Digital Humanities in Early Online Archives

Lela Ellison
Katie Risseeauw
Walker Sampson

University of Texas at Austin
Digital Humanities in Early Online Archives

Recent research in humanities is taking on new depth through more interactive and collaborative research projects using materials physically hard to access. This creates new contexts for previously static texts and images. Compared to studying Walt Whitman's work, evaluating Rossetti's art and poetry, or moving through the finding aids of an archive's Civil War records thirty years ago, humanist scholars in the past fifteen years have more resources due to the development of online digital collections published on the World Wide Web. In the early 1990s, the establishment of a series of digital collections first showed the promise and potential of improving research and to make a lasting impact on humanities scholarship.

History

Even though the rapid expansion of electronic formats and digital archive collections found on the World Wide Web is recent, the attempt to merge computers and humanities has been an ongoing effort since Father Roberto Busa's concordance project *Index Thomisticus* began in 1946. A daunting linguistic analysis of the written words of St. Thomas Aquinas, including conjunctions, prepositions and pronouns, Busa sought automation to speed up the pace of the project (Schriebman et al., 2007). The first foray into merging the humanist studies with an automated machine created a field of study that would be revisited through the next half century.

It was not until three basic requirements were fulfilled that computer technology became instilled in the web of literary and humanities scholarship: 1) advanced computing equipment that supplied greater computing power and accessibility, 2) sophisticated software that let the user configure the terms of analysis, and 3) access to high quality texts and material. When these circumstances manifested themselves, there was a corresponding rise of digital collections that sought to transform the methodology of traditional scholarly research into a dynamic and interactive system that allowed new types of scholarship.
Early computer projects, first centering on concordance works, progressed to more complicated textual studies. Record keeping and record making were the first uses of computing in the humanities. Generating lists of textual patterns that acted as supplementary support for research was the closest computing came to the facilitating substantive scholarship. It was not until the appearance of the personal computer and email that the humanities field began to entertain the possibility of a new kind of scholarly work within their methodologies, one specific to the digital medium. The development of email lent itself to increased collaboration, which resulted in even more complex projects. Scholars were no longer as restricted by the limitations of real-time communication, and the ease of asynchronous collaboration encouraged more expansive interdisciplinary research. Humanities scholars and computer scientists regularly collaborated to develop programs and software designed for humanist research.

Increased computing led to a wider effort to create standards and rules to instill some kind of methodology into the display of text, its meaning and relationships on an automated machine. At this time a number of markup languages were adapted by scholars in any grammar that could suit their needs. The differing approaches to markup led to confusion and a general feeling that one must repeat considerable rote work with each digital project. This naturally hindered interoperability.

Standard Generalized Markup Language (SGML), whose first draft was published in 1980, is a text description language standard based on the Generalized Markup Language ("SGML User's Group," 1990). Subsequent drafts became ANSI and ISO standards. SGML created the perimeters of humanities related markup language, and with its ability to feature multiple nested element sets, it was flexible enough to be adapted for a variety of subjects. SGML paved the way for a major development in humanities computing: the Text Encoding Initiative (TEI), whose first official release came in 1994. TEI is an encoding scheme for texts that was initially expressed in SGML and is presently expressed in XML. The standard coincided with a slew of libraries posting digital facsimiles of their collections online, many using TEI from the start. Of course, these developments did not preclude persistent interoperability problems stemming from both limitations of SGML and XML when applied to the
idiosyncratic humanist and creative texts. However, their utility as unifying and adequately flexible frameworks for machine readable markup was immense.

The development of HyperCard, a model of computerized “file cards” linked in specific ways that allowed the user to skip around a collection, held the limelight as a development and teaching tool that promoted new ways of accessing literature. The different aspects of a literary work, the author, the history, and other associated information could be displayed non-linearly from a centralized location. An example of a HyperCard project is the *Beowulf* Workstation, established in 1991. The *Beowulf* Workstation featured the original Anglo-Saxon version of the epic with a link to the modern English version, accompanied by linguistic and contextual annotations from various Beowulf experts. This was designed to aid students in preparing translations and explications of the poem (Connor, 1991). This anticipated present day websites by popularizing the notion and use of hypermedia (Schriebman et al., 2007). Jerome McGann, director of The Rossetti Archive, states, “E-text completely redefines our textual limits” and is a “radical re-imagination of textual space” (1994). For the first time, the user of such a collection could map out their own paths of discovery within an archival setting. The HyperCard projects gave way to a more expansive and accessible medium: the Internet, and more specifically, the Wide World Web.

The formation of TEI along with these pioneering projects worked together to create new digital endeavors that arose with popularity of the