Scholarship in the Age of Abundance: Enhancing Historical Research With Text-Mining and Analysis Tools

1. Significance and Goals of Project

The Problem of Abundance and the Future of Historical Research

Consider the plight of the future historian of the Clinton White House who faces the daunting task of looking through the 40 million email messages produced by the administration. If this researcher could read one email message every minute, it would still take 76 years to finish the entire corpus. Not even the most patient publisher or tenure committee could wait for the book that would result from such scholarly persistence.

Indeed, researchers already confront the problem of panning vast digital archives for historical gold and analyzing that data for patterns of meaning. JSTOR's archive of more than 2 million scholarly articles from the past century, the 14 million pages of newspapers in ProQuest's Historical Newspaper Collection, the 9 million items in the Library of Congress's American Memory Project, or the projected 15 million volumes in Google Book Search present both exciting opportunities and major challenges to traditionally trained scholars.

In the last decade the library community and other providers of digital collections have created an incredibly rich digital archive of historical and cultural materials. Yet most scholars have not yet figured out ways to take full advantage of the digitized riches suddenly available on their computers. Indeed, the abundance of digital documents has actually exacerbated the problems of some researchers who now find themselves overwhelmed by the sheer quantity of available material. Meanwhile, some of the most profound insights lurking in these digital corpora remain locked up. We believe the absence of appropriate methods and interfaces is largely to blame: digital content providers have not yet developed the kind of sophisticated and flexible search, extraction, and analysis tools capable of capitalizing on this vast investment in a digitized cultural heritage—a considerable outlay not just by the libraries engaged in digitization but also by those who pay for access to gated collections.

These more sophisticated methods and interfaces require what computer scientists call "text mining and analysis," which involves both the micro-analysis of texts (picking apart a chosen text by its constituent words) and macro-analysis of texts (using software and algorithms to reveal patterns and relationships within and between documents). This sort of computational analysis makes possible advances in the current research process for history and the humanities. Harvard University Library's Dale Flecker, who is working with Google to digitize Harvard's public domain books, recently told an audience at the American Library Association's annual meeting that the Google project has made him realize that "'text mining' will be an important part of research" in the near future.[1] (*Citations are in Appendix A.*)

For the humanist struggling with the problem of abundance, useful text mining and analysis promises major advances in three critical phases of the research process:

- *Locating documents of interest.* As collections grow exponentially, it becomes harder for individuals to find specific texts to study in detail for a particular research topic. Techniques such as document classification and clustering (finding similar documents in a collection) and recommendation systems (automatically finding documents based not on keywords but on what a scholar and his or her colleagues have already looked at) have shown potential on a variety of websites (such as Google's "similar pages" technology or the *New York Times*'s "related articles" sidebar) but have not been widely applied in academia. This category of tools combines the power of digital search with the long tradition in the humanities of the close reading of selected

texts. Experiments such as David Mimno's "Virtual Shelves," in which books that touch on the same theme but that are far apart in a physical library (because of their cataloging numbers) are grouped together on a web page based on a researcher's interests, demonstrate the promise of this approach. For instance, someone studying the history of canals might see economic books on canals, social histories of canal towns, and first-person travel narratives in a way that is not possible except through text mining.[2] But extending these methods to large bodies of digitized materials will not be possible without the application of more sophisticated text-mining techniques.

• *Extracting and synthesizing information.* After locating documents or an entire collection worth examining, possibilities other than close reading are now available using computational techniques. These techniques are far better and faster than a human reader for highlighting entities such as all place names in a large diary or all royalty mentioned in a set of early modern plays. Extracting these entities enables a much quicker appraisal of certain hypotheses—e.g., how far the diarist traveled during her lifetime or the theatrical portrayal of nobility in a certain decade. Such extraction also allows a researcher to interweave this information with other digital documents, such as maps, timelines, dictionaries, or glossaries. A scan of the personal accounts in the *September 11 Digital Archive* for the words "I prayed," followed by the extraction of the location of those praying and a placement of those locations on a digital map, shows the strong hold of prayer in the outer suburbs of American cities (and a commensurately low incidence in the inner suburbs and urban cores). Other techniques in this realm, in combination with more traditional methods and the close reading of texts, hold similar promise to make the invisible visible. For instance, it took decades for Hemingway scholars to notice a telling switch in the narration of "The Snows of Kilimanjaro" in which the author suddenly stops introducing new events with "And now," using "And then" instead, thus possibly signifying the death of the main character. Finding significant words, phrases, and other entities, or anomalies like Hemingway's telling "And then" can be done virtually instantaneously using text-mining techniques.

• *Analyzing large-scale patterns.* One of the unique, and potentially revolutionary, opportunities presented by digital collections and computational techniques is to examine trends over time, space, or other vectors in a large volume of texts. Such analyses are impossible to pursue in the humanistic tradition of close reading since no single person (or even group of scholars) could hope to read thousands of texts in a short time. In what Franco Moretti has called "distant reading," the goal is to make systematic assessments of a corpus, as opposed to the necessarily anecdotal assessments scholars make when they have to choose a small subset of texts to read to support their theses.[3] These trends may be topical, linguistic, or related to metadata such as dates or locations where a text is written. For instance, on Brigham Young University's site for text mining *Time Magazine* from 1923 to 2007, a search for "race relations" shows a quick rise to prominence of this issue in the 1950s and 1960s followed by a trough in the 1980s and a strong resurgence in the 1990s.[4] What might a student make of this pattern, especially the seesaw of the 1980s and 1990s that followed the period best known for the civil rights movement? Charts, graphs, 2- and 3-D visualizations of words, topics, events, people, and other entities are not interpretations or hypotheses by themselves, but they can aid the researcher who is "prospecting" for new information, insights, or interpretations to pursue.

Unsurprisingly, virtually all of the tools available to deal with the problem of abundance through *locating*, *extracting*, and *analyzing* have been produced by the same computer scientists who have pioneered text mining and clearly reflect that origin. Most of the tools require a high level of technical sophistication to use, and many do not have the kinds of interfaces that humanists find useful. Moreover, while the computer scientist often engages in research in text mining and analysis to improve upon algorithms or test computational theories, the humanities scholar has very different requirements and goals—establishing a new or better understanding of an era, event, person, or issue.

We thus currently find ourselves in the unsatisfactory position where much of the work in this area has been done on the "supply side," i.e. by computer scientists regularly producing new software to address their needs and by librarians and data providers digitizing vast archives and collections. Almost nothing has been done on

the "demand side"—to see what humanities scholars really need in text-mining and analysis tools and to design software and analytical techniques with those needs in mind.

How We Propose to Address the Problem

This proposal addresses those scholars and their needs in the age of abundance by finding out more about how digital resources are currently being used, making them easier to use in more sophisticated ways, and exemplifying some promising approaches to text mining and analysis. As historians, we will focus largely on the needs of historical research, but we believe that the outcomes of this project will be widely beneficial to scholars in a variety of humanities fields. Specifically, we will do this by:

1. providing a baseline of knowledge about how historians are using new digital resources (and what problems they are encountering);

2. developing a test bed of tools and resources that will allow ordinary historians with no technical training to experiment with text mining and analysis;

3. having a virtual working group of fifteen historians test these tools and resources and write "research diaries" about their experiences with them;

4. undertaking two extended case studies that will illuminate the possibilities (and limitations) of advanced text mining and analysis for historians;

5. disseminating the results of this research and demonstration work to the broader scholarly and digital library community.

Our more detailed work plan for accomplishing these goals is laid out in Section 4 below.

