daniels**, Director of New Media, American Experience, WGBH Boston: Web site associated with broadcast video have become a growing instrument for conveying historical materials to a broader community, with Web materials providing additional depth to the video while Web sites associated with video productions become key portals for further study. Collaboration with the American Experience allows us to integrate the needs of this audience into our design from the start.

- **Kenneth Haynes**, Assistant Professor of Classical Studies and Core Curriculum and Assistant Director, Editorial Institute: Two Perseus Project members have taught courses on electronic publication and structured markup for Boston University's Editorial Institute. The Editorial Institute is actively investigating ways in which emerging technologies augment the traditional role of textual editors.
Projects such as the one proposed here blur the boundaries between digital library production and editorial practice, but in essence, librarians and editors have a solid basis for collaboration, with editors performing more detailed markup that builds on the basis established by librarians. Thus collections such as those in American Memory aim at “level 4” TEI markup (capturing the structure [Friedland, 1999 #1301]) and leave the final, open ended “level 5” markup task to editors such as those in the Editorial Institute. This project thus provides a mechanism whereby to enhance conversations between programs such as Simmons’ GSLIS and BU’s Editorial Institute.

Sustainability.

Boatwright Memorial Library (BML) at the University of Richmond is committed to developing and maintaining digital collections to support, teaching and research both locally and throughout the scholarly community. Our commitment to supporting scholarship includes developing and acquiring high-quality, coherent digital collections that respond to the needs of students and scholars and that at the same time provide new avenues for intellectual inquiry and discovery. Our goal is to build, share, and preserve digital collections that are useful to the worldwide scholarly community. We are building the foundation for such collections by identifying the tools, strategies, and resources needed to provide the content, access, and services necessary to support digital library collections. Our goal is to apply the core values of the library, providing free access to organized, searchable digital content, while taking advantage of the speed, convenience, and the powerful, flexible capabilities technology offers.

Our strategic plan acknowledges that advances in information technology indicate that we should provide more learning opportunities that extend beyond transmitting knowledge from faculty to students. Digital library collections will help support an environment of intellectual curiosity and discovery.

The University of Richmond’s libraries have purchased and have begun implementation of the Endeavor Information Systems Corporation’s ENCompass software, a suite of tools specifically designed to assist in the creation of a digital library. A digital library committee composed of individuals from the Library, Information Services, the Center for Teaching, Learning & Technology, the University Museums, our Law School, and several departments has been formed. This committee has been meeting regularly to discuss issues of policy and management. Members of this group keep each other informed on initiatives and opportunities within our university and across the nation.

We have committed the time of a number of librarians and instructional web designers to this grant project because we want to ensure that the work will be maintained after the grant period. Furthermore, these librarians and others collaborating on the project will leverage the experience and knowledge gained in this project to undertake new projects of value to our faculty and of value to the scholarly community at large.

The University of Richmond is organized such that the University Librarian reports to the Vice-President for Information Services, Kathryn Monday. Ms. Monday sees the digital library as a synergistic melding of the knowledge and talents of the varied professionals in Information Services and is committed to its success now and in the future.

The University of Richmond has an annual Program Improvement Process whereby ideas that have been piloted and proven successful can apply for permanent funding. We will submit the digital librarian position initially funded here for permanent funding after this grant has been completed. We have had a great deal of success with this process in the past.

The University of Richmond has a robust technology infrastructure. We have ongoing budget allocations in place to refresh the majority of our servers and computing equipment on a 3-year replacement cycle. Our strategic plan speaks to our need to keep up with rapid changes in technology and to effectively integrate emerging technologies into teaching and learning. The plan articulates our objective to “...invest in infrastructure, including library resources, specialized equipment in the sciences, and cutting-edge computer technology, to support opportunities for student and faculty scholarship and creative activity.” In sum, the University of Richmond is ready and eager to undertake this project. We are committed to its success and to the ongoing maintenance of the Civil War Era Newspaper Repository.