

Penn Libraries

Scholarship at Penn Libraries

University of Pennsylvania

Year 2007

Digital Scholarship and
Cyberinfrastructure in the Humanities:
Lessons from the Text Creation
Partnership

Shawn Martin
University of Pennsylvania, shawnmar@upenn.edu

Postprint version. Published in *Journal of Electronic Publishing*, Volume 10, Issue 1, January 2007.

Publisher URL: <http://hdl.handle.net/2027/spo.3336451.0010.105>

This paper is posted at ScholarlyCommons.

http://repository.upenn.edu/library_papers/54

Digital Scholarship and Cyberinfrastructure in the Humanities: Lessons from the Text Creation Partnership

Shawn Martin

Abstract

Electronic technology has changed the way scholars in the humanities do their work, creating two distinct groups of scholars: first, those who perform leading-edge humanities computing research (a relatively small number); and second, scholars who perform traditional humanities research with new electronic tools (a fairly large number). How is it possible to bring these two groups together? The Text Creation Partnership at the University of Michigan provides one way of providing services to both. And as the electronic publishing community looks for ways to provide reliable cyberinfrastructure in the humanities, the Text Creation Partnership provides a model for building large digital collections that meet the needs of future scholars.

Few would disagree with the fact that electronic technology has changed the way the academic community does its work. Collections like Early English Books Online (EEBO) from ProQuest, Evans Early American Imprints (Evans) from Newsbank-Readex, and Eighteenth Century Collections Online (ECCO) from Gale have allowed scholars and students to perform research in a matter of seconds that previously would have taken a lifetime. According to Mark Sandler, former Collection Development Officer at the University of Michigan, “the widespread, electronic dissemination of hundreds of thousands of rare books, and their easy availability for undergraduate students, represents a revolution of sorts in higher education. . . . For humanists, these collections represent an intellectual analog to the role played by the cost-effective Model T in unleashing a culture of ubiquitous automotive transportation.” [\[1\]](#)

This revolution has manifested itself in a variety of scholarly projects, including, among other, things such as sociolinguistic mapping of ideas in different geographical locations and the creation of interactive learning environments, such work is pushing the boundaries of scholarship and teaching. Though this revolution may have already occurred for many in academe, there are in fact two revolutions going on. The first is being led by a small number of dedicated academics, librarians, and technologists at a handful of institutions. Edward Ayers at the University of Virginia has called this “a revolution led from above.” [\[2\]](#) The second and much slower revolution involves the rest of the scholars, graduate students, undergraduates, librarians, and others who use electronic technology for the “traditional” scholarship they have been practicing for many years. One bridge between these two groups is the Text Creation Partnership (TCP) at the University of Michigan. The TCP, a joint project between librarians, scholars, and publishers at over 150 participating institutions around the world, lets both digital revolutionaries and traditional scholars participate in digital humanities, which, roughly defined, is an “interdisciplinary core. . . illustrated by examination of the locations at which specific disciplinary practices intersect with computation.” [\[3\]](#)

The recent report of the American Council of Learned Societies (ACLS) Commission on Cyberinfrastructure called digital scholarship “the inevitable future of the humanities and social sciences,” and said that, “digital literacy is a matter of national competitiveness

and a mission that needs to be embraced by universities, libraries, museums, and archives.” [4] Yet relatively few humanities scholars are relying heavily on electronic tools and methodologies. In fact, some scholars even seem hostile toward electronic resources. [5] So the Commission’s assessment would therefore seem like little more than a pipe dream. Thus the question remains, how do the library and academic communities create an infrastructure that makes digital scholarship an “inevitable” (or at least a prospective) future? Though the TCP may not be the model for cyberinfrastructure, it provides an instructive example of how the electronic-publishing community can achieve the larger goals of electronic resource creation and mass digitization.

Background

The TCP began in 1998 when ProQuest Information and Learning developed Early English Books Online, a database containing digital images of nearly every book printed in English or in England between 1470 and 1700. Though it had obvious advantages over microfilm, EEBO only duplicated what Michigan and many other libraries had in their collections. The real power of EEBO was not in searching catalog records in the computer (which researchers could already do to some degree); rather, it was to allow full-text keyword searching of those titles. Librarians at the University of Michigan and Oxford University believed that if scholars and students could search for references to Aristotle, mentions of the Great Fire of London, or even curse words, entirely new avenues of scholarship could be opened. The cost of turning the pictures of the pages into words in a database was, however, prohibitive. ProQuest could not run optical character recognition (OCR) software on the collection because the early Gothic fonts were beyond OCR’s capabilities, and to have the entire collection hand transcribed would cost more than they felt institutions would be willing to pay. So they decided to sell the collection as it was: a set of pictures of pages of books.

