UC Office of the President
CDL Staff Publications

Title
Creating the Mark Twain Project Online

Permalink
https://escholarship.org/uc/item/7xt0z56t

Journal
Learned Publishing, 22(3)

ISSN
09531513

Author
Schiff, Lisa

Publication Date
2009-07-01

DOI
10.1087/2009305

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Learned Publishing, 22: 191–8
doi:10.1087/2009305

CASE STUDY

Creating the Mark Twain Project Online

Lisa SCHIFF
California Digital Library

ABSTRACT. The Mark Twain Project Online (MTPO) was launched in October 2007 and, in this first phase, digitally published several thousand of Twain's letters, almost 100 facsimiles of original documents, and over 28,000 records of other correspondence. The partners involved in the building of MTPO – the Mark Twain Papers, the California Digital Library, and University of California Press – created the site on the premise that Web-accessible versions of this content, enhanced by innovative design and site architecture and a suite of research tools, would greatly improve scholars' ability to discover and work with this unique material. Anecdotal feedback supports this premise. Early data analysis confirms the usefulness of facsimiles and browsing functionality. Users tend to prefer simple keyword over complex searches; however, those conducting advanced searches are using the tool flexibly, as we had hoped. Further analysis is expected to improve our understanding of how researchers use these tools in combination.

Introduction

In October of 2007, the Mark Twain Project Online (MTPO, http://www.marktwainproject.org) was released to the public. The launching of the site marked the culmination of several years of work, the last 18 months of which required an intensive focus among the project collaborators: the Mark Twain Papers, the California Digital Library, and University of California Press.

MTPO is a digital critical edition of the writings of Mark Twain, including the rich body of editorial material developed over many years by the staff of the Mark Twain Papers at the Bancroft Library. This first release of MTPO consists of correspondence, providing access to more than 2,300 complete texts, over 28,000 records of other known items, and almost 100 facsimile images. A second release, currently under development, will provide access to two literary works: *Adventures of Huckleberry Finn* and *Huck Finn and Tom Sawyer among the Indians*.

While the development of a digital critical edition of a single author can be idiosyncratic, MTPO nevertheless shares common aspirations and challenges with many other nascent digital scholarly publishing projects; thus, our findings will hopefully be of use to those embarking on the creation of new digital publications of archival and other scholarly materials.

Goals of the MTPO digital critical edition

MTPO was designed to support innovative access to a critical edition, in terms of both timely publication and robust search and display technologies. We sought to create a system that could break away from the delays associated with traditional publishing timeframes for new editions of these works. Particularly in the case of Twain's edited letters, new letters could previously only be

© Lisa Schiff 2009
This article is published under the Creative Commons 'Attribution-Noncommercial-No Derivative Works' 2.0 UK – England and Wales licence (http://www.creativecommons.org/licenses/by-nc-nd/2.0/uk/).

Lisa Schiff

brought to the public's attention when a sufficient number had been edited to warrant the production of a new print volume, a process that often took years. We recognized that a digital site could publish new letters individually, as each emerged from the editorial process. This would result in a continually updated set of materials for scholars and interested lay people to have at their disposal and, theoretically, would be much less costly than the production and distribution of an entirely new book. It is important, however, to qualify the latter assumption that digital items are significantly less costly to produce, distribute, and perhaps sell, than their physical counterparts. It seems too early in the day of these new forms of digital publication to make such a determination, particularly in light of the variations in digital formats, which have, as yet, thwarted the types of production efficiencies that publishers enjoy with printed works.

MTPO was conceived to better meet the needs of scholars and lay enthusiasts interested in the writings of Mark Twain

In addition to the goal of more rapid publication, MTPO was also conceived to better meet the needs of scholars and lay enthusiasts interested in the writings of Mark Twain. While book technology has served this audience exceedingly well, as evidenced by the amount and quality of extant scholarly and popular work on Mark Twain, digital versions of source material offer new possibilities for research and discovery. The partners involved in creating MTPO posited that expanding the pool of individuals who could access the material would increase the amount of academic work using these same items, in particular the facsimiles of letters. In addition to opening up access beyond those with budgets permitting travel to the archives in Berkeley, California, a digital edition could provide enhanced access to those materials through exposure of the carefully crafted metadata used to describe the material, along with state-of-the-art technology for searching and browsing full texts of documents. Finally, a digital edition could offer unique opportunities for displaying the original text and the editorial notes and textual apparatus, allowing for a more integrated visual connection between any given piece of editorial matter and the original text to which it refers. These goals provided ample motivation for tackling the sometimes daunting issues involved in actually creating the site.