What Has Been Done So Far and What More Needs to Be Done

Historians, librarians, and other keepers of digital corpora who want to fulfill the promise of a great leap forward in research and scholarship in the digital era confront the problem that methods for searching and analyzing these new digital corpora seem frozen in time, limited to simple keyword search and failing to take advantage of the latest technology for scanning text, revealing patterns and related documents, and ranking search results. Moreover, most of the work on text mining and analysis remains confined to scientific and technical fields.[5] Virtually all such work in the humanities appears in literary studies journals such as *Literary & Linguistic Computing*, where the most common approach—close analyses of word usage and patterns in limited numbers of canonical and often highly structured texts—differs greatly from the questions we will address and which interest historians. Our questions will focus particularly on large bodies of primary and secondary sources that typically appear as uncorrected transcripts from optical character recognition (OCR)—the products of mass digitization that comprise the vast majority of texts online. Moreover, while most current text-mining tools focus on the "micro-analysis" favored by literary scholars, far less effort has been committed to helping a much bigger group of scholars with quandaries like the "White House email problem"—the problem of documentary abundance rather than that of closely analyzing a single or a few texts.[6]

Very recently, some leaders in the digital library field such as Clifford Lynch have issued calls for "open computational access" to large corpora of texts, arguing that it offers "truly stunning" opportunities that "point toward entirely new ways to think about the scholarly literature (and the underlying evidence that supports scholarship) as an active, computationally enabled representation of knowledge that lives, grows and interacts with its contributors." But, as Lynch also acknowledges, the potential benefits are still "largely

speculative and unproven" because we lack basic research.[7] To prove the case, we need research on how scholars will operate in the new digital environment provided by this higher level of access—how will our historian of the Clinton years do her work and reach her insights into the politics of the 1990s?[8]

Scholars today typically approach digital resources through a keyword search interface, but keyword search is unfortunately a relatively blunt instrument. For instance, in seconds a historian can find the 3,013 articles in the *New York Times* from the 1860s that include the phrase "Central Park." More dauntingly, the full run of the *Times* contains more than 333,000 articles that mention "Central Park," and many of these articles will discuss the park only in passing, or address subjects of little interest to a researcher looking for insight into a specific problem. At the same time, he or she would miss large numbers of relevant articles that make no explicit mention of Central Park—for example, articles about the use of carriages by wealthy New Yorkers in the 1860s, one of the prime ways that the park was used initially.

Researchers similarly have had access to only the simplest methods for analyzing the patterns within digital corpora, and again, these patterns have tended to focus on the frequency or usage of single keywords. For instance, a historian interested in the controversial question of the Founding Fathers' commitment to the Christian faith might count the number of "hits" on "Jesus" in the papers of George Washington, John Adams, and Thomas Jefferson. But she would miss more complex patterns of potential interest—for example, the co-occurrence of mentions of "God" in the context of discussions of death. Such patterns are the basic leads or evidence that historians and many other researchers in the humanities look for when building a thesis, yet this kind of prospecting is thwarted by the inadequate search tools provided by libraries and other holders of digital collections.

With the right tools, questions that were once addressed in a merely anecdotal fashion can be given a far more comprehensive treatment, perhaps even challenging established theories in a field. For instance, much of Victorian intellectual history focuses on a "crisis of faith," or the deep challenges to traditional Christianity such as the loss of the Bible as a central reference point among intellectuals. Scholars have pointed to major intellectuals such as John Stuart Mill and T. H. Huxley as exemplars of this crisis.[9] With the right tools, however, we can now search not just the texts of these leading lights but also the writings of thousands of other, lesser-known Victorians. If text-mining services could show every use of the Bible in every Victorian book, would the incidences in this much larger sample truly decline over the nineteenth century? Might other patterns appear that deserve further investigation, e.g., a rise in the use of the Old Testament versus the New?

Is there a way to map these interesting conjunctions of meaning and see where they lead? Can we do better than keyword search and simple word counts? Can we harness the power of vast new digital libraries to discover new knowledge—findings and understanding not possible with traditional methods? These are questions that librarians and researchers are increasingly asking.

Scattered attempts to allow scholars to perform computational research on digital corpora have shown promise. The National Center for Supercomputing Applications (NCSA) at the University of Illinois at Urbana-Champaign (UIUC) has produced Text to Knowledge (T2K), a complex Java package that provides a number of interchangeable text analysis modules (T2K has recently become part of the Software Environment for the Advancement of Scholarly Research {SEASR} project, funded by the Mellon Foundation). The MONK project (formerly NORA), also based in part at UIUC, has used T2K as well as several other text-mining tools to perform tests such as identifying words of affection in Emily Dickinson's poems. Open-access visualization tools like IBM's Many Eyes provide exciting new opportunities to create graphs and charts of the words in a text or texts. Another promising project is "MALLET: A Machine Learning for Language Toolkit," an integrated collection of Java code useful for statistical natural language processing, document classification, clustering, information extraction, and other machine learning applications, which the Computer Science Department at the University of Massachusetts, Amherst has developed.[10]

Yet however pioneering, these tools remain unlikely to become part of the landscape of most libraries' portals or included in most scholars' tool kits. First, they generally require a significant amount of technical knowledge to install and run. For instance, Philomine, the University of Chicago's text-mining extension to their full-text search tool Philologic, requires knowledge of the programming language Perl, command-line installation and configuration skills. Indeed, a majority of text-mining tools require knowledge of a specific programming language and an understanding of how to connect the program to existing collections of text. Second, even technically sophisticated users will find that each program has a different interface and preferred input type that often requires preprocessing texts (i.e., putting texts into a particular format before analysis can being). Third, these tools have not yet been tested in the field—with rare exceptions these tools are generally the products of software developers unaware of the needs of the humanist end user. In short, these tools require users to think like computer scientists, not like humanists. (One exception is the Canadian TAPoR project which has done a much better job than most other text-mining projects by linking texts to literary and linguistic tools—concordance makers, parts-of-speech taggers, and word frequency analyzers.)[11]

Even when these tools are more fully tested with scholars—and SEASR has some excellent plans for working directly with researchers—they will probably not satisfy the needs of all humanists. Historians, in particular, yearn for more expansive text-mining tools that look not at the turn of phrase in an Emily Dickinson poem or at Herman Melville's use of adjectives, but rather at patterns that may span hundreds, thousands, or even millions of documents written by different authors. They also often care more about locations, events, and topics than specific words or types of words. MONK's new online interface allows for incredibly fine-grained assessment of each word in a document and how much it indicates a certain feeling (such as affection)—a boon for literary scholars who might not be aware of hidden word usages by a poet—but a focus that may be too narrow for most other researchers.

Indeed, text-mining projects with a literary bent, like MONK or TAPoR, while important for certain scholars who focus closely on the use of language, often fail to address the broader needs of historians and others dealing with abundance, especially those at the early stages of research. Those non-literary researchers would prefer text analysis that significantly augments the four elementary user tasks identified in the library community's Functional Requirements for Bibliographic Records (FRBR)—that is, locating and acquiring entities that are relevant to their research.[12] In other words, before the "literary" form of text mining can occur, or indeed before any kind of research (digital or otherwise) can begin, most historians must find documents of interest on which they wish to perform these operations.

Historians have always engaged in the process of finding and assessing documents that may be of value to their research, but the possibilities for significantly enhancing that process through text mining and analysis and different user interfaces are not well understood outside of computer science and human-computer interaction labs. At the same time, the research techniques and work flows employed by humanists are foreign to those software developers typically charged with creating text-mining and analysis tools.

Nevertheless, we sense a widespread sentiment that a wall has been hit in the use of digital collections. In our conversations with non-technically savvy scholars as well as those at digital humanities and digital library centers, it has recently become clear to us that problems relating to digitization and access have largely been solved (though to be sure, there is much to be done to get more online), while the possibilities relating to discovery, search, and meaning extraction remain unrealized. Two conferences this year involving librarians and humanities scholars most active in text mining and analysis—one at Tufts University and another at the University of Chicago, jointly focused on "What Do You Do With a Million Books?"—both identified bringing text mining to the average collection and scholar as the critical goal of the next few years.[13]

Historians and the libraries that serve them—the central audiences for our research—are particularly interested in learning how they can leverage the power of, or better use, their newly digitized collections for scholarly research. As the just-released report of the American Council of Learned Societies on the emerging "cyberinfrastructure for scholarship" underlines, "Digital cultural heritage resources are a fundamental data-

set for the humanities: these resources, combined with computer networks and software tools, now shape the way that scholars discover and make sense of the human record, while also shaping the way their findings are communicated. . . . Even greater transformations are on the horizon."[14]

But to achieve those transformations, we need research into the process of digital scholarship, tools that make such scholarship possible, and demonstrations of the effectiveness of using text mining in serious scholarship. That is what we propose to do in this project.

2. Background of Applicant

The Center for History and New Media is well positioned to carry out this project. Since 1994, CHNM (http://chnm.gmu.edu/) has used digital media and computer technology to democratize history—to incorporate multiple voices, reach diverse audiences, and encourage popular participation in presenting and preserving the past. CHNM combines cutting-edge digital media with the latest and best historical scholarship to promote an inclusive understanding of the past as well as broad historical literacy. CHNM's work has been recognized with major awards and grants from the National Endowment for the Humanities (NEH), the Department of Education, the Library of Congress, the Institute of Museum and Library Services (IMLS), the National Historic Records and Publication Commission, and the Sloan, Mellon, Hewlett, Rockefeller, Gould, Delmas, and Kellogg foundations.