Librarians at Michigan and Oxford believed that they could get support to capture full electronic text for at least a portion of the EEBO corpus. Thus the TCP was born. Mark Sandler, the original visionary behind the project, proposed that the two institutions could provide the infrastructure to fully tag about 25,000 texts. To fund the project, libraries that already subscribed to EEBO provided money for conversion, which ProQuest (which was to own the system) then matched. The more money the project was able to bring in, the more text TCP could create, and the cheaper each volume became for the libraries involved. Furthermore, every volume that TCP converted would enter the public domain five years after TCP finished producing them. Therefore these works would not be locked up in their digital forms by commercial companies. Much of the English literary and historical record would eventually be freely available to all universities and to the general public, but at the same time the investment of the scholarly publisher (in this case ProQuest), would be protected by allowing them time to sell the EEBO product with text for a period of five years. [6] This model has now been extended to two other scholarly publishers and products: Evans from Newsbank-Readex and ECCO from Gale. In all, these three publishers offer commercial products that contain over 400,000 titles printed in England and the Americas between 1470 and 1800. TCP believes that there will be enough support from the library community to fund at least 40,000 texts from these

collections. Depending on how much support TCP can get, it may be possible to finish 400,000. However, TCP plans to complete at least a canonical core of texts, and then move on to other materials.

In many ways, this is an ideal model. It provides TCP with over \$1 million per year, more money than could ever be generated from a single or even a combination of grants. It involves publishers as an essential part of the scholarly communication enterprise, partners with whom university libraries should collaborate rather than compete. It draws on the monetary resources and the expertise of over 150 libraries worldwide. [7] It provides high-quality texts in a cost-effective manner (at about \$2 per title for a partner institution), and, most importantly, it provides a model of digitization that can be extended to other digital library enterprises.

TCP has three great strengths. First, it can generate a large amount of capital to be used to fund digital libraries. The overall annual budget of the TCP is nearly \$1.4 million. That capital extends not only to dollars; it also generates a great deal of goodwill between collaborating libraries and publishers, and between the scholars who are using the text and the libraries that are creating it. This is the second strength of the TCP, the community that builds around it. A common interest in TCP text brings together academic institutions dealing with scholarly communication, and TCP meetings let them explore how to work together in new ways. To date TCP has helped to foster over a dozen online scholarly projects in which scholars are taking TCP text as a base on which they can add more material and build further tools for research and teaching. Finally, TCP links scholars, publishers, and libraries, but it also bridges different kinds of scholars. Faculty members use EEBO, Evans, ECCO, and the TCP regularly, some to check references in their students' papers, to double check a catalog entry before they go to the British library, or to find a picture they wish to use in a lecture. Others write scripts to mine the corpora for specific words or phrases, download the images into interfaces they have created, or edit new editions and hyperlink them to other digital libraries. TCP can serve both kinds of scholars.

Digitization in the Age of Google

When universities like Michigan, Stanford, Harvard, and Oxford, along with the New York Public Library, announced their partnerships with Google, the dream of the universal digital library suddenly became a possibility. Certainly now all public domain works from the 19th century would become accessible, and possibly even the works currently under copyright would have at least some degree of searchability. What does this mean for TCP? On the surface the Google partnership has absolutely no effect on the work of the TCP. Google will generally not be converting and including the special collections of the libraries involved (where most of the works in EEBO, Evans, and ECCO reside), and even if they did so, it would not provide the sophisticated search engines that ProQuest, Readex, and Gale provide. That is only scratching the surface of the issue, however. What the Google partnership provides, among other things, is a model of mass digitization. If Google could feasibly start scanning the collections of the British library and indexing them, what point would there be in TCP continuing its cost-

effective—though still very expensive—way of doing the same thing that Google is doing. Why bother?