Building MTPO

Who: a collaboration

MTPO was built as a collaborative effort between three central partners: the Mark Twain Papers & Project at The Bancroft Library of UC Berkeley (the Papers, http://bancroft.berkeley.edu/MTP/); University of California Press (UC Press, http://www.ucpress.edu/), the Papers' primary publisher; and the California Digital Library (CDL, http://www.cdllib.org) of the University of California. Although areas of responsibility overlapped, there were clear centers of expertise for each of the partners. The Papers provided the content (including XML-encoded material and extensive metadata) and editorial expertise. The Press provided an imprint (thus granting the publication an essential marker of credibility on a par with the Press's traditional and highly valued print offerings), secured qualified individuals to participate in user assessment, and marketed via their traditional channels. The CDL provided core technical development, project management, and information architecture and design resources. Each of these components was essential to the eventual publication of MTPO, and each component fell far beyond the scope of any one of the three partners, so that collaboration was essential.

What: the ingredients

MTPO was built using content resulting from years of careful textual editorial work by the Papers. Prior to the release of MTPO, this work had been made available only in print form, apart from a Web interface that allowed scholars to search through one of several databases maintained by the Papers. For the purposes of creating a digital critical edition that same content had to be transformed into electronic resources that could be served up through a Web-based application. This involved transforming the following classes of content into appropriate digital formats: the texts (including all editorial matter); the metadata (meaning all

descriptive information about a given piece of content); and any images, including facsimiles of manuscripts, artwork in the literary works, and photographs.

The texts – the written materials and accompanying notes and textual apparatus – were encoded according to the Text Encoding Initiative standard (TEI, http://www.tei-c.org/index.xml), which is a widely accepted set of XML guidelines and conventions for describing texts for processing by computers. Basic encoding of the letters and the literary works was outsourced to a third party, who worked against a set of detailed guidelines established by the Papers. Such guidelines are necessary because of the idiosyncratic nature both of the original texts and the editorial practices. For instance, the Mark Twain Papers uses its own transcription system called 'Plain Text' (http://www.marktwainproject.org/about_hirst_essay.shtml), which is a set of specific graphical markers and conventions employed by the editors to designate various deletions, insertions, comments, and other notations (http://www.marktwainproject.org/xtf/view?docId=letters/MTDP00005.xml;style=letter;brand=mtp). All of these indicators have been used in the print publications of the letters, and therefore needed to be translated into the digital edition. To accomplish this, the staff at the Papers had to determine how to express the editorial signs and indicators in the TEI encoding of the text. For example, they had to decide how to describe the various types of braces that spanned multiple lines on either side of a page, as in the examples below. This was made more difficult by the number and variation of such signs, the incomplete coverage of these signs and symbols in readily available Web fonts, and the varying degree of screen (as opposed to print) legibility for certain symbols.

In addition to requiring the texts themselves to be in machine-readable format, MTPO, like any robust search and display system, also requires a good set of descriptive data about the texts that it provides access to, in order to provide more and better ways to discover and work with the materials. Descriptive data (metadata) had been captured and developed over time by the editors of the Papers as they worked with the texts. Originally stored in several different databases, these metadata required a significant amount of work to rationalize and combine the disparate sources into one new database that could both more effectively store the information and also present it in a variety of different ways. The metadata records for each content object within MTPO have been generated from that database and expressed according to the Metadata Encoding and Transmission Standard (METS, http://www.loc.gov/standards/mets/), a standard used widely within the library community and elsewhere for capturing and sharing metadata about numerous types of objects. These metadata files include administrative information about the objects, such as the unique identifier, and descriptive information, ranging from the title of the work to the addressee of a letter.

Facsimiles of original letters make up the third major component of the material currently available to scholars through MTPO. These facsimiles are JPEG derivatives of high-quality TIFF files that were generated by scanning the original letters and any related items such as envelopes and included notes.