CHNM maintains more than four dozen online history projects directed at diverse audiences, making them available at no cost through its website. In 2006, CHNM's websites had 300 million hits and 10 million visitors—making it one of the busiest non-commercial history education sites on the Web. CHNM's staff of more than forty people (including eleven with doctoral degrees) includes those skilled in all aspects of digital history—database architecture, software engineering, web development and design, preservation, and archiving. CHNM has an annual budget of more than $1.5 million and a $2 million endowment (achieved with the assistance of a $500,000 Challenge Grant from NEH), which guarantees the long-term stability of CHNM.

CHNM has been a leader in using digital media to improve the understanding of history by students and the general public; in helping libraries and museums explore new and more effective ways to use digital materials and methods; and in building innovative, easy-to-use digital tools for accessing historical content. CHNM maintains several very popular and award-winning web-database projects, including educational projects such as *History Matters*, *Liberty, Equality, Fraternity: Exploring the French Revolution*, *World History Matters*, and *Historical Thinking Matters* as well as collections-based public history projects such as the *September 11 Digital Archive*, the *Hurricane Digital Memory Bank,* and *Gulag: Many Days, Many Lives*. CHNM is also the lead partner in four major Teaching Traditional American History grants from the Department of Education and a partner in two others. CHNM has won (or shared) the American Historical Association's James Harvey Robinson Prize three times for its "outstanding contribution to the teaching and learning of history."

CHNM also has extensive experience working with museums and libraries on digital projects. For example, CHNM worked closely with the Smithsonian Institution's National Museum of American History to provide an interactive "tell your story" component for NMAH's award-winning "September 11: Bearing Witness to History" exhibit. We have also worked closely with the Library of Congress, which accessioned the *September 11 Digital Archive* as its first major digital acquisition. In partnership with the University of Maryland, we have received a grant from the Library's National Digital Information Infrastructure and Preservation Program (NDIIPP) to preserve the digital history of dot-com boom.

Finally, CHNM has emerged as a leading developer of free tools and software for scholars, librarians, museum professionals, educators, and students. Through the *Echo: Exploring and Collection History Online*—

Science, Technology, and Industry project funded by the Sloan Foundation as well grants from NEH, IMLS, and the Mellon Foundation, CHNM's freely available tools have attracted hundreds of thousands of users. CHNM is also currently developing digital tools for teachers and students under a grant from NEH's EDSITEment program. Our most recent software release, an innovative, open-source reference management and note-taking called Zotero (http://www.zotero.org), was named among PC Magazine's "Best Free Software." It will also be receiving the American Political Science Association's "Best Software" award at the Association's 2007 meeting.

3. History and Background of Project

This project grows broadly out of the Center for History and New Media's extensive experience in building digital collections, developing innovative tools for operating in the digital environment, and introducing scholars, especially historians, to digital research.

CHNM has undertaken more than a dozen online projects that present historical documents for students, scholars, or the general public. Most of our digital archives—including *History Matters*, *World History Sources*, *Women and World History*, *Making the History of 1989: Sources and Narratives on the Fall of Communism*, and *Children and Youth in World History*—are aimed at the classroom. For example, *Liberty, Equality, Fraternity: Exploring the French Revolution* presents more three hundred textual documents, more than two hundred images, and dozens of songs and maps in an online, student-friendly archive. More than a half million unique visitors use the site every year.

But we have also worked on creating digital archives for a more scholarly audience. Our most ambitious of such projects is the *Papers of the War Department, 1784-1800*, which will present more than 50,000 documents from the most important agency in the Early American Republic. Other CHNM archival projects include the just-launched *Bracero History Project*, which will work collaboratively to document the guest-worker program and experience of the post-World War II years, and *Probing the Past*, an archive of probate inventories from the Chesapeake region of Maryland and Virginia for the period of 1740 to 1810.

CHNM has also worked extensively on gathering and archiving the history of the recent past through the creation of what we call "digital memory banks." Our largest project in this area has been the *September 11 Digital Archive*. In addition to collecting more than 150,000 digital objects, we encouraged people to write their own accounts of 9/11 and an astonishing 18,000 individuals responded. The *September 11 Digital Archive* has become the first major digital acquisition of the Library of Congress, and the World Trade Center Memorial Museum is likely to make it central to its exhibits. We have also provided occasions and motivations for a range of people—followers of information theorist Claude Shannon, people affected by the 1960s and 1970s New York blackouts, women scientists and engineers, and leaders of open source software efforts such as the Mozilla project that has led to the Firefox browser—to write and record their own historical accounts. Most recently, we developed an extensive site, the *Hurricane Digital Memory Bank*, to allow those affected by Hurricanes Katrina and Rita to documents their experiences.

All of these projects—both the teaching and scholarly archives as well as digital memory banks—have led us to think about and work on questions of text mining and analysis. How do we make our digital resources accessible and useful to the largest audience? What can you learn from these digital collections that you cannot as readily grasp from a paper archive? Thus, starting in 2003, we began to experiment with data mining the personal accounts in the September 11 collection and came up with some interesting preliminary results by mapping the locations of people who talked about "prayer" in their accounts of September 11th and comparing those who talked about George W. Bush with those who talked about the attacks in more personal terms.

CHNM's interest in developing easy-to-use tools for text mining and analysis also grows out of our extensive work in developing digital tools for historians over the past several years. Some of these tools have been directed at facilitating the teaching and research work of historians. For example, *Survey Builder* allows those without programming experience to collect history online. And the *Web Scrapbook* (or its later iteration as the EDSITEment teaching tools) makes it easy for teachers to create interactive online exercises. Some of our other tools efforts have directly involved us in problems of text mining and analysis. For example, the *Syllabus Finder* links a search algorithm with Google's application programming interface (API) to make it possible for searchers to find and data mine course syllabi on the web. Scholars have found this an extremely useful resource in planning their courses, and it has already been used over 700,000 times. A more experimental project, *H-Bot: The Automated Fact Finder*, uses text-mining and analysis algorithms to answer natural language queries such as "When did Charles Darwin write *The Origin of the Species*?" [15]

Our most extensive tools effort has taken us directly to the heart of the digital research process—one of the immediate points of entry for this project. In 2006, we released the first beta of *Zotero*, a free, open source, easy-to-use research tool that helps scholars gather and organize resources (whether bibliographies or full-text articles), and then annotate, organize, and share the results of their research. It includes the best parts of older reference manager software—the ability to store full reference information in author, title, and publication fields and to export it as formatted references—and the best parts of modern software such as del.icio.us or iTunes, such as the ability to sort, tag, and search in advanced ways. Using its unique ability to sense when one is viewing a book, article, or other resource on the web, Zotero will—on most major research sites—find and automatically save the full reference information in the correct fields of the researcher's personal database. By the spring of 2007, more than 150,000 users had downloaded Zotero. One of the motivations for the project proposed here is thinking about how we could build text-mining and analysis tools into what is emerging as the standard reference and note-taking package for serious researchers.

Finally, we have worked extensively with scholars and students seeking to understand how to do history in the new digital environment. These efforts have included a decade of offering graduate courses at George Mason University on history and new media, seven workshops on doing digital history for historians of science and technology funded by the Alfred P. Sloan Foundation, and the publication of the first book-length guide to digital history (*Digital History: A Guide to Gathering, Presenting, and Preserving the Past on the Web* by Dan Cohen and Roy Rosenzweig). These activities have impressed on us that for new approaches to be successful, they must avoid technical barriers and speak directly to the needs of working historians. There is little prospect that a typical historian will adopt a research tool that requires command-line programming.

These prior activities in building archives, developing tools, and working with historians have set the stage and determined the overall approach for this project. But we have also taken more specific steps that feed directly into this project. For example, we have experimented with n-gram analysis, a methodology that examines texts at an extremely granular level. We compiled an open-source proof-of-concept Java application capable of comparing individual documents against each other and against an entire corpus. By feeding this software journal articles from *French Historical Studies*, we located networks of related articles that ordinary keyword search could not identify. By employing the tool on our colleague Steven Weinberger's *Speech Accent Archive* (http://accent.gmu.edu), we generated a new mapping of English-speakers according to their relative accents that informs taxonomies of speech. (*For more on n-gram analysis, see Appendix B.*)

Equally key to adoption of text mining is the creation of partnerships and agreements that will allow us to field test our tools and to make them available to a broad audience of scholars. We began to talk with potential partners so we could do some pilot studies of the problems involved with text mining. Given how important JSTOR's archive is for the humanities, we approached them to see if we could get a text corpus to begin some initial tests using a variety of tools and we now have a formal agreement with them that allows such experimentation.