Bridging communities and bringing people together sounds like a good answer, but it is not sufficient. If Google did the same thing TCP is doing and for less money, it would likely create a de facto community that would eventually use the collection out of necessity. Though doing the same thing more efficiently (looking up references, reading books, and tracing arguments) might be Google's goal, it is not TCP's. Instead, TCP wants to help scholars harness information technology in creative ways that may not be obvious today. As researchers concluded at the recent summit on digital tools for the humanities, "revolutionary change is only possible when the foundational processes of scholarship change." [8] In other words, humanities scholars in the vanguard want to do research that cannot be done even with standard digitization.

Google could perhaps provide some tools that would help scholars do new work. They have, for instance, created the Google Scholar tool, which "aims to sort articles the way researchers do, weighing the full text of each article, the author, the publication in which the article appears, and how often the piece has been cited in other scholarly literature." [9] It is not clear how this might be done with Google's processes. Google is probably using simple OCR on Google books, the only cost-effective way to scan great numbers of books. Missing some occurrences of a word within the text is certainly acceptable for a cursory look at a book. But if a scholar in digital humanities wishes to copy and paste quotations from a Google book into an article, annotate a book for use in an electronic syllabus, and hyperlink particular words in a book to other information on the Web—actions that are common in digital humanities—missed instances of a particular word become problematic. It seems unlikely that Google would ever wish to take on the work that would be required to enable this precision for the same reasons that ProQuest, Readex, and Gale saw no reason to do so. To create accurate and searchable texts would cost Google more than the benefit they would likely gain from the small number of academic researchers who would use such functionality.

TCP: A Case Study in New Scholarship

In its six years TCP has learned that sophisticated functionality is essential to new digital humanities scholars. For instance, researchers at Northwestern University are seeking to map variant spellings that are common in early modern printed works to their modern counterparts (e.g., saynte, saint, saint). The word "saint" occurs in the EEBO-TCP database 82,579 times; including all of the variant spellings brings the total number of occurrences to 193,527. [10] The research, called Virtual Modernization, provides a better searching experience, particularly for novice researchers who would otherwise have to search on all of the variant spellings of a word—a daunting experience for undergraduates. Perhaps more important, this project exemplifies the new, previously impossible research enabled by electronic technology. Martin Mueller, director of the Virtual Modernization project, studies historical linguistics. He looks for instances of words and how they changed their meanings over time. He looks for every instance of words like "saint" over 400 years and compares the context of those words throughout

the entire corpus. [11] By mapping TCP texts to modern spellings and loading those texts into PhiloLogic, a linguistics interface tool developed at the University of Chicago, Mueller and scholars like him have the ability to ask entirely new kinds of questions. One can only imagine the new methodologies that will develop in the humanities as it adapts to this new reality: computational modeling, for instance. With large corpora like TCP, it is possible to pick a word such as “God” or “Man” and generate distributions of that word in particular kinds of literature or at certain periods of time. With those distributions, a scholar can speculate on why people would have referred to God more frequently at certain times than at others. For instance, during times of calamity, crop failures, or plague it might be more common to plead for God’s help. During the Scientific Revolution, one might expect to hear praise of Man’s achievements. Computers allow scholars to test these hypotheses by displaying frequency of words according to particular variables. They also allow scholars to map words in geographical locations to see if one location uses a particular word more than others. They also allow scholars to trace ideas geographically to see if of a word like “God” is more common in one part of the world than in another, and whether that usage migrates from one place to another. [12]

Other examples abound. Joseph Loewenstein at Washington University in St. Louis created dynamic textual editions using TCP texts. [13] Ben Schneider at Lawrence University searched for references to classical authors like Cicero and found them cited by Christian theologians. [14] Jennifer Danby’s work at the SUNY-Graduate Center traced the career of 17th-century actor Michael Mohun by searching in cast lists and even medical books. [15] Ian Lancashire at the University of Toronto created a lexicon of early modern English. [16] The Renaissance English Knowledgebase at the University of Victoria brought together primary, secondary, audio, visual, and other materials on the early modern period and created a dynamic environment in which scholars and students could do research and learn about that period in literature. [17] None of these projects relies on traditional methodologies of scholarship. All of them require materials much more accurate than Google would ever be likely to create. None of them is possible without electronic technology. A model like TCP is an essential component in creating a new electronic publishing paradigm.

