Advancing Scholarship through Digital Critical Editions: Mark Twain Project Online

Lisa R. Schiff[1]

[1]California Digital Library, University of California
415 20th Street, 4th Floor, Oakland, California, United States
e-mail: Lisa.Schiff@ucop.edu

Abstract

Digital critical editions hold the promise of supporting new scholarly research activities not previously possible or practical with print critical editions. This promise resides in the specific ability to integrate corpora, their associated editorial material and other related content into system architectures and data structures that exploit the strengths of the digital publishing environment. The challenge is to do more than simply create an online copy of the print publication, but rather to provide the kind of resource that both eases and extends the research activities of scholars. Authoritative collections published online in this manner, and with the same rigor brought to the print publishing process, offer scholars: the ability to discover more elusive, granular pieces of information with greater facility; tighter, more obvious and more accessible connections between authoritative versions of texts, editorial matter and primary source material; and continually corrected and expanded "editions," no longer dependent upon the print lifecycle. This paper will explore these benefits and others as they are instantiated in the recently released Mark Twain Papers Online (MTPO) (http://www.marktwainproject.org), created and published as a joint project of the Mark Twain Papers & Project at The Bancroft Library of UC Berkeley (the Papers), the University of California Press (UC Press), and the California Digital Library of the University of California (CDL). This current release of MTPO is comprised of more than twenty three hundred letters written between 1853 and 1880; over twenty eight thousand records of other letters with text not held by the Papers; nearly one hundred facsimiles; and makes available the many decades of archival research on the part of the editors at the Papers. Of particular focus in this discussion will be several key features of the system which, despite the many challenges they presented in development, were felt to be essential pieces of a digital publication that could support scholarship in new and significant ways. Those features include facets, which create intellectual structure and support serendipity; advanced search, which provides a means for researchers to apply their own analytical frameworks; citation support functionality, which serves to secure and record the outcomes of research exploration; and complex displays of individual letters, which allow detailed inspection by collocating the pieces of the authoritative object. These features together maintain the integrity and stability of the collection, while concurrently allowing for fluidity in the continued expansion of the material. In this way, MTPO hopes to succeed as a digital critical edition that will support and extend the research activities of scholars.

Keywords: Mark Twain Project Online; digital critical editions facets; search; citation support

1. Introduction

Digital critical editions hold the promise of supporting new scholarly research activities not previously possible or practical with print critical editions. This promise resides in the specific ability to integrate corpora, their associated editorial material and other related content into system architectures and data structures that exploit the strengths of the digital publishing environment. The challenge is to do more than simply create an online copy of the print publication, but rather to provide the kind of resource that both

eases and extends the research activities of scholars. The vision this challenge speaks to has been compellingly articulated by Robinson [1], an early and consistent explorer in this terrain, with his work on such efforts as the Canterbury Tales Project [2] and the development of tools such as the XML publishing application *Anastasia* to bring digital editions online, who calls such works essential for the future of humanities scholarship.

Authoritative collections published online in this manner, and with the same rigor brought to the print publishing process, offer scholars: the ability to discover more elusive, granular pieces of information with greater facility; tighter, more obvious and more accessible connections between authoritative versions of texts, editorial matter and primary source material; and continually corrected and expanded "editions," no longer dependent upon the print lifecycle and able to take advantage of a larger pool of knowledge found in both the scholarly and lay communities. D'Iorio [3] provides a dramatic example of the value of such possibilities in his account of the impact of a mistranslated word in a collection of aphorisms by Nietzsche, a mistranslation which he argues undercuts an interpretation of Nietzsche's concept of the "will to power" by the influential philosopher Deleuze [3 p.2]. Had the original manuscripts been available to the scholarly community long before, this translation error might have been noticed due to the presence of so many more individuals examining the facsimiles and their associated transcriptions, translations and analysis.

This paper will explore these benefits and others as they are instantiated in the recently released Mark Twain Papers Online (MPTO) [4], created and published as a joint project of the Mark Twain Papers & Project at The Bancroft Library of UC Berkeley (the Papers) [5], the University of California Press (UC Press) [6], and the California Digital Library of the University of California (CDL) [7]. This current release of MTPO is comprised of more than twenty three hundred letters written between 1853 and 1880; over twenty eight thousand records of other letters with text not held by the Papers; nearly one hundred facsimiles; and makes available the many decades of archival research on the part of the editors at the Papers. MTPO draws from 30 volumes of previously published material—including the critical apparatus created by the Mark Twain Papers—which have been encoded in XML according to an extended version of the Text Encoding Initiative P4 (TEI P4) [8] customized by technologists at the Papers just for this collection. The metadata gathered by the Papers has also been generated by them according to the XML-based metadata standards METS and MADS. Together the TEIs, METS and MADS served as the primary inputs to the MTPO system – an implementation of CDL's eXtensible Text Framework (XTF), a robust and flexible platform for providing search and display solutions for collections of digital content. UC Press has long been the publisher of the Paper's critical editions and serves in the same capacity by offering their imprint for the digital critical edition.

Of particular focus in this paper will be several key features of the MTPO system which, despite the many challenges they presented in development, were felt to be essential pieces of a digital publication that could support scholarship in new and significant ways. Those features include faceted browsing, advanced search, citations, and complex displays of individual letters.

2. The State of Digital Critical Editions

Critical editions, digital or otherwise, are the authoritative texts that textual editors, produce and that other scholars use for their work. The material difficulties and philosophies and approaches towards creating such texts continue to be widely debate, as can be seen in the works of both central participants such as Tanselle [9] and new entries into the profession such as Gunder [10]. While such theoretical matters are significant, they are not germane to the discussion at hand, which is focused on how one produces such a digital edition in a way that accurately reflects the philosophy of its textual editors, while presenting itself usefully and meaningfully to the literary scholars that come after. For this discussion then, a very simple definition of the notion "digital critical edition" can suffice, specifically, a coherent collection of content

which has been assembled by qualified textual editors and that serves as the authoritative versions for scholars wishing to work with original texts.