A final ingredient is, of course, the financial support. Any project of this scope and complexity requires tremendous resources. Funding for MTPO came from a variety of sources, including the Bancroft Library, the Papers' parent organization; the CDL; and the National Endowment for the Humanities. MTPO does not have a cost-recovery model at this point; the site is available without charge to any who wish to use it.

any project of this scope and complexity requires tremendous resources

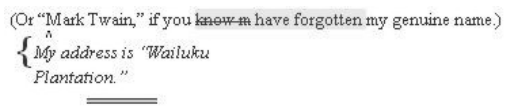

Figure 1. Encoding and rendering editorial symbols: left brace

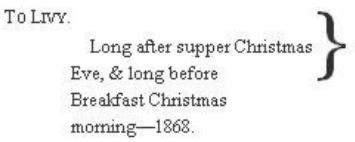

Figure 2. Encoding and rendering editorial symbols: right brace

scholars wishing to work with such a large collection need intuitive means for filtering and creating smaller, meaningful subsets

How: the tools

MTPO was developed using the open source eXtensible Text Framework (XTF, http://xtf.wiki.sourceforge.net/), a search-and-display application created by technologists at the CDL. XTF is a Java-based program built on top of two other open-source tools, the search engine Lucene and the XML parser Saxon. XTF allows robust searching of documents, including features such as relevance-ranked search results, spelling correction for queries, and faceted browsing (see below). While it also provides basic display of content and search results, customization can be easily achieved through the use of XML stylesheets, meaning that no Java programming is required to develop unique, content-specific applications that meet the identified needs of targeted user populations. For this project, the user population consisted primarily of Twain scholars, with a secondary audience among lay enthusiasts, paralleling the audiences for the print materials. Fortunately the needs of these communities, especially the former, were relatively well known to us as we could draw on the nature of the questions individuals have brought to the Papers over the years, along with decades of research by the library community on the scholarly communication needs and practices of academics. Because of the flexibility of XTF, the primary developer working on the MTPO site was more readily able to build and experiment with features intended to respond to these needs, such as the dynamic interaction between the editorial notes and the transcribed text, in which hovering over a note in the right pane of the browser window would cause the associated text in the main frame to be highlighted.

Essential features

MTPO is a feature-rich site, in terms of both functionality and visual design. Early in the development process, four features in particular were identified as essential to the publication of the digital critical edition: faceted browsing of the content; advanced searching; complex displays of editorial matter and text; and citability at a granular level.

Faceted browsing

MTPO contains many thousands of items, including full-text letters, records of other known pieces of correspondence, and editorial essays. Scholars wishing to work with such a large collection need intuitive means for filtering and creating smaller, meaningful subsets. For the letters within MTPO, researchers want to be able to search by, for example, the date of correspondence; the writer and addressee(s); the place of origin; the availability of text and/or facsimiles; and

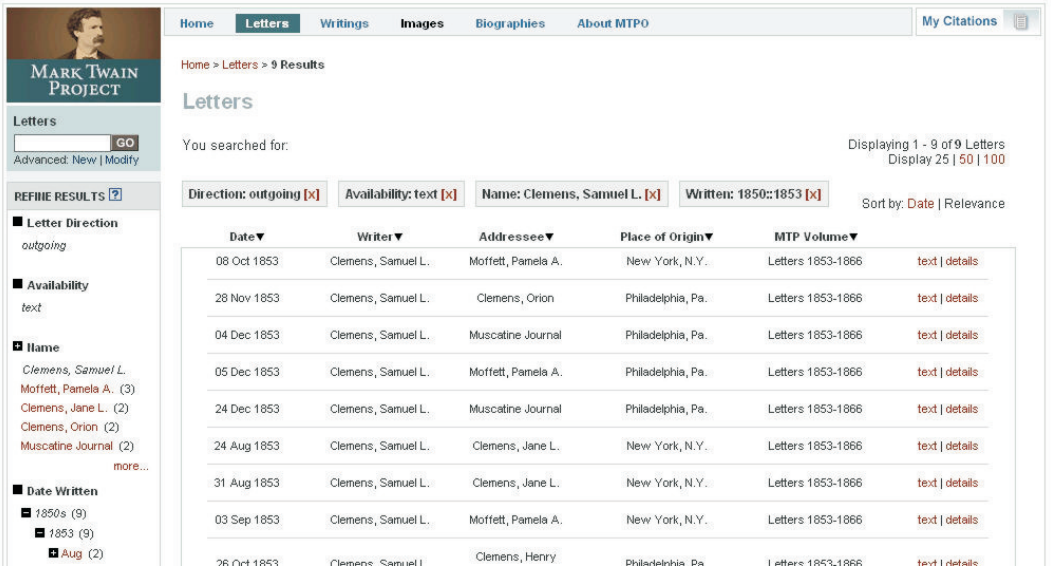

Figure 3. Browsing the collection using facets

the institution housing the material. Users are able to see the available values in each of these dimensions and choose among them, either to browse or to narrow their search results. In this way, facets present scholars with an immediate overview of the content available, while supporting serendipitous discoveries and avenues of exploration that go beyond the results to be expected through sophisticated search strategies alone (although these are also supported by MTPO).