Building Bridges

In 2003 TCP decided to investigate some of the ways scholars were using its resources and, with the help of researchers from the School of Information at the University of Michigan, interviewed several scholars about their use of TCP. [18] Generally scholars found TCP extremely helpful for their work. However, they were reluctant to use it in their teaching because of poor connectivity in classrooms and fears about plagiarism, both of which are common concerns for faculty who are not yet comfortable with electronic resources. Perhaps most telling was the faculty member who avoided electronic resources because they did not offer new insights, just an easier way to do the same old things (at the cost of having to learn the new technology). “I need to see something new,” he said. Another faculty member added “I am up for tenure this year; I don’t have time for this electronic stuff.” Scholars who are focused on their research need help understanding what digital humanities can do for them; they need to know what is

possible. And librarians' courses on information resources are not the answer. A study of information-seeking behavior among historians showed a preference for learning about information resources from colleagues or print literature in their fields. Only two percent of information discovery was from librarians. [19] The way to scholars' brains is through their colleagues, not their computers or their librarians.

In September 2006, the TCP provided just such an opportunity. The TCP conference, "Bringing Text Alive: The Future of Scholarship, Pedagogy, and Electronic Publication," brought together librarians, publishers, traditional scholars, digital scholars, and students to discuss TCP in particular and electronic resources in general. [20] Many of the papers went into detail about the kinds of projects EEBO, Evans, ECCO, and the TCP have enabled (such as those mentioned above). Some researchers discussed how they are searching within the TCP. Others told how they teach using these databases. In all, the conference did exactly what scholars in the TCP study wanted: It built bridges among colleagues and featured new applications.

TCP provides a cyberinfrastructure for the humanities and social sciences; a corpus of material for scholars to plumb; a community around that corpus developing new projects, supporting old ones, and discussing future needs; and most importantly, a stable, accurate means of producing text on which scholars, librarians, publishers, and students know they can rely. Arguably Google or any commercial entity could provide the first component. Scholarly societies and conferences have long provided the second. The third is the most difficult. Commercial publishers get bought and sold and are often unreliable in providing stable services over time, but libraries have been centers of research activity since the beginnings of the university in the middle ages. Even today they have led many of the changes in digital technology and continue to preserve the written record both in print and electronically. A library can provide the infrastructure that bridges the gap between communities. It can build a project like TCP, and after its initial purpose is concluded (completing 40,000 texts from EEBO, Evans, and ECCO), it can continue to support TCP.

Conclusion

In 1945 Vannevar Bush wrote "As We May Think," about a computer called the Memex that would allow a scientist to store "all his books, records, and communications," and use its "mechanized" functions to consult it "with exceeding speed and flexibility." Bush said that the Memex was "an enlarged intimate supplement to [a scientist's] memory." [21] Bush understood that the human mind would need to be the primary navigator through this labyrinth, making connections and tracing the important arguments needed to sort through a vast amount of information. In essence, EEBO, Evans, ECCO, TCP and many other electronic resources have become today's Memex, a complete record of nearly every book that has existed.

Martha Brogan identifies three criteria that undergird the most successful collaborative humanities computing projects, specifically those in American literature. These criteria are sustainability over time, accessibility to a large audience, and usefulness for a variety

of different purposes. [22] The University of California at Berkeley came to similar conclusions, and added peer review as another important criterion. [23] The American Council of Learned Societies' Commission on Cyberinfrastructure identified five criteria for successful humanities projects: collaboration; policies for tenure and promotion; workshops for scholars and teachers; opportunities to bring scholars and technologists together; and support for solving software, data storage, and technical problems. [24] TCP meets most of these criteria. It has proved a sustainable project over the past six years. It is accessible to an audience of scholars in a variety of departments at over 150 universities. It enhances collaboration and has spawned many collaborative scholarly projects. It brings together scholars, teachers, and technologists for a common goal, and it provides technical support for its texts to all of its users. Therefore it is a viable model for cyberinfrastructure that could be applied to other electronic publishing projects. Though Google, ProQuest, Readex, Gale, universities around the country, and scholarly societies all provide infrastructure of various kinds, none of these by itself is complete enough to be a Memex for humanities computing or a model accurate enough to meet the criteria necessary for cyberinfrastructure. TCP is. It is one possible way to unleash a culture of ubiquitous intellectual engagement in the humanities.