2.1. The Goals and Requirements of Digital Critical Editions

While digital critical editions should adhere to the high standards of the practices of textual editors, they need not, and indeed should not, simply replicate the printed volumes which have preceded them. The digital medium offers different constraints and opportunities, which must be openly acknowledged and addressed. Hillesund [11] has described the difficulties and possibilities presented at this moment when, despite our immersion in the Web, we are still quite grounded in print processes and technologies, so much so that many of our digital efforts are merely electronic re-presentations of print works, drawing from the same workflows and perspectives that go into making print publications. Hillesund argues that we need to shift from such a "print text cycle" to a "digital text cycle", in which "texts are produced, distributed and read with the aid of computers, networks and monitors in a predominantly digital environment." He identifies two hallmark attributes of the digital text cycle : the separation of the storage of the text (or data) from its manifestations (increasing the likelihood of a variety of presentations) and a significantly increased amount of digital reading, as opposed to print reading.

The qualities embedded in Hillesund's notion of the digital text cycle quickly become obligations for digital critical editions, appearing as goals that those working in the field aspire to attain and requirements that those awaiting such texts expect to be met. The first of these is easy access to more original manuscripts. Along these lines, Tanselle [9] reasonably insists that a "hypertext" scholarly edition is "inadequate if in addition to transcriptions and editorial matter, it does not offer images of the original documents, both manuscript and printed. Important physical evidence will obviously still be unreproduced, but at least the range of paleographical and typographical evidence made available will be far greater than has been customary in editions of the past—even in "facsimile" editions, which have usually been limited to single documents." Tanselle even goes so far as to point to the need for regenerated texts, via the assemblage of bits and pieces: "Indeed, the point can be made more positively: that critically reconstructed texts ought to be included within the collection of texts available in a hypertext edition. Readers can of course make their own choices among variants, using whatever bases of judgment they wish, just as they have always been able to do with other forms of apparatus—though with hypertext they can more easily produce a smooth reading text of their own construction."

The availability of original documents, and especially of the types of recreations Tanselle calls for, leads to a second requirement and goal, which is the wider exposure of the editorial work to a broader scholarly community (and the well-informed lay community as well), intensifying the gaze on all aspects of the text thereby increasing opportunities for corrections of transcription and translation errors, and the bridging of gaps created by incomplete collections of materials. D'Iorio [2] gives a gives a resoundingly convincing argument of the need for digital critical editions in his gloss of the problematic history of a non-existent work attributed to Nietzsche, The Will To Power, in which he describes the various publications of this constructed work, the errors of translation and transcription which they include, and the resulting problematic scholarship, as mentioned earlier. D'Iorio's argument is that much of this could have been avoided through access to the original manuscripts. We can also add that access to the trail of scholarship for which these documents have served as the genesis would also have greatly enhanced the conversation about this text.

Continual accretion of the edition by the entire community of interested parties is yet another goal and expectation theoretically more feasible with a digital critical edition. For instance, D'Iorio's vision [2, p.4] with the HyperNietzsche[12] project is of an edition that allows for correction and debate over primary sources, but that also serves as a focal point for contributors of vetted scholarly analysis:

> *"...one may consider HyperNietzsche as the integration of a public archive, which*

allows free access to primary sources; a public library, which allows free access to critical editions and other scholarly contributions; and a non-profit academic publisher with a prestigious editorial board and rigorous procedures of peer review."

The significance of expanding the participation within a scholarly community of interest is so great that it has been taken on (in addition to other goals) in en EU initiative called Interedition [13], which aims to create "international digital infrastructure for scholarly editorial work."

Certainly these opportunities are significant for MTPO, as the Papers receive and acquire new letters in a continual stream, a flow which they wish to extend to include a regular incremental dissemination of such texts once the editorial process has been completed. The goal is to move further into the "digital text cycle" realm and build an infrastructure that allows the digital collection to grow and be made available apace at which the work on small groups or even individual letters is completed, as opposed to having to wait until there are a print-volume-worthy number of texts ready for public consumption.

2.2 The Challenges of Digital Critical Editions

The opportunities, goals and expectations of digital critical editions are not without their challenges. The constraining effect of print technology on how we imagine digital publications has almost become a trope. More interesting are the ways in which particular pieces of the work of producing digital publications (critical editions and other works) act in this liminal space. Hillesund [11] provides a very compelling analysis of the constraining effects of a backbone technology for digital publishing, XML, showing that while it achieves the goal of separating storage and representation, it can be seen as a limiting transitional technology as it is so frequently tied to print displays, especially in such standards as TEI:

"This promotion of XML structures will have the paradoxical (and obviously unintended) consequence that conventions of print will dominate digital publishing for a long time, especially the parts based on cross-media publishing and single sourcing. These production workflows will lead to print-based content structures being contorted to fit new media, while new genres which exploit the potential of digital media will not be developed."

Robinson [1] provides a different perspective on the challenges to digital critical editions, locating the major obstacles on the lack of usable tools and the unwillingness of major publishers to put forth these works. Formulated as goals instead of shortcomings, Robinson argues:

"Our goal must be to ensure that any scholar able to make an edition in one medium should be able to make an edition in the other. Further, that an edition in either medium should be equally assured of appropriate distribution: just as once a library has bought a print edition it can be used by any member of the library for years to come, so too should it be for electronic editions."

The call for tools has been both echoed and responded to by those involved with such systems as NINES [14] and the related Collex system [15], which, as described by one of the developers, Nowviskie [16], offers a standard platform for producing scholarly editions and end-user oriented tools for working with the resultant collections.