Advanced searching

Effective searching is increasingly achieved by using simple keyword search boxes. Due to the influence of Google and the fortunate ubiquity of search engines using relevance-ranking algorithms, entering a word or phrase into most search interfaces yields very good results for many researchers. However, scholarly users working in their own area of expertise both want and are able to formulate much more precise search queries. For instance, with MTPO we felt that a basic requirement of the digital critical edition must be the ability for users to distinguish between Twain's words – the original edited text – and the editorial material generated by the Papers. Our strategy was to offer an advanced search feature that allowed users to choose one or all of the broad categories of materials available to be explored, including the text itself, explanatory notes, introductions and documents,

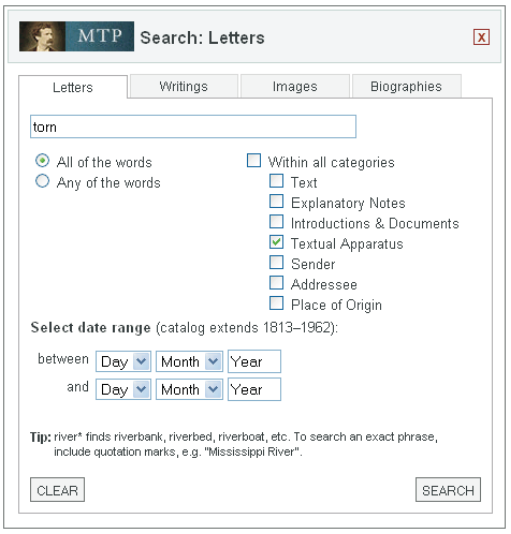

Figure 4. Using advanced search to find torn letters

and the textual apparatus. Offering this type of choice allows users to make the important distinction between Twain's own words and those of the Papers' editors; Figure 4 indicates how a researcher might find editorial references to letters that were torn, as opposed to correspondence in which a writer used the word 'torn'.

scholarly users both want and are able to formulate much more precise search queries

Complex displays

Digital critical editions can, and thus arguably should, present transcribed material, editorial material and facsimiles in close proximity, so that the various components of a given intellectual object (e.g. a letter) have the least possible visual and conceptual distance between them. Various strategies

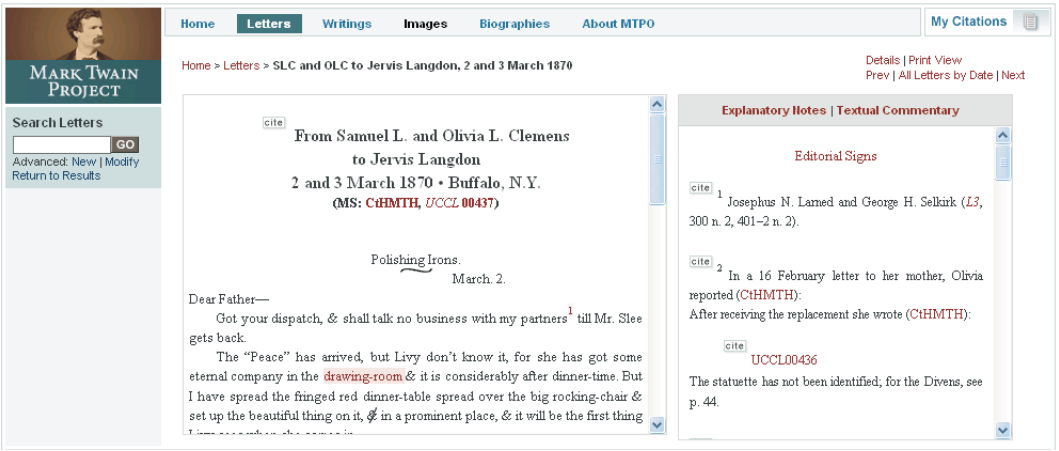

Figure 5. Coupling text and editorial matter

exist for accomplishing this convergence. MTPO's solution was to present the transcribed text always alongside the explanatory notes and textual apparatus, further enabling toggling back and forth between a display of the editorial matter and a facsimile where available. While we recognized the importance of supporting certain useful research habits from the print publications, such as allowing the text or editorial matter to be read and explored separately (e.g. a user can use the scroll bar to read all of the explanatory notes and textual apparatus for a given letter without changing the position of the letter text), the digital environment also allowed for a more dynamic expression of the relationship between editorial material and text. By clicking on an editorial note in the display, users automatically highlight the associated material in the main text, and vice versa. These features, among others, were developed in order to facilitate the reading of editorial notes and texts simultaneously, a task rarely possible when consulting a print publication.