4. Methodology and Standards: How We Will Carry Out the Project

Four Key Problems and Four Strategies for Addressing Those Problems

Historians stand at the edge of a transformation in their research practices. Although the Internet (or its predecessors) has been around for almost four decades and personal computers are more than a quarter century old, historians have only been making serious use of digital resources in their research (as opposed to teaching) for less than a decade. And, indeed, many are only just now discovering the potential of these materials. Four factors characterize the current primitive state of the use of digital historical resources:

- Librarians and other data providers know very little about how scholars are actually using digital resources—for example, what they value about them and what problems they are encountering in making effective use of them—and have, therefore, done little to make their resources available in ways most amenable to serious researchers;

- The tools for serious text mining and analysis are either very primitive or only available to those with considerable technical background;

- Ordinary working historians are relatively uninformed about the potential of digital historical resources and are especially ignorant of more advanced methods of text mining and analysis;

- Even historians with a particular interest in digital materials or quantitative approaches have not done more than engage in preliminary explorations of the possibilities of text mining and analysis—indeed, they have few tools or resources for such explorations.

This research and development project responds to these four problems with a four-part plan of interlinked activities designed to jump-start the use of text mining and analysis by historians and to advance our very primitive state of understanding and resources. In particular, we propose the following activities (each of which is explained in more detail below):

1. a **survey** of current use of digital resources in historical research, which will provide a baseline of knowledge about how historians are using new digital resources (and what problems they are encountering) and inform our own activities as well as others working in this realm;

2. a **test bed** of tools and resources that will allow ordinary historians with no technical training to experiment with text-mining tools that assist with *locating, extracting, and analyzing;*

3. a year-long **working group** (both in person and virtual) in which fifteen historians learn about and apply these tools and resources to their own research and document their research processes for the benefit of tool builders, librarians, and other researchers;

4. two advanced **case studies** that will illuminate the possibilities (and limitations) of advanced text mining and analysis for historians.

Activity One: Survey Use of Digital Resources in Historical Research

The appearance of massive bodies of digital resources is transforming historical research. Yet, surprisingly, we know remarkably little about how this is happening specifically—e.g., what kinds of resources are most used, how those resources affect research findings and historical interpretations, and what barriers scholars encounter in using digital materials. There have been some works—notably, *Use and Users of Digital Resources:*

A Focus on Undergraduate Education in the Humanities and Social Sciences by Diane Harley, Jonathan Henke, Shannon Lawrence, et al.—that talk about the use of digital materials in teaching in the humanities. And there have been some excellent studies of what services libraries should be providing to scholars in the digital era such as the University of Minnesota Libraries report, *A Multi-Dimensional Framework for Academic Support* (2006).[16] But no one has talked extensively and systematically with historians about the impact on historical research of the availability of digital resources.

One limited and preliminary effort, which has lead directly to this project, was a pilot study of about fifty history graduate students carried out two years ago by Co-PI Roy Rosenzweig as part of his work as a member of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. That pilot study yielded rich material on emerging practices among graduate students, including the impressionistic finding that these resources are reshaping dissertation research more thoroughly and quickly than expected—for example, many graduate students described them as central rather than supplementary to their research work; yet, they also described frustrations with the current interfaces for using these collections. This convinced us that a more widespread survey effort would be valuable as a baseline for further work in text mining and to guide our efforts in this research and development effort. As a result, our work on the survey will take place largely in first six months of this project.

For the purposes of capturing emerging research practices, we do not think that we need a random sample of practicing historians, an approach that would add considerably to the expense of this portion of the project. By recruiting participants broadly rather than randomly, we will likely wind up with more younger scholars rather than very senior scholars, and more Americanists than scholars of other cultures—but we do not see this as posing any serious problem for drawing general conclusions from the resulting data. Every historical research project is to some extent unique yet also representative of larger issues, and the likely sample will probably represent the leading edge of those encountering the new digital corpora (and who may have the most to tell digital content providers and librarians). Our goals do not, however, include providing an estimate of the percentage of historians using digital materials, an effort that would require a random sample.

Nevertheless, we will actively recruit to achieve diversity across factors like age, gender, location, and differing methodological schools (e.g., intellectual, social, political historians). We will also collect limited demographic information (e.g., gender, date of receipt of Ph.D., area of specialization) and will examine how such factors affect views and practices. But because this is not a study of the historical profession as a whole but rather of those currently making use of digital resources, we do not need to make a particular effort to reach those who remain wedded to traditional sources and research methods. Instead, we will focus on encouraging wide participation. At the same time, this extensive but qualitative study will be useful in guiding the crafting of a smaller set of targeted questions that could be included on a national probability survey of historians, when and if such a survey were fielded. That is, this study might help in the creation of a digital content module that could be a part of a larger survey of historians in the future.

To attract a broad array of participants, we will publicize the survey extensively through: notices in professional newsletters from groups such as the American Historical Association and Organization of American Historians; postings to H-Net lists; advertisements and articles in *History News Network*, a CHNM-affiliate that reaches more than 13,000 historians three times a week with its newsletter and whose website receives about 3 million unique visitors per year; encouraging notices in the more than six hundred history blogs; and sending notices to other digital and print outlets.

As befits the topic, the survey will be web-based and will ask a limited number of open-ended questions. Before we begin widespread dissemination of the survey, we will conduct a preliminary test of the questionnaire to determine how well the questions capture what we intend them to capture, how long the survey takes to complete, and to locate any problematic questions or features of the survey instrument. The

precise list of questions and their wording will emerge out of this pre-testing, but in the pilot testing, we received useful responses from such questions as:

- Which online or digital resources have you used in your scholarly research and writing?

- Please give examples of how digital resources and tools are allowing you to do research (and to learn things) that you couldn't have done ten or twenty years ago.

- Are there barriers or problems you have encountered in the use of online research resources?

- Are there digital tools or forms of access that would enhance your ability to do historical research and writing?

CHNM has extensive background in managing online surveys through its work on creating digital memory banks, which have collectively gathered tens of thousands of responses from diverse audiences. In addition, Co-PI Roy Rosenzweig has substantial experience in organizing nationwide surveys through his work in the 1990s on an NEH-funded survey of more than 1,500 people on their views of historical resources and the place of history in their daily lives. That survey resulted in the prize-winning book (co-authored with David Thelen), *The Presence of the Past: Popular Uses of History in American Life*, which has been widely read and has also become the basis of similar surveys carried out in other countries, including Australia and Canada. Finally, we will be advised in this project on survey methodology by Scott Keeter, the director of survey research for the Pew Research Center in Washington, DC. Keeter also served as adviser to the *Presence of the Past* survey.

We will formally write up and disseminate the results of our survey in a white paper. But we will also be making use of the information we gather immediately, as we build the test bed of tools and plan the virtual working group.

Activity Two: Develop Test Bed of Tools

Some text-mining and analysis tools can reside on a personal computer, scanning or sorting a local collection assembled by a researcher. Other tools involve similar activities but operate on a much larger corpus, and must by necessity reside on a server. We plan to test a full range of both types of tools, including a comparative assessment of the strengths and weaknesses of these two main types and research into when it is best to use a server for processing and when local processing provides sufficient power and flexibility.

Types of Tools

Whether client- or server-based, our planned tools will fall into the three main categories: *locating*, *extracting*, and *analyzing*.

Locating Tools

a) *Relevance*. Relevance algorithms provide a "more like this" list of documents or texts using n-grams and other forms of computational analysis. An obvious use case is when a researcher finds a document that is useful for his or her research and wishes to see if there are similar documents to buttress a budding theory. While reading a journal article or archival document, a researcher will be able to find other documents "like" this one. Using extremely granular analysis, this tool will be able to locate relevant documents that keyword search would not identify.

b) *Summarizer*. A summarizer can help to create an abstract of a text that has none and allow researchers to quickly scan the contents of many texts during a discovery process. Autosummarization is still an immature

science and obviously no substitute for a human reader, but part of the reason to explore this possibility is to assess the value of machine summarization.

c) *Translation Services*. Like machine summarization, translation remains a computational field of limited (and occasionally comical) value to scholars in the humanities, but like the scholar who has just enough reading knowledge of a language to scan a table of contents to see if a text is worthy of further (perhaps labored) reading, machine translation may prove useful for prospecting.

d) *Complex Searching*. Finds texts with certain words within a certain distance of each other ("love" near "passion" in a set of diaries), texts with synonyms for a keyword (a particular failing of keyword search is that it doesn't account for the change in word choice over time), or cascading criteria (i.e. multiple criteria about elements in the text or metadata that can be put together in any sequence).

e) *Similarity Measurements and Ranking*: When searching for documents online, search results will rank results according to their relative similarity with each other, not just according to keyword frequency—today's typical norm in academic settings.

f) *Topic Clustering*: Through either search or browse interfaces, researchers will find documents "clustered" according to topical similarity, again not relying on keyword frequency but rather on granular textual analysis. Mimno's "Virtual Shelves" will serve as the model for this tool.