Many have taken on the challenge to produce editions that move us along towards achieving the goals described above. Many notable examples of digital critical editions currently exist, particularly excelling at the provision of access to manuscript facsimiles. The most recent and perhaps most popularly known of these is the Charles Darwin site [17], which received over 7 million hits to its website on the day of its public release, April 17th, 2008. A selection of other examples that focus on facsimiles include the Rosetti Archive [18], the Walt Whitman archive [19], the Blake Archive [20] and HyperNietszche [12].

In addition to addressing the problems involved as producers of digital critical editions, a major area of

concern must also be with the usability of such works on the part of the scholar or lay researcher. This concern resonates with Nowviskie's [21] identification of the need for a theory of interface for critical and scholarly editions, noting that little thought has been given to "editions as interface, though, because we as readers are so accustomed to jacking into the codex." While this arena has received less attention, it is a lynchpin in the success of such works. Some sites that have taken on the challenge of prioritizing the end-user experience (with or without an underlying theory of that interface) include the NINES based sites which use Collex and the Carlyle Letters Online [22]. These publications have been experimenting with end-user tools arenas, such as facets, tag clouds, advanced search functionality and support for citation activities, tools widely available in other online arenas. Similarly, interface was a motivating factor in many of the decisions regarding MTPO as will be shown the discussion that follows.

3. Mark Twain Project Online

The Mark Twain Project Online is the result of an effort to meet the challenges described above, an attempt to provide a scholarly resource that would significantly enhance the research activities of those interested in Mark Twain. The original vision for the complete works of Samuel Clemens (SLC) online began in 2001, when members of the Papers realized digital editions were a requirement of the future [23] and soon began the process of having the texts encoded in an extended form of TEI P4 [24].

At this point, a natural collaboration formed between the Papers, their publisher UC Press, and the CDL, in which each party contributed in distinct ways to the achievement of the aforementioned goal. The interface, known in CDL parlance as the "User Experience" or "UX," became a defining aspect, a touchstone for decision making regarding feature selection, representation of information, and content selection. Developing the UX was a lengthy group process that involved balancing the differing allegiances to various aspects of the project, namely the purest reflection of the material and the editorial work that had gone into producing the edition; the intuitiveness and ease of the interface; and the overall production qualities of the design and content scope.

A year and a half later, the beta version of MTPO was launched, complete with components the collaborative team deemed essential for successfully using MTPO. Those components included: 1) faceted browsing functionality, in order to provide a stable structural view that would allow users to easily grasp the scope and nature of the collection even as it grew transparently over time and that would also support purposeful exploration and provide opportunities for serendipitous findings; 2) robust advanced search functionality providing support for general searching activities as well as precisely formed research tasks, allowing the user to apply his or her own analytical framework to the collection; 3) citation support, by allowing users to easily generate and collect citations at useful levels of granularity; and 4) wholistic presentations of letters, providing integrated access to interdependent layers of the text (e.g. transcriptions, facsimilies, notes and apparatus entries). Each of these components will be discussed below.

3.1 Essential Features: Facets

Facets are the various dimensions of a piece of content, such as, in the case of a letter, the sender, receiver and date of composition. When facets can be defined across an entire set of documents, as in a collection of letters, they provide an effective means for users to peruse a large corpus, browse through a number of items across one or more dimensions, and alternately, hone in quickly on a specific slice of a collection.

Although facets are commonly discussed from the perspective of end-users, the notion was originally developed by the librarian S.R. Ranganathan [25] in the early 1930's as a more effective way for catalogers to describe the subjects addressed within a document. Ranganathan's approach was powerful in that it advanced a composite approach to describing a document, an approach in which individual concepts are

identified and then coexist as a set providing a richer description of the document than can be achieved by having to choose among pre-grouped concepts, adhering to a more brittle classification scheme. This more atomistic, dynamic approach allows for specificity and agility by allowing for terms to be combined as needed, as opposed to in advance of use.

Ranganathan's conception of facets (and that of those who followed him, namely Bliss [26]) was still a highly formalized structured, which has of late been overtaken by a more organic approach as an increasing amount of content is being provided digitally and thus the demands for creating more effective descriptions in order to promote discoverability are breaking down that formalism. The power of facets has been discovered by a range of people, from ecommerce sites, hoping to more effectively pair goods and consumers, to information retrieval scientists in their continued effort to improve our ability to find "relevant" information from large masses of content. Facets are useful along this spectrum because they provide improved intellectual access to such objects by breaking out discrete, meaningful attributes, making it easier to both to describe and discover content. In addition, they are flexible and adaptable in that facets can be added (or eliminated) as they are deemed useful dimensions on which to describe a collection and because they support multi-dimensionality as they can be combined as required for both the descriptive and discovery tasks in which we work with objects.

Contemporary discussions of facets are frequently focused on automatically generating close approximations of handcrafted facets through activities such as clustering (Hearst [27]) and on how to effectively present facets in interfaces to better support the activities of end users in discovery. Hearst's Flamenco browsing system [28] has shown that facets, particularly hierarchical facets support browsing and unstructured exploration of large collections, allowing for serendipity in a way that searching does not. The Flamenco system not only presents facets but allows users to move up and down a ladder of specificity within any dimension, using a hierarchy within each facet where appropriate. In another application of Flamenco, Yee et al [29] have demonstrated that hierarchical facets are not only useful for discovery of textual objects, but of images as well. The significance of hierarchical and interdependent facets as opposed to simply flat "bucket" categories has also been discussed by Ben-Yitzhak et al [30] in work on the use of faceted "search" that supports the complex, enterprise decision making.