by clicking on an editorial note, users automatically highlight the associated material in the main text

Granular citations

A final essential feature of MTPO is the citation functionality. The research value of a scholarly work is dependent on its citability: other researchers must be able to re-create and evaluate the argument that is being presented to them. Many digital sites have content that includes page numbers imported from print versions or other place indicators, such as paragraph or line numbers, which because they are reliably stable can be used by scholars in their own subsequent publications. The materials in MTPO did not have any such stable markers, as the page numbers from the original printed volumes were not transcribed and, furthermore, newly edited letters that are digitally published would not have page numbers. Paragraph or line numbers were also not options, given the dynamic nature of the editorial work on the texts. The solution we developed to accommodate these complexities involves a widget that generates a citation at the level of predefined 'citable chunks' (as decided by the editors), using identifiers in the TEI encoding. Those citable chunks include a letter as a whole (e.g. a citation for Clemens' July 3, 1874 letter to his wife) and individual notes. In the next release, which includes two of Twain's literary works, users will be able to cite at the paragraph level.

Challenges

Building any large website is a complex endeavor, and MTPO was no exception. The areas of complexity included establishing and managing partnerships; achieving a robust yet intuitive information architecture; maintaining standards of quality; and providing appropriate and powerful tools.

Partnerships

The challenges of partnerships are well known. For MTPO, two areas in particular were particularly complex. The first concerned the process of creating the design of the site: both the information architecture (i.e. what activities should be supported and what should be the logical flow from activity to activity within the site) and the visual design (the branding, look and feel, font, etc.) were developed and worked out through a large committee with multiple representatives from each of the three partners in the collaboration. On the one hand, this process ensured that many perspectives, important details, and central concerns were brought to every meeting. On the other hand, we discovered that we spent a good deal of time at each meeting in discussion, debate, and negotiation – a necessary process but one that becomes increasingly difficult as the numbers involved increase.

The second area of collaboration that we found to be a challenge was the technical development. The technical work related to the content and metadata was all handled by technologists at the Papers, while the Web development and the XTF implementation of the design based on the content was the responsibility of the technologists at the CDL. Complications with this division of labor arose because of the difficulty of separating these two dependent development efforts. Additionally, as designs were implemented in the Web application, decisions about functionality and display sometimes

needed to be revised, thus requiring changes both in the encoding of the metadata or content and in the application itself.

Information architecture

The information architecture of MTPO was challenging in itself, not simply because it was developed with such a large group. Determining how to provide intuitive, aesthetically pleasing access to such a large, complex, interdependent body of work poses many problems. For example, much time was spent attempting to devise a way to allow a researcher looking at a single letter to see that letter's position in the editorially determined chronology of all correspondence. The original thought was to display the chronology as a set of hierarchical links in a left pane, similar to a table of contents, thus visually sandwiching the letter text between the chronology and the editorial matter. However, usability testing indicated that this presentation was too similar to a table of contents without actually being one, resulting in ambiguity and confusion for potential users. Our solution was to create a separate page presenting all letters in order, available from a link at the top of each individual letter. We also allowed users to step backwards and forwards in the chronology from the point of a given letter, through 'Prev' and 'Next' links.

The problem of the letters' chronology highlights another significant challenge: continuing to enable practices already well supported by the print volumes (e.g. using a table of contents to understand and access text), while avoiding the display constraints of those same volumes. We needed to respect the comfort and ease that persists in the use of print publications, but still to exploit the advances enabled by digital display of content. This difficulty has been discussed and written about extensively – and yet it remains. One component of the problem is our continued grounding in the world of print; this is explored by Hillesund in his discussion of the slow transition into what he terms the 'digital text cycle' (as opposed to the 'print text cycle'), a transition that is both enabled and constrained by backbone technologies such as TEI.[1] Robinson has explored how the continued lack of general-purpose, easy-to-use development tools maintains a high hurdle for creating digital editions, despite the many decades he and others have spent attempting to devise such tools.[2] Finally, Nowviskie has rightly brought the reader's experience to the foreground by making the compelling case that a solid theory of interface must propel the building of digital editions, if they are to successfully support scholarly users in their research efforts.[3]