Extracting Tools

a) *Word Usage and Frequencies*: Researchers will be able to calculate the most frequent words used in a document under consideration, exempting language-specific "stop words," i.e. words that occur so frequently that they are never statistically significant and so are removed from the results to highlight truly important or unusual usages.

b) *Statistically Significant Phrases*: Building on the popularity of Amazon.com's "Statistically Improbable Phrases" feature, researchers will be able to view at a glance terminology indicative of a given document. These phrases often provide a surprisingly quick insight into the tone and orientation of a text. For instance, a text that uses "capitalist aggression" is likely of a different stripe than one that uses "communist propaganda."

c) *Named Entity Extractor*: By searching for capitalized phrases, our server-based tools will extract, rank, and even resolve named entities (personal, corporate, and institutional names; locations; dates) to their canonical identities.

Analyzing Tools

a) *Timeline Builder*. Plot the dates of chosen texts on a timeline, or events that have been extracted from a text or texts onto a timeline, possibly overlaid with other events and information. Timelines have long been used to order and map the proximity of people, events, and texts.

b) *Faceted Browsing*. Browse through a collection by "facets" or elements of a text, such as fiction vs. non-fiction, author, or country. Faceted browsing can rapidly cut down on the number of texts a historian may decide to study in depth, or, more powerfully, reveal patterns or anomalies such as a high number of manifestos written by artists in the decade before the first World War.

c) *Network Analysis and Mapping*: How are authors or figures in a text related to one another? What do patterns of communications, common themes, and locations tell us? A corpus containing the correspondence of hundreds of Victorian scientists might reveal the importance of a little-known figure who published little but

functioned as the nexus of information exchange for a particular field.

d) *Trends.* Takes the results of an extraction tool and provides a frequency analysis over time, place, or other vector. Brigham Young University's *Time Magazine* text-mining tool provides a good example here—the rise and fall of telling terms such as "race relations," "global warming," or "feminism."

Client-based Testing with Zotero

The text-mining tools that we will offer on the scholar's own machine we will be built on top of our Zotero research tool, which is an extension to the popular, open-source Firefox web browser (*see Appendix C for a fuller description of Zotero*). Zotero allows researchers to collect, manage, and cite materials they find online and offline. The software already has a large following, thanks in part to its easy-to-use interface that looks much like iTunes, which should help overcome the common problems with non-technical users approaching highly technical software.

Zotero was built from the start to facilitate the development of "utilities" to extend its functionality. An API for the client software already exists that enables any software developer to request metadata or entire texts to be extracted from a local Zotero collection. This extraction can be bundled up and sent using web protocols to web-based services (much like TaPOR does, connecting texts with analysis software that is on the web), or to a desktop application for local processing. A number of Zotero utilities are already in development by third parties, including utilities that synchronize Zotero metadata with social bookmarking and research sites such as Connotea and del.icio.us. The Zotero programming environment is both flexible and powerful: text-mining and analysis tools can be written in object-oriented Javascript and executed right within the Zotero environment; other programming languages like Java and C++ are also available. In short, Zotero provides an extensible and powerful foundation on top of which developers can place digital research tools that operate on personal collections of texts gathered by individual scholars. In addition, because Zotero is free, open source, and cross-platform compatible, the tool set will be available to all scholars and will not require any highly specialized installation.

Zotero's iTunes-like interface can be used by utilities to display or sort results of any processing. For instance, if a text analysis tool ranks the relevancy of each letter in a correspondence corpus to a research question or previously selected letter, the middle column of Zotero can arrange the results from top (most relevant letter) to bottom (least relevant letter). Should a larger canvas be necessary or should the results not accommodate themselves well to the Zotero interface, the main browser window can easily display the results instead. Graphs, maps, and other visualizations relationships could take advantage of this main, more expansive window. (*See Figure 1.*)

We will embed a suite of text-mining and analysis tools in a major Zotero utility that our testers will be able to download from the Zotero site and plug into their Zotero installations. The utility will layer itself on top of the standard Zotero software and add a "Digital Research" drop-down menu. Testers will first select a subset of their research items (or their entire research library) and then choose from this menu to apply a specific text-mining or analysis tool to the selections. The utility will then automatically combine, package, and transform this personal corpus and send it to web services or desktop applications or process it right within Zotero. For a large corpus or complex analysis, the package will likely be passed to a server for processing, while smaller text bundles or less complicated analyses will occur locally.

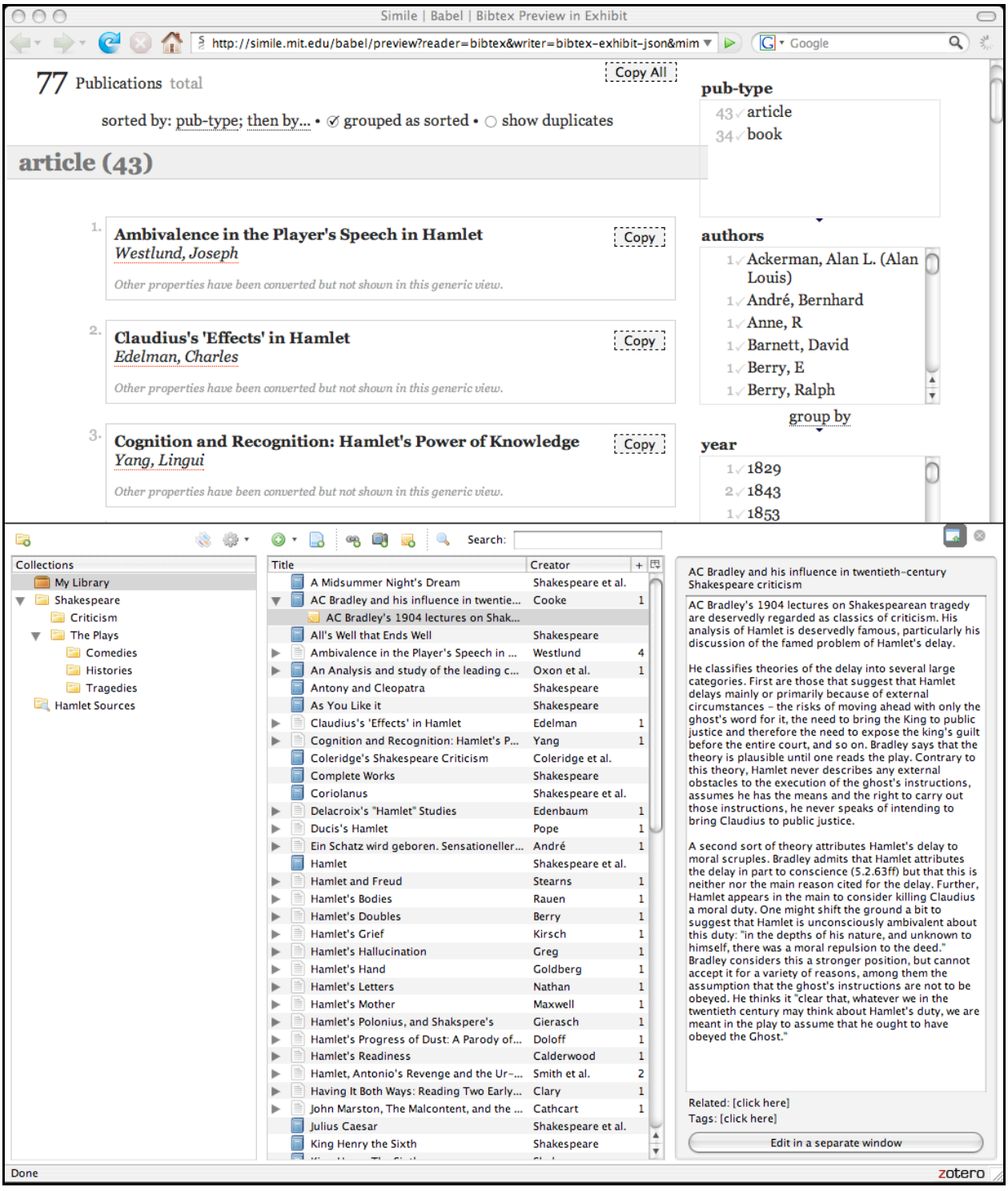

Figure 1. A screenshot of Zotero running in the Firefox web browser in a prototype text-mining scenario. In the bottom half of the window is Zotero's iTunes-like interface showing a personal collection compiled by a researcher. In the top half of the window is Firefox's regular display region, in this case showing MIT's Exhibit faceted-browsing tool processing the local Zotero collection.