Facets are powerful, but that power is dependent upon the quality of the metadata that populates any given dimension, the degree to which that metadata is available across an entire collection and how clearly it is displayed. Additionally, although facets can seem intuitive if built properly, their use and manipulation can become complex putting a serious burden on the interface to clearly guide the user. For MTPO, facets were essential as the only way users would be able to explore such a large volume of texts and records. The key dimensions that were ultimately built were: letter direction (i.e. whether the letter was sent or received by Clemens); availability (i.e. whether the transcription and/or facsimiles were online); name (i.e. named persons in the "envelope" of a letter; date written; the place of origin; the repository holding the original letter; (when the letter was written); and print volumes in which the letter is referenced.

Using the CDL's eXtensible Text Framework (XTF) [31] as its backbone enabled MTPO to be built with facets that are both hierarchical where appropriate (with dates, for example) and interdependent, meaning that as a user makes a choice within one facet, the possible set of values within all other facets is updated as well. For instance, if a researcher choices "Elmira, NY" from the "Place of Origin" facet, not only are the letters presented in the results pane only those letters originating from Elimra, NY, but the remaining choices within the facets are refined appropriately, so the people that can be selected in the "Name" facet are now only those people who are associated with letters that originated in Elimra, NY, and so on.

The challenges in providing such a faceting feature lay not in the technology of the XTF application, but rather in questions of consistency in the data (because unlinked variations of names or other data will simply appear as one more choice, and will not return to the user all the associations he or she might

expect), but more importantly in: 1) determining formulations of that metadata that were acceptable to the editors of the Papers, feasible to create and work with for the technologists on the team, and meaningful to end users; and 2) identifying where and how in the interface users would explore and maniuplate the facets.

The seemingly innocuous example of the "date written" facet serves as a good illustration. In this case, the goal is to allow the researcher the dimension of time in which to examine the letters, supporting activities such as using dates to find specific letters, or letters within certain time periods, or to combine questions of date of writing with other aspects, such as developing an understanding of how Mark Twain's correspondents changed over the course of time. Providing an ability to ask and get answers to such questions involved first ensuring that the data actually was available for all letters where possible, devising solutions for those letters for which there was no date, and secondly maintaining the editorial integrity of the date information by coming to agreement upon a notation that could also display reasonably well.

This second challenge is a good example of the work that goes on in this intervening space between the text and digital text cycles. Taking an individual letter, such as SLC's letter of February 13, 1869 to his future mother-in-law [32] as a case study, the editors at the Papers recorded this date as "1869.02.13" which was converted in the METS metadata record to an ISO 8601 compliant date format [33] "1869-02-13" but is presented in the print volumes as "13 February 1869." The first formulation is not end-user friendly and the third formulation does not fit into the narrow facet sidebar of the website, nor does it lend itself to easy visual parsing by end users for narrowing down by decade, year, month and day. Thus while the canonical date form was recorded in the metadata, it had to be transformed. This meant agreeing on how data information would be stored in the metatdata so that the indexing and display software could distinguish it from other date information in the metadata. It also required coming to agreement on the ultimate representation of that data, specifically month abbreviations, the use of periods, and the ordering of each of the year, month and day components that together make up a complete date.

A secondary challenge with facets is how to indicate to users when they've made a choice, which choices they've made, and how to back out of those choices. This interface difficulty is created by a combination of limited screen real estate and likely visual overload. Our final solution, derived at through a detailed heuristic analysis by the CDL's User Assessment team, involved several pieces of functionality and some specific visual cues including: ensuring that the list of letters presented in the main frame of the site reflect the result of the current set of facet choices; presenting selected terms within facets in a line of labeled boxes (one per chosen facet) just above the main frame of results, each of which could be independently removed by clicking on an "X" box; greying out and italicizing the chosen value within a given facet as it appears in the left sidebar in the facet area; providing dynamically updated hit counts for all values each time a facet choice is made. The ultimate effect has been to make a close visual link between the user's choices and the subsequent letters that appear in the result set in the main pane. Testing with users indicated that individuals understood how to use these tools and that the resulting sets of letters were what they were expecting.

3.2 Essential Features: Advanced Search

Search activities are becoming increasingly simplified as online applications strive to provide the stripped-down Google experience that so many people are now expecting. For some academic work these searches may be sufficient, especially for those newly exploring a discipline and are attempting to map out an unfamiliar terrain. Such an approach can also be strewn with invisible difficulties, as signficant work within a given area may not appear in results generated by commercial search engines, a liklihood increased by the lack of knowledge of key concepts and central individuals. For more advanced scholars who bring more depth and breadth to their research, search tools must be calibrated more finely to support more

precision in working through a collection. Such precision offers scholars search results that are more likely to be relevant to their research interests as opposed to having a mere coincidental relation to each other.

The issue of expertise extends to both knowledge of a domain and facility with online discovery tools, of which search tools are no doubt the most heavily used. Wildemuth [34], in a study of the efficacy of searches by medical students as they progress through their education, has shown that domain expertise is tied to the ability to form more precise, complex searches, thereby producing more satisfactory search results. In a study that adds another dimension to the question of expertise, Hsieh-Yee [35] has shown that not only does search experience trump domain knowledge in effective search practices, but that search experience is a key to effectively employing domain expertise, finding that novice searchers were unlikely to use their subject knowledge (as expressed by synonyms, etc.) to improve and enhance searches, whereas those individuals with expertise in both the field in question and online searching did avail themselves of their domain knowledge. In a confirming study, Bhavnani [36] observed individuals with online skills and expertise in healthcare or online shopping search both in and outside their areas of knowledge, revealing that domain expertise combined with search expertise yields better results, particularly because of the ability to effectively use domain specific tools. One explanation for such findings is that naïve searchers presumably do not have a good grasp on how the systems that they are searching actually work, and therefore are unable to usefully bring their domain knowledge into play. Palmer [37] notes that humanities scholars gravitate towards indexing tools within their subfields, but need help finding online search tools that are robust and precise in their fields. Turning this around, digital tools built for humanities scholars must at a minimum provide the same kind of utility that scholars rely on with print resources. These findings presents a compelling argument for translating the intellectual structure of content within a domain (e.g. "text," "editorial notes," etc. in the case of literary scholarship) in a way that is obvious to knowledgeable users within that field and learnable on the part of those with little or no domain expertise.