Quality standards

The editors at the Papers quite rightly maintain high standards for all of their work, whether it appears in print or digital form. This helps to ensure that what is ultimately presented to the public is a solidly established text of Mark Twain's writings. The difficulty lies in translating those standards to the realm of digital publication. An example of this was the negotiation regarding the representation of information within the date facet, which enables the narrowing of results by date. The editors have a particular way of recording letter dates (e.g. 1869.02.13), but our usability requirements made clear the need to use month abbreviations. A standard had to be agreed upon that was sufficient in terms of quality, but that would work well within a narrow left pane of a website (e.g. 'Feb' as opposed to 'Feb.').

the challenge is for scholars to feel comfortable using it as a legitimate academic resource

Scholarly legitimacy

The ultimate challenge for MTPO, or for any digital critical edition, is to make the site sufficiently robust for scholars to feel comfortable using it as a legitimate academic resource. The requirements for success in this regard include, as discussed earlier, a solid information architecture that is intuitive and has a compelling design; maintenance of editorial standards; tools to make it easier to work with the content; access to facsimiles as well as digitized versions of original documents; and a means of creating reliable citations. Designing and building the site so that all of these aspects were sufficiently developed, without one compromising the other, involved a constant

process of experimentation, negotiation, and iterative design and development.

Conclusion

Early feedback indicates that researchers are able to work with the material more easily than before. For instance, a graduate student has reported how MTPO's search and citation functionality have aided doctoral research. This individual has been able quickly to discover and read correspondence with particular individuals of interest, and also to capture and save the references in digital form. This is in contrast to the often frustrating exercise of combing through texts, manually crafting references, and then being unable to return to the original items later on due to copying errors and omissions.

we need to determine how much our user communities are taking advantage of the tools

However, anecdotal information is never sufficient, especially considering the significant resources contributed by all partners to make this site a reality. We are still gathering and analyzing usage data to determine how well the site is accessed and by whom. Results to date show that there is a substantial amount of viewing of facsimiles and significant use of faceted searching to explore the collection, particularly by date, status of online availability, and letter direction (i.e. was the letter sent or received by Twain). While keyword searches are much more heavily used than advanced searches by almost six to one, we can see that users are indeed taking advantage of the ability to search within the original text only or within the editorial material, which demonstrates the importance of this type of refinement for scholars in this area. Users also seem to be actively generating citations, but unexpectedly they are not emailing those citations to themselves (or others) to facilitate the creation of references in a paper or for further reading and analysis at a later date. One possible explanation for this 'underuse' may be that individuals simply copy and paste the list of citations into a file saved on their personal computer; this behavior could be confirmed with more user testing.

As we continue to collect and analyze our data, we will closely examine our users' behavior on the site, looking particularly at whether the available features are used in combination, and if so how. Collecting this data will help us to evaluate our own suppositions about the relative need for features such as faceted browsing, advanced search capabilities and sub-object level citation widgets. We need to determine the degree to which our user communities are actually taking advantage of the tools we have built for MTPO as the relevance of this site is dependent upon offering scholars an intuitive and compelling alternative to the traditional print publication of the critical edition of Mark Twain's works. Such an analysis is especially important as the site moves into new phases. In the near term, as mentioned earlier, both *Adventures of Huckleberry Finn* and *Huck Finn and Tom Sawyer among the Indians* will be brought onto the production site. Future, as yet unscheduled releases will likely make available to users additional literary works, a rich collection of images, and a selection of secondary materials.

References

1. Hillesund, T. 2006. Digital text cycles: from medieval manuscripts to modern markup. *Journal of Digital Information* 6. Available at http://journals.tdl.org/jodi/article/view/jodi-164/65 (accessed 17 February 2009).
2. Robinson, P. 2005. Current issues in making digital editions of medieval texts – or, do electronic scholarly editions have a future? *Digital Medievalist* 1. Available at http://www.digitalmedievalist.org/journal/1.1/robinson/ (accessed 17 February 2009).
3. Nowviskie, B. 2000. Interfacing the edition. Available at http://www.iath.virginia.edu/~bpn2f/1866/interface.html (accessed 17 February 2009).

Lisa Schiff, Ph.D.
Technical Lead, eScholarship Publishing Group
California Digital Library
300 Lakeside Drive #744
Kaiser Center, Oakland, CA 94612, USA
Email: lisa.schiff@ucop.edu
Website: www.cdlib.org