Server-based Testing

While the Zotero text-mining and analysis utility will focus on processing the unique collections of individual users, a researcher also needs to find other, new relevant material or to discover key sections or patterns in an entire collection, such as an online archive or literature database. Furthermore, many digital services for researchers must reside and function on servers, since these tools will usually require the preprocessing of much larger collections than would normally be stored in a Zotero library.

For instance, a researcher examining the changes in historiography in the twentieth century might identify JSTOR's comprehensive collection of historical journals, including hundreds of thousands of articles, as a valuable corpus to probe for evidence of change over time. Yet researchers could hardly download the entire corpus, and even if they could, it is highly unlikely that they would know how to process the collection to locate these patterns, nor would their personal computer have the storage or processing power to run these analyses.

In building these server-based tools and resources, we will exploit our extensive experience building, searching, and displaying large collections, such as the *September 11 Digital Archive*. Our server-based tools will enable researchers to locate, extract, and analyze corpora of documents too unwieldy for desktop machines. Moreover, our server-based tools will be able to incorporate visualization and mapping technology too resource-intensive for the typical client machine.

Using the Center for History and New Media's experience with mounting extensive collections on the web as well as our initial forays into text processing, we will build a set of test bed websites that will provide interfaces for a series of server tools in the three main areas of *locating*, *extracting*, and *analyzing*. To lay the groundwork for this research, the Center has already made agreements with JSTOR, the Alexander Street Press, the University of Virginia Library, the Project for American and French Research on the Treasury of the French Language (ARTFL), the Internet Archive, and other collections to explore these kinds of research questions.

These two main types of tools, client- and server-based, are of course not mutually exclusive. Since Zotero lives in the web browser it can call upon web-based tools for some or all of its text-mining processing power. Obviously for large online collections such as JSTOR, locating tools must reside on the server and be accessed through web interfaces. Large-scale pattern assessment and mapping will also require server-based tools. However, if a researcher has assembled a subset of interesting documents, extracting entities and analyzing patterns within that collection could occur on the scholar's personal computer. The line between potential client- and server-based text-mining and analysis tools thus is not entirely fixed, and nor should it be: there will be cases when it makes better sense to harness the computational horsepower of servers for reasons of scale or simply because a researcher is looking for resources that she does not yet have locally on her own computer. In other cases, a client-based solution may better serve a researcher because she is working with a personal collection of documents easily analyzed through a local set of tools. In the course of a research project, it is easy to imagine a researcher moving back and forth between client- and server-based text-mining and analysis tools, and one of our research objectives is in fact to identify optimal relationships between the two types.

Activity Three: Organize Virtual Working Group

Even before these resources are fully constructed, we will begin to assemble a team of historians to experiment with and evaluate different tools and resources. Our primary strategy for this will be a virtual "working group" on digital historical research. It will consist of fifteen scholars who will participate in an in-person workshop in which we will explain the use of the text-mining and analysis tools and also brainstorm

about possible strategies for using these tools in research projects. In return for a modest honorarium these scholars will keep a "research diary" describing their research process and especially their use of the text-mining and analysis tools; they will also participate in an online forum. We will make an effort in recruiting the members of the virtual working group to ensure diversity of methodological approaches. The data gathered from the online forums and the research diaries will allow us to research the comparative utility of different text-mining and analysis tools.

The working group will begin with a two-day in-person workshop in September 2008 at the Center for History and New Media at George Mason University, where we have good facilities for such gatherings. The first day will focus on introducing the tools and resources assembled in the test bed. The participants will get a chance to experiment directly with the tools. We will also discuss examples of text-mining and analysis projects in other disciplines and whether these approaches are useful for historians. The second day will focus on the projects being developed by the fifteen scholars in the workshop. Each will introduce the historical project they are undertaking and then we will collectively brainstorm about how text mining and analysis can be applied to it.

During the following year, the scholars will report at least once every month on their progress and problems. They will also be required to comment on the projects being carried out by other participants. Each participant will be given a blog to post comments to, and all of the blogs' feeds will be aggregated via RSS to the project's main website. In addition, a private wiki will be set up to collaboratively build a users manual and knowledge base for the tools and to describe the possibilities and limits of each tool. These ongoing research diaries, commentaries, and knowledge base will provide the basis for our assessment of the comparative value of different text-mining and analysis tools as well as of the broader possibilities of these approaches for historians.

Activity Four: Develop Two Detailed Case Studies

Philosophes and the Encyclopedia

Since the 1980s, the Project for American and French Research on the Treasury of the French Language (ARTFL) has digitized thousands of French texts of historical and literary interest, the most ambitious of which by far is Diderot and d'Alembert's great Enlightenment endeavor, the Encyclopedia. Containing some 72,000 articles written by over 140 contributors, the Encyclopedia claimed to be nothing less than a comprehensive collection of human knowledge.

Ordering this massive compendium of knowledge was at least as important as its content, and the Encyclopedia employed several finding strategies: it suggested that all human knowledge fell into the three decidedly secular categories of memory, reason, and imagination; readers could locate individual articles alphabetically; articles on any given topic directed readers to other articles through a system of cross references. All these strategies were intended to democratize and decentralize knowledge.

Ultimately the ambition of the Encyclopedia outran the capacity of its authors and editors: its proposed taxonomy of knowledge largely vanished into a sea of content; the practical considerations of publishing a multi-volume set over the course of two decades strained alphabetical ordering; phantom cross references frequently led readers to dead ends.

Diderot, d'Alembert, and their fellow Encyclopedists thus faced much the same problem that digital researchers and content providers confront today: the abundance of content can ultimately serve to obscure without appropriate tools to find and order. By applying text-mining and analysis tools to the Encyclopedia, we seek to achieve three goals. First, we will create a new web-based finding aid that will provide a flexible system of ordering that will allow readers to identify new webs of knowledge within and across topics and

articles, an objective far beyond the technical means of the eighteenth century. Second, we will use this tool to evaluate the scholarship that has investigated the Encyclopedia's successes and limitations in its effort to collect and order human knowledge. Third, we will make far more accessible the deep knowledge embedded in the Encyclopedia to researchers interested in the cultural, social, political, intellectual, and economic history of the Enlightenment.

We have recently enabled Zotero to allow researchers to save articles from Diderot and d'Alembert's Encyclopedia: the tool saves all relevant publication data, resolves author names using an authoritative index, and automatically generates subject tags based on the Encyclopedia's original taxonomy. The tool also prepares a full-text index of the saved text and preserves a "snapshot" of the original article that researchers can digitally annotate. With this case study, we will vastly expand on the current functionality: researchers will be able to locate relevant articles based on their research libraries; trace networks of related information; visualize not only word frequencies but topical frequency. By comparing a given article's suggested cross references with a network of related documents located by text mining, for example, a researcher might identify surprising slippage between conceptual relevance and editorial intent.

This case study will result in two articles. The first will be a review article on Enlightenment taxonomies of knowledge, evaluating how text-mining and analysis tools sustain or provide new insight into existing understanding of how knowledge was ordered in the eighteenth century. The second piece, to be published on our site, will explain how scholars of the Enlightenment can utilize our finding aid in their research, as well as apply our methodology to other relevant materials.

Project Director Sean Takats will lead this case study.

Victorians and the Bible

The Open Content Alliance (OCA), a consortium of the Internet Archive, Microsoft, Yahoo!, several large libraries, and other major companies and institutions, is in the process of building a combined collection of public domain texts. The collection is particularly strong in Victorian texts, currently holding 50,000 books and growing rapidly. Unlike the books in the Google Book Search project, these books are available not only as page images but also as OCRed text. In short, for the first time it is possible to scan the full text of thousands of works from the Victorian age.

Scholars of Victorian history have commonly held notions about the era they study that could be examined with new eyes with the proper text-mining and analysis tools. For instance, much of Victorian intellectual history is based on the "Secularization Thesis," or the sense that the age witnessed a decline in religiosity and especially a decline in organized religion. Anecdotally, scholars have noted fewer uses of the Bible as the nineteenth century wore on.

We will scan the entire Victorian corpus from OCA to extract references to the Bible. We will then build a website that scholars can use to trace biblical usage and see if this is helpful in revealing new evidence for existing theories or instead challenges those theories. What does the overall usage pattern of the Bible in the nineteenth century look like? Were some parts brought to the fore (e.g., Revelations, the Sermon on the Mount) while others moved into obscurity? What do such peaks and valleys say about the concerns of authors who quoted the Bible? Which English translations, with their significant differences and biases, were used throughout the Victorian era?

This case study should provide a helpful foundation for other, similar projects that scholars in history and the humanities might wish to pursue. For instance, using the same techniques (and indeed, much of the same technology), a literature scholar could trace the usage of Yeats's poem "The Second Coming" from its origin in 1919 to the present day. The case study will also likely encounter common problems experienced in this kind of text mining. Many quotations from the Bible (as well as most other texts) are much briefer than an

entire verse; some quotations may be only five or six words since authors assume a knowledgeable reader will pick up canonical references. N-gram analysis is perfect for finding passages of this small scale and thus could create a concordance that is far more comprehensive than anything in existence. Moreover, analyzing the use of the Bible will require us not only to extract passages but also shorthand for those passages, such as standard biblical notation.

The case study on the Victorians and the Bible will result in three articles. The first will be a review article in a scholarly historical journal on the "Victorian crisis of faith" that will look at the most important books and articles on that subject to see if the new techniques and data from the case study support or challenge each work's thesis. The second article, to be published in a broadly available journal (such as the *Chronicle of Higher Education*), will describe in a plainspoken way the methodology of the case study in a way that is comprehensible to humanities scholars with little or no technical knowledge. A third article, to be published on our website, will explain in greater technical detail how to reproduce the site for other case studies.

Co-Principal Investigator Dan Cohen will lead this case study.

Dissemination

The results of this research and demonstration project will be disseminated in multiple forms, as outlined below.

Evaluation

We will involve our distinguished group of consultants *(see Appendix D)*, who will also constitute the project advisory board, in both formative and summative evaluation. At the start of the project, we will convene the advisory board (electronically) to discuss the project work plan as a whole. The advisory board will help us to design the survey, to finalize an exact list of tools to be created for both our server and Zotero tests, and to plan and recruit the virtual working group. We will also draw on their expertise to define benchmarks for successful tools, such as the ability to reveal new information, reduce searching time, and enable flexible interaction with a collection.

At the same time, we will also draw on different individuals to provide more specialized advice on different aspects of the project. For example, survey specialist Scott Keeter will advise us on the questions and methodology for the overall survey. Digital library specialists (Dan Chudnov, Mark Olsen, and Josh Greenberg), who are currently exploring whether and how to add text-mining tools to library servers and user interfaces, will advise on how to craft the tools to be most easily implemented by digital libraries and optimally useful for their patrons. Greg Crane, William Turkel, Mark Olsen, and Michael O'Malley have experience creating digital tools and will help us in the prototyping and programming phase. Then, at the end of first nine months of the project—when the tools test bed and the initial survey are complete—we will convene another electronic meeting of the advisory board to evaluate the tools and the results of the survey.

We will convene a final meeting in the last month of the project to provide an overall evaluation. We will again evaluate our progress against the benchmarks to see how well we met the original goals and to understand which tools could have been better designed and how. We will also ask the advisory board to make recommendations for next steps in improving and enhancing digital historical research.

5. Work Plan

Phase I: Project Planning, Survey Research and Development of Test Bed (1/1/08-9/1/08).

January 2008
- Management: Project Launch Meeting with key staff, consultants, and advisors focusing on the survey and tools test bed design
- Survey: Design and pre-test survey with help of Scott Keeter

February 2008
- Tools Test Bed: Develop design for Zotero extension
- Tools Test Bed: Begin building Zotero extension
- Tools Test Bed: Perform preliminary survey of existing server-side tools in consultation with Dan Chudnov and Josh Greenberg
- Survey: Recruit participants and perform ongoing publicity to attract respondents

March 2008
- Survey: Continue to recruit participants and perform ongoing publicity to attract respondents
- Tools Test Bed: Preliminary testing Zotero extension with focus group
- Tools Test Bed: Recruit Zotero extension testers
- Tools Test Bed: Revise Zotero extension based on focus group testing
- Tools Test Bed: Customize programming/plugins for server-side tools

April 2008
- Survey: Preliminary analysis of existing data
- Tools Test Bed: Launch Zotero extension
- Tools Test Bed: Begin ongoing publicity for Zotero extension
- Tools Test Bed: Continue customizing programming/plugins for server-side tools

May 2008
- Tools Test Bed: Begin designing interface for making server-side tools available to researchers through single portal

June 2008
- Survey: Close survey
- Survey: Conduct data analysis
- Tools Test Bed: Continue designing interface for making server-side tools available to researchers through single portal

July 2008
- Survey: Draft summary survey results
- Tools Test Bed: Test interface for server-side portal with focus group
- Planning: Recruit 15 scholars to participate in Virtual Working Group

August 2008
- Survey: Revise summary of survey results
- Tools Test Bed: Revise server-side portal interface based on testing
- Planning: Arrange site and logistics for Virtual Working Group Workshop

September 2008
- Evaluation: Electronic Meeting of the Advisory Board and Consultants to review the results of the Survey and the Tools Test Bed
- Survey: CHNM publishes summary survey results on the CHNM website
- Tools Test Bed: Launch server-side portal
- Tools Test Bed: Begin on-going publicity server-side portal

Phase II: Virtual Working Group and Case Studies (9/1/08-9/1/09)

September 2008
- Virtual Working Group: Conduct 2-Day Workshop at GMU for the Working Group
- Virtual Working Group: CHNM builds portal for Virtual Working Group interaction

October 2008
- Virtual Working Group: Participant historians submit research diary detailing their progress on their text-mining project
- Case Studies: Plan Case Studies

November 2008
- Virtual Working Group: Participant historians submit research diary detailing their progress on their text-mining project
- Virtual Working Group: Participant Historians comment on their peers' research diaries
- Case Studies: Begin Research

December 2008
- Virtual Working Group: Participant historians submit research diary detailing their progress on their text-mining project
- Case Studies: Continue Case Studies

January 2009
- Virtual Working Group: Participant historians submit research diary detailing their progress on their text-mining project
- Virtual Working Group: Participant Historians comment on their peers' research diaries
- Case Studies: Continue Case Studies

February 2009
- Virtual Working Group: Participant historians submit research diary detailing their progress on their text-mining project
- Case Studies: Continue Case Studies

March 2009
- Virtual Working Group: Participant historians submit research diary detailing their progress on their text-mining project
- Virtual Working Group: Participant Historians comment on their peers' research diaries
- Case Studies: Continue Case Studies

April 2009
- Virtual Working Group: Participant historians submit research diary detailing their progress on their text-mining project
- Case Studies: Continue Case Studies

May 2009
- Virtual Working Group: Participant historians submit research diary detailing their progress on their text-mining project
- Virtual Working Group: Participant Historians comment on their peers' research diaries
- Case Studies: Begin Writing up Results

June 2009
- Virtual Working Group: Participant historians submit research diary detailing their progress on their text-mining project
- Case Studies: Continue Writing up Results

July 2009
- Virtual Working Group: Participant historians submit research diary detailing their progress on their text-mining project
- Virtual Working Group: Participant Historians comment on their peers' research diaries
- Case Studies: Finish Writing up Results

August 2009
- Virtual Working Group: Participant historians submit research diary detailing their progress on their text-mining project

September 2009
- Virtual Working Group: Participant Historians comment on their peers' research diaries

Phase III: Working Paper, Dissemination, and Evaluation (7/1/09-12/31/09):

July 2009
- Working Paper: CHNM Staff begin reviewing the user data from Tools Test Bed

August 2009
- Working Paper: CHNM Staff drafts white paper on Tools Test Bed usage

September 2009
- Dissemination: CHNM publishes a white paper on the CHNM website on Tools Test Bed usage

October 2009
- Dissemination: CHNM publishes article for scholars of the Enlightenment based on the findings of the first case study
- Dissemination: CHNM publishes an article on the CHNM website about how scholars of the Enlightenment can use the finding aid from the first case study in their own research

November 2009
- Dissemination: CHNM publishes review article on the "Victorian crisis of faith" in a peer-reviewed journal
- Dissemination: CHNM publishes technical article on the CHNM website on the methodology of the case study on the Victorians and the Bible
- Dissemination: CHNM publishes an article in the *Chronicle of Higher Education* (or similar publication) on the methodology of the second case study on Victorians and the Bible
- Evaluation: CHNM staff draft summative white paper

December 2009
- Evaluation: Electronic Meeting of Advisory Board and Consultants to review the case studies, the conclusions drawn from the work of the Virtual Working Group, and the summative white paper.
- Dissemination: CHNM publishes the summative white paper on the CHNM website

6. Staff and Management Responsibilities

Key Staff and Responsibilities

Co-executive Producers (Primary Investigators): Dan Cohen and Roy Rosenzweig have overall responsibility for program vision, budget and grant management, and long-term sustainability. Cohen will take primary responsibility for overseeing work related to the development of the client- and server-based tools, the detailed case studies, and the dissemination, evaluation, and work of the advisory board. Rosenzweig will have primary responsibility for overseeing the survey and the virtual working group. (*See bios and CVs in Appendix D.*)

Project Director and Manager: Sean Takats will coordinate day-to-day personnel management; serve as lead software development architect; coordinate with digital libraries, data providers, virtual working group, and consultants; plan the organization of research and the analysis and dissemination of research results.

Consultants/Advisory Board: CHNM will be assisted by a distinguished group of consultants drawn from the worlds of libraries (Dan Chudnov, Library of Congress, and Josh Greenberg, New York Public Library); digital humanities (Greg Crane, Tufts University, Mark Olsen, University of Chicago, and William Turkel, University of Western Ontario); and historical scholarship (Michael O'Malley, George Mason History Department, and Olsen), and survey research (Scott Keeter, Pew Research Center).

Programmer: CHNM will hire a contract programmer to work on the development of the Zotero utilities and the web-based resources. Our previous extensive experience in software development projects suggests that this is the most cost-effective strategy for securing high quality programming assistance. Cohen and Takats, who have extensive programming experience, will work closely with the contract programmer.

Web Developer/Designer: Experience has taught us that a clean and intuitive interface is vital to encouraging participation by non-technically oriented historians. CHNM Web developer Jeremy Boggs will have responsibility for building the interface for the Zotero-based utility and the server-side tools.

Project Assistant: The project assistant—likely a graduate student drawn from Mason's doctoral history program—will assist the Project Director in all phases of the work—including data collection, recruitment of scholars, data analysis, and dissemination.

Organization of Work

CHNM's extensive experience in digital projects equips us to manage the planning, development and dissemination of this project. The project builds directly on three core strengths of CHNM: our work in developing widely used digital tools, including the open source Zotero; and our combined expertise in digital libraries, digital tools development, and traditional historical research; and our extensive experience working with historical scholars without technical expertise. In other words, we are ourselves scholars and understand the core research questions that are being investigated here.

The day-to-day work of the project will be overseen by Project Director Sean Takats and will proceed by way of weekly meetings and ongoing collaboration via email, blogs, instant messaging, and conference calls. Here,

Takats will be particularly assisted by Co-PIs Cohen and Rosenzweig (who, in fact, sit in offices adjoining Takats in CHNM's office suite in the Mason Research Building). The project assistant will provide support to Takats in day-to-day project tasks, including publicizing the survey, recruiting for the working group, setting up advisory committee meeting, and disseminating project findings. Open communication will ensure that production benefits from the advisory committee's expertise, and that we meet all of our project deadlines. This system will result in project staffing that makes the best use of personnel expertise, while emphasizing communication and flexibility at every stage of the project.

While Takats will take overall management responsibility, each member of the leadership group, as noted above, will have particular areas of responsibility. For example, Cohen, who has an extensive programming background and has taken the lead on all of CHNM's digital tools projects, including Zotero, will have primary oversight of the tools development. Rosenzweig, who has extensive survey experience and is very well connected in the historical community, will oversee the survey and the set up of the virtual workshop. Takats and Cohen will carry out the two detailed case studies, which fall within areas in which they have extensive historical training and have published widely. Cohen and Rosenzweig, who are well known in digital history and digital humanities, will oversee dissemination of the results of our research to scholars, digital librarians, and digital humanists. One crucial feature of this project is that two of the leading figures—Cohen and Takats—have extensive experience and credentials in both technical and historical work. Takats, for example, has years of professional experience in technical consulting and web application development. We have found that this combination of technical skill and advanced historical training is essential to a project like this one. We don't want our users to need technical expertise to explore new dimensions of historical research, but our team needs to be well grounded in the possibilities and problems of new technology.

7. Dissemination

We will disseminate the results of this project in multiple formats, including an overall white paper on the entire project, conference presentations, blog postings, and journal and newsletter articles. The white paper will report on the state of digital research in history and the possibilities for text mining and analysis. It will offer not "best practices," but "promising practices"—advice to librarians on leveraging the digital resources now available for the benefit of research in history and the humanities. Because the white paper will not be completed until the end of the grant period, we will also present ongoing reports on the research results through blog postings and conference presentations. For example, the first piece of the research to be completed will be the survey of digital research practices, and we will provide the results of that as soon as the data is analyzed. In addition, the results of the working group and the two case studies will be embodied in works of historical scholarship, which will be published in standard historical journals as well as online. Finally, the results of our research will embodied in more than just written documents. They will also be available online in the form of digital tools and digital corpora that researchers will be able to use.

The Center for History and New Media is well positioned to disseminate the results of this project to its two key audiences—scholars, especially historians, and digital librarians—because of its extensive track record of work, its reputation, and its connections in these communities. Leaders at CHNM have regularly presented at conferences attended by historians and by those involved in digital libraries—including the Coalition for Networked Information (where Co-PI Roy Rosenzweig sits on the steering committee), the WebWise Conference on Libraries and Museums in the Digital World, the Digital Library Federation, and the Association for Research Libraries —and we will use those forums for disseminating the results of this project. CHNM also sponsors a number of blogs that are well read in the digital library community, including the blog of Dan Cohen, which has over 600 subscribers and is syndicated on the Code4lib library tech website. We will set up a blog specifically focused on this project, as we have done for Zotero. We will also publish more sustained commentaries in publications directed at the research and library community (e.g., *D-Lib*, *First Monday*, and the *Chronicle of Higher Education*), as we have already done with other CHNM projects.[17] Another key link to the library community will be through Josh Greenberg, a longtime CHNM

staff member who has just become the Director of Digital Strategy and Scholarship at the New York Public Library. Greenberg will serve as a key adviser and liaison to the digital library community.

CHNM also has deep connections with the scholarly community in history, the other key audience for this research. Project Co-PIs Cohen and Rosenzweig are co-authors of the standard work on digital history, and CHNM has had a longstanding involvement in scholarly associations in history. Rosenzweig, for example, headed the Electronic Advisory Committee of the Organization of American Historians for many years and was until recently Research Vice-President of the American Historical Association (AHA). CHNM has won the AHA's James Harvey Robinson Prize an unprecedented three times for its digital projects. And, as an affiliate of the AHA, CHNM has a regular slot on the AHA program. Finally, CHNM affiliate *History News Network* has a three-times per week newsletter that reaches 13,000 historians. Our own website receives more than 10 million unique visitors (accounting for 300 million hits) per year. We will use all of these communication channels—conference presentations, blogs, scholarly articles, our website, *HNN*—to disseminate the results of this research project to librarians, data providers, and scholars.

CHNM's ability to sustain the digital tools and resources that are part of this project are demonstrated by more than a dozen years of work in the field of digital history. In addition to its dozens of internally developed projects (all of which are still maintained), the creators of major digital history projects—such as the award-winning *DoHistory*—have given their sites to CHNM to ensure a permanent home. CHNM's stability stems from several sources. With the help of an NEH Challenge Grant, CHNM has raised a $2 million endowment, providing for CHNM's long-term viability as a unit of the George Mason University history department. The history department in turn has made "new media" a central focus of its new Ph.D. program, ensuring a steady stream of graduate students—more than ten of whom currently work at CHNM—interested in sustaining existing projects and creating new ones. Mason also has a strong institutional commitment to CHNM, which has a $1.5 million annual budget and a staff of more than forty. CHNM has also recently received an Academic Excellence Equipment Award from Sun Microsystems, which has allowed us to build a robust server that is housed in Mason's secure data facility. Equally important, the text-mining and analysis testing platform developed here will be linked to our Zotero project, which has already raised more than $1.3 million in external funding, and will have its own state-of-the art server. Zotero is an open source project with its code and documentation exposed and available on the Zotero website. We will fold our proposed client-based text-mining utilities documentation into this existing structure. Zotero has already build a robust open source community that is supporting the product. This substantial endowment, demonstrated fundraising success, strong institutional support, significant open source experience, and solid technology base allow CHNM to guarantee that our proposed research (and research tools) will provide an ongoing basis for continuing research into the future of text mining and analysis in historical scholarship.