With this body of knowledge pointing the way, the MTPO team faced the challenge of providing researchers with a search framework appropriate to their discipline and that could accommodate, and perhaps nurture, a range of online searching skills. We wanted to allow scholars to distinguish among the identifiable pieces of the edition they would need to search independently without overwhelming any user with overly precise, lengthy lists of options. The difficulty resided in determining how to translate those distinctions into a managable set of meaningful "content slices" that would fit on a small search form, and then how to implement the search technology to support the complex queries users would want to create.

Providing search access to the essential descriptive fields of the letters (i.e., the "envelope" attributes of sender, address, and place of origin) was a clearcut decision, particularly since it provides users the opportunity to distinguish between sender and addressee, something not possible by browsing with the facets. However identifying distinct components of the edition was much more difficult, becoming again a balancing act between adhering to editorially integrity, which meant specificity, and usability for both literary, as opposed to textual, scholars and lay enthusiasts, which meant larger groupings. Ultimately this required the grouping of very well defined pieces of the editorial matter into somewhat larger buckets. For instance, the editors were clear that Clemens' own text needed to be distinctly searchable apart from any of their own work. Secondly, as part of maintaining and exemplifying the standards of textual scholarship, the distinction between "Explanatory Notes," which provide contextual information and the "Textual Apparatus," which provide information such as provenance and emendations, also needed to be retained, as each offer unique sets of data to which scholars of all sorts would require separate access.

An example can demonstrate how these "content slices" serve the needs of the end-user, (expert or lay), addressing and validating the concerns of the editors. For instance, a scholar who was interested in knowing more about the physical condition of the letters might want to know about certain types of damage, which letters were known to have be torn or have sections missing, for example, a physical

attribute that would be captured by an archive. A simple keyword search on the word "missing," which searches a combination of all metadata fields, original and editorial text, returns 143 results, too many to look through in a reasonable amount of time. Searching on that same term in the various content slices discussed above demonstrates how that term appears with different meanings depending upon what is being searched. Looking for the term within the original texts retrieves 56 letters in which there are both inline notes from the editors indicating missing portions, but also such letters as Clemens' 17 January 1869 letter to his wife Olivia in which he notes that "Your Iowa City letter came near missing—it arrived in the same train with me." [38] This use of the word is, of course, entirely different than what the researcher in this scenario would be interested in. Further, while 56 letters is a more manageable set than 143, it is still quite a lot to work through. Searching only within the Textual Apparatus returns a more tractable 27 letters, which when examined include exactly the sort of letters pertinant to this scenario, as reflected in such critical notes as this, from an October 1865 letter from Clemens to Orion and Molly Clemens [39]: "All four leaves are creased and chipped, especially along the right edge where one crucial fragment is missing (324.34)."

While the facets of MTPO discussed in the previous section present a structure in which researchers can browse the edition, the advanced search framework, by separating the various components of a critical edition, provides scholars the ability to uncover information related to their more individually tailored concerns. Serendiptious findings are possible in both settings, but a primary goal of such a search tool is the support of targetted exploration of a collection.

3.3 Essential Features: Citation Tools

Any scholarly artifact must be citable. Without the ability to reliably refer to a specific "location" that others can independently access, the conversation among scholars is halted. Print material and citation structures are closely married, as evidenced by the numerous standards for referring to any manner of printed item. With the rise of online primary and secondary source material, conventions for referring to resources on websites are beginning to be established, as indicated by The Modern Language Association Guidelines that insist that "Authors of scholarly writing on the Web should number paragraphs or other sections in texts of significant length so exact locations may be cited." [40]

Accompanying the greater frequency of online sources requiring citation has come the development of tools to ease the generation and management of those references. These tools range from informal folksonomy tagging tools such as "del.icio.us" [41] to more formal but still web-based, web-friendly bibliographic management systems such as CiteULike [42] and Zotero [43] that can produce citations adherent to well known standards such as The Chicago Manual of Style [44] or the MLA Handbook for Writers of Research Papers MLA [45]. Palmer [46] explored the issues presented by bibliographic web services in the development of his CiTEX citation management, identifying the importance of persistent identifiers such as DOIs and Handles and the need for more work to eliminate tedious copying and pasting. Similarly, Hitchcock et al [47] in their work on the Open Citation project discuss how in the web environment there are increasing possibilities for citations to become actionable reference links, positing that such a feature will become essentialized as part of scholarly publication and communication. Bier et al [48], who have taken web-based citation management one step further by building a document-centric research support tool, argue that citation extraction and management is central to supporting the overall research efforts of scholars, which include such activities as in-depth reading, following reference chains, and tracking which documents have or have not been read. Indeed, many of the existing digital critical editions provide citation support at the top level of the object, either focusing on providing persistent URLS, as in the case of HyperNietzsche's predictable URL structure for referring to page numbers within a text; the Carlyle Letters, which supports the output citation data in the formats required by a number of standard citation managers (e.g. EndNote and BibTeX); and Collex/NINES, which suports

citations of the Collex based search results.

None of those conventions or approaches, however, is sufficient for dealing with digital texts that lack attributes transferred from print such as page numbers or that are likely to be modified in a non-linear manner as new material comes to light, thus eliminating the possibility of stable paragraph or line numbers. Thus the MTPO team was confronted with the need to support citations for individual texts that would likely expand but whose previous components would remain stable. An additional concern was the issue of monograph length works that had no usable markers from the print world leaving only the most easily attainable citations for individual chapters or the entire work, either of which were unacceptable in that they are at too coarse of a level to support scholarly arguments in a comfortable fashion.

MTPO addressed this complexity by developing the notion of "citable chunks," which meant determining which encoded chunks that could be automatically identified by the system should have an associated citation widget that when clicked, would automatically generate a reference that included a persistent URL to that exact location. Enabling this involved determining at which level such chunks made sense, from the perspective of scholars who would want to use them; from the perspective of the technology pieces that would have to come together to generate the citations; and from the perspective of end-users who would have to understand how to use the citation tool and not have it interfere with their work with the texts.

Because MTPO's first release includes letters as opposed to the longer works, the decisions were somewhat clearer. For the letters, users can click on an icon to gather a citation for a letter in its entirety and can do the same for editorial notes at the individual note level. In the next phase of MTPO, the infrastructure is in place to support citations at the paragraph level, which could provide useful precision for citing scholars, but could also introduce some interesting interface challenges as one imagines a page of dialog strewn with citation widget symbols at the beginning of nearly every line. Still, as there are ways solutions to address this issue (e.g. enabling users to turn off and on the citation widgets), our tendency at the moment is towards supporting the greatest degree in citation precision, as this seems to be a significant advantage of a digital publication.

Generating citations is the first part of the equation especially for original sources for which the particular citation structure either may not yet exist and even if it does, is not as easily remembered as the traditional (online or offline) journal article. For MTPO for instance, the editors devised citation structures for the various types of texts that could be referenced, ranging from the letters themselves to records of letters to editorial notes. A the development of new web-based reference managment tools indicates, in addition to having access to acceptable citation structures, scholars need to be assisted in tracking what items they are interested in, either before or after having read them. To that end, MTPO includes the ability to gather citations of interest to a web page within the system (called "My Folder") that is available during any individual session. Citations can be added and removed, seen in a compact or full format, and emailed to oneself or any other interested party.

Offering tools to help researchers as they work with the texts is yet another distinguishing feature that can set digital critical editions apart from their print-based antecedents. At this point, development teams have the option of supplying those tools within the sites they are building, as MTPO chose to do, or to integrate with generic reference management tools. The best choice, of course, is to make no choice at all, but rather to build sites that produce appropriate citation structures, offer good tools for scholars to use while doing their research on the site and that can also be used with the predominant tools in existance, which is where MTPO hopes to be in future phases.

3.4 Essential Features: Complex Object Views

MTPO, like other digital critical editions, faced the challenge of how to represent the overlapping layers of detailed, precise information that together make up a discrete work within the edition. For MTPO, a letter was the defining object in question and the pieces that need to be made available and put under the control of the reader qua user were transcriptions, facsilimies, editorial notes, the textual apparatus, and other related editorial material such as the descriptions of the text and provenance, and the guide to editorial signs

Similar to the manuscripts presented in Gallica Proust [49], letters were presented in two left and right panes. In MTPO, the larger left pane always contains the transcribed text, which is consistently available for any letter for which there is text. The narrower right pane defaults to a facsmilie, if it exists, and if not, then to a display of the Editorial Notes and Textual Apparatus that are associated with that letter. For letters with a facsmile, users can magnify the image for closer inspection and can also select a link to switch back and forth in the right pane between the manuscript image and the editorial matter. In the notes view, when users glide the mouse over the transcribed text, notes and apparatus entries associated with that text will be shaded in the right-hand pane. If a user begins to read one note and then decides to read further, clicking on an editorial note will sync up the reading pane with the portion of the text associated with that note or entry and for the apparatus entires only, will additionally shade the portion of the text to which the apparatus entry applies. The asynchronous movement of each of the panes supports well-known research habits of perusing beyond the original note of interest, yet another way serendiptious research behaviours can be supported in the online environment. Additionally, as desired users can call forth in a secondary window a "print view" of the letter with all of the Expalanatory Notes and Textual Apparatus entries appended to the end. Finally, from the letter view, a researcher can add a citation for the letter on screen at the moment to his "My Folder" page and can also wander backwards or forwards along the editorially determined sequence of letters.

Such functionality transfers as much control over the reading experience as possible. All the components of the edition are made apparant and available to the user whenever she may require them without having that information overwhelm the central text itself.

4. Conclusion

Scholarly practices are being transplanted, re-imagined, and newly developed as the digital environment expands in both depth and breadth, becoming a richer arena in which scholars can discover and work with pimary and secondary resources. Researchers are gaining a better understanding of how to use existing tools to work within this dimension and at the same time, new tools and infrastructure are being dreamed up and crafted to aid them. Per Hillesund, the digital text cycle has not wholly overtaken the print text cycle, but advancements toward that end are being made. Digital critical editions such as MTPO are examples of the transitional works and associated embedded tools and practices that represent significant movement towards that new paradigm. Although the data the edition publishes online was originally created for print presentation, it is now stored in such a way that many alternate presentations and uses are possible, as evidenced by the publication of MTPO itself.

As an original online edition and not simply an electronic print reproduction, MTPO offers unique access to a rich, dynamic collection of authoritative material that provides scholars easier entry into that material from many vantage points, easier means in which to closely inspect it, and easier methods for recording the trail of that investigation. Facets and advanced search together provide a rich set of ways for researchers to explore and discover the material and information in which they are or might be interested. By providing the ability to browse the collection via elemental dimensions and by exposing the structural elements of textual scholarship, MTPO supports both serendiptious discovery and more precise research strategies.

Bringing these approaches together, by allowing individuals to avail themselves of that overall structure to further winnow down their findings as they traverse the collection according to their own terms of interest, MTPO supports the dynamic and creative aspects of research that draw on the expertise of scholars within a field. By presenting a visually managable, interactive composite view of the many layers of information which together comprise an individual work (i.e. letter) within the edition, MTPO provides a richer and easier environment for detailed inspection of the content. Finally, by creating automatically generated, reliable citation formats for the unique materialsin the edition, and a means for gathering those citations and saving them (through emailing them), MTPO provides a means to safeguard and track the results of valuable, often irreproducable, research efforts.

To briefly recapitulate, facets create intellectual structure and help to maintain the persistent integrity of the edition as it appears to users over time; advanced search and support serendipity provides opportunities for researchers to apply their own framework; citations are the outcome of these together and serve to secure and record those outcomes; and complex displays of objects allow detailed inspection by collocating the pieces of the authoritative object. These features together maintain the integrity and stability of the collection, while concurrently allowing for fluidity in the continued expansion of the material. In this way, MTPO hopes to succeed as a digital critical edition that will support and extend the research activities of scholars.

5. Acknowledgements

A great many people contributed over several years to the ultimate publication of MTPO, a complete listing of whom can be found on the site [50]. However, a core project team, of which I was one member, worked very closely together during the year and a half leading up to publication. I would like to directly acknowledge those individuals, as the result of their work is what made this paper possible: Laura Cerruti, Erim Foster, Sharon Goetz, Benjamin Griffin, Kirk Hastings, Catherine Mitchell, and Leslie Myrick.

6. Notes and References

[1] ROBINSON, Peter. Current issues in making digital editions of medieval texts—or, do electronic scholarly editions have a future? *Digital Medievalist* [online]. Spring 2005, vol. 1, no. 1 [cited 28 April 2008]. Available from Internet: <http://www.digitalmedievalist.org/journal/1.1/robinson/>. ISSN: 1715-0736.

[2] http://www.canterburytalesproject.org/

[3] D'IORIO, Paolo. Nietzsche on New Paths: The HyperNietzsche Project and Open Scholarship on the Web. [online]. [cited 1 April 2008]. In FIORINI, MC., and FRANZESE, S. *Friedrich Nietzsche. Edizioni e Interpretazioni*. Pisa : ETS, 2006. Available from Internet: <http://www.hypernietzsche.org/doc/doc.html>

[4] *Mark Twain Project Online* [online]. Edited by Fischer, Victor, Frank, Michael B, and Smither, Harriet Elinor. Berkeley: University of California Press, 2007 [cited 28 April 2008]. Available from the Internet: <http://www.marktwainproject.org>.

[5] http://bancroft.berkeley.edu/MTP/

[6] http://www.ucpress.edu/

[7] http://www.cdlib.org

[8] www.tei-c.org/

[9] TANSELLE, G.Thomas. Critical Editions, Hypertexts and Genetic Criticism. *The Romantic Review.*

1995, vol 86, p. 581-593.

[10] GUNDER, Anna. 2002: Forming the Text, Performing the Work - Aspects of media, navigation, and linking. *Human IT*. [online]. February – March, 2001.[cited 26 April 2008]. Available from Internet: <http://www.hb.se/bhs/ith/23-01/ag.htm>. ISSN: 1402-150X.

[11] HILLESUND, Terje. Digital Text Cycles: From Medieval Manuscripts to Modern Markup. *Journal of Digital Information.* [online] 2006, vol 6, no. 1 [cited 26 April 2008]. Available from Internet: <http://journals.tdl.org/jodi/article/view/jodi-164/65>. ISSN: 1368-7506.

[12] *HyperNietzsche*. [online]. [cited 1 April 2008]. Available from Internet: <http://www.hypernietzsche.org>.

[13] http://interedition.huygensinstituut.nl/?page_id=2

[14] http://nines.org

[15] http://nines.org/collex

[16] NOWVISKIE, Bethany. A Scholar's Guide to Research, Collaboration, and Publication in NINES. *Romanticism and Victorianism on the Net*. [online]. August 2007. vol 47. [cited 5 May 2008]. Available from: <http://www.erudit.org/revue/ravon/2007/v/n47/016707ar.html>. ISSN: 1916-1441.

[17] The Complete Work of Charles Darwin Online. *[online]. Cambridge:* University of Cambridge*, 2008. [cited 28 April 2008]. Available from Internet:* <http://darwin-online.org.uk/>.

[18] *The Complete Writings and Pictures of Dante Gabriel Rossetti*. [online]. Edited by McGann, Jerome J. [cited 28 April 2008]. Charlottseville: IATH, in production. Available from Internet: <http://www.rossettiarchive.org>.

[19] *The Walt Whitman Archive.* [online].edited by Folsom, Ed and Price, KennethLincoln, NB: Center for Digital Research in the Humanities, 15 February 2007. [cited 28 April 2008]. Available from Internet: <http://www.whitmanarchive.org/>.

[20] The William Blake Archive. [online]. edited by Eaves, Morris, Essick, Robert N. And Visomi, Viscomi. 13 November 1997 [cited 28 April 2008]. Available from Internet: <http://www.blakearchive.org/>.

[21] NOWVISKIE, B. 2000. Interfacing the Edition. [cited 5 May 2008] Available from Internet: <http://www.iath.virginia.edu/~bpn2f/1866/interface.html>.

[22] *The Carlyle Letters Online.* [online]. edited by Kinser, Brent E. [cited 28, April 2008]. Available from Internet: <http://carlyleletters.dukejournals.org/>.

[23] The Making of Mark Twain Project Online. *Mark Twain Project Online* [online]. Berkeley: University of California Press, 2007 [cited 28 April 2008]. Available from Internet: <http://www.marktwainproject.org/about_makingMTPO.shtml>.

[24] *Mark Twain Project Online* [online]. Berkeley: University of California Press, 2007 [cited 28 April 2008]. Available from Internet: <http://www.marktwainproject.org/about_technicalsummary.shtml#contentTransform>.

[25] RANGANATHAN, S. R. *Colon Classification*. 7th ed. Bangalore: Sarada Ranganathan Endowment for Library Science, 1987.

[26] BROUGHTON, Vanda. Faceted classification as a basis for knowledge organization in a digital environment; the Bliss Bibliographic Classification as a model for vocabulary management and the creation of multidimensional knowledge structures. *New Review of Hypermedia and Multimedia*. 2001, vol 7, no. 1, p. 67-102.

[27] HEARST, Marti A. Clustering versus faceted categories for information exploration. *Communications of the ACM*. 2006, vol 49, no. 4, p. 59-61.

[28] __ et al. Finding the flow in web site search. *Communications of the ACM.* 2002, vol 45, no. 9, p. 45.

[29] YEE, Ka-Ping, SWEARINGEN, Kirsten, LI, Kevin, et al. Faceted metadata for image search and browsing. In *CHI '03: Proceedings of the conference on Human factors in computing systems*. 2003, p. 401-408.

[30] BEN-YITZHAK, Ori et al. Beyond basic faceted search. In WSDM '08: Proceedings of the international conference on Web search and web data mining. 2008, p. 39.

[31] XTF is a full-text and metadata search and display technology, initially focused on documents, and built on top of the open source Lucene search index. For MTPO, that index was built out of the full-text of the transcribed TEI encoded letters, and the METS encoded metadata. For more information, please refer to <http://www.cdlib.org/inside/projects/xtf/>.

[32] SLC to Olivia Lewis (Mrs. Jervis) Langdon, 13 Feb 1869, Ravenna, Ohio (UCCL 00249). [online] In *Mark Twain's Letters, 1869*. Edited by FISCHER, Victor Fischer, FRANK, Micahel B. and ARMON, Dahlia. *Mark Twain Project Online* Berkeley: University of California Press. 1992, 2007. [cited 5 May 2008]. Available from Internet: <http://www.marktwainproject.org/xtf/view?docId=letters/UCCL00249.xml;style=letter;brand=mtp>.

[33] http://www.w3.org/TR/NOTE-datetime

[34] WILDEMUTH, Barbara M. 2004. The effects of domain knowledge on search tactic formulation. *Journal of the American Society of Information Science and Technology*.2004, vol 55, no. 3, p. 246-258.

[35] HSIEH-YEE, Ingrid. Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information Science*. 1999, vol, 44, no. 3, p. 161-174.

[36] BHAVNANI, Suresh K. Domain-specific search strategies for the effective retrieval of healthcare and shopping information. *CHI '02 extended abstracts on Human factors in computing systems*. Minneapolis, MN: ACM. 2002, p. 610-611.

[37] PALMER, Carole L. Scholarly work and the shaping of digital access: Research Articles. *Journal of the American Society of Information Science and Technology.* 2005, vol 56, no. 11, p. 1140-1153.

[38] SLC to Olivia L. Langdon ... , 17 Jan 1869, Chicago, Ill. (UCCL 00234). [online] In *Mark Twain's Letters, 1869*. Edited by , Victor Fischer, FRANK, Micahel B. and ARMON, Dahlia. *Mark Twain Project Online*. Berkeley: University of California Press. 1992, 2007. [cited 5 May 2008]. Available from Internet: <http://www.marktwainproject.org/xtf/view?docId=letters/UCCL00234.xml;style=letter;brand=mtp>

[39] SLC to Orion and Mary E. (Mollie) Clemens, 19 and 20 Oct 1865, San Francisco, Calif. (UCCL 00092). [online] In *Mark Twain's Letters, 1853-1866*. Edited by BRANCH, Edgar Marquess, FRANK, Michael B., SANDERSON, Kenneth M. et al. *Mark Twain Project Online*. Berkeley: University of California Press. 1988, 2007. [cited 5 May 2008]. Available from Internet: <http://www.marktwainproject.org/xtf/view?docId=letters/UCCL00092.xml;style=letter;brand=mtp>.

[40] Minimal Guidelines for Authors of Web Pages. [online]. Modern Language Association. [cited 28 April 2008]. Available from Internet: <http://www.mla.org/web_guidelines>.

[41] http://del.icio.us/

[42] http://www.citeulike.org

[43] http://www.zotero.org/

[44] *The Chicago Manual of Style*. 15th Ed. Chicago: University of Chicago Press, 2003.

[45] GIBALDI, J. *MLA Handbook for Writers of Research Papers*. New York: Modern Language Association of America, 2003.

[46] PALMER, James D. Exploiting bibliographic web services with CiTeX. *Proceedings of the 2007 ACM symposium on Applied computing*. Seoul, Korea: ACM. 2007, p. 1673-1676.

[47] HITCHCOCK, Steve et al.. Developing services for open eprint archives: globalisation, integration and the impact of links. *Proceedings of the fifth ACM conference on Digital libraries*. San Antonio, Texas, United States: ACM. 2000, p. 143-151.

[48] BIER, Eric, GOOD, Lance, POPAT, Kris et al. A document corpus browser for in-depth reading. *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*. Tuscon, AZ, USA: ACM. 2004, p. 87-96.

[49] *Gallica Proust*. [online] Edited by CALLU, Florence. [cited 28 April 2008]. Available from Internet: <http://gallica.bnf.fr/proust/>.

[50] Contributor Credits. *Mark Twain Project Online* [online]. Berkeley: University of California Press, 2007 [cited 28 April 2008]. Available from Internet: <http://www.marktwainproject.org/about_contributorcredits.shtml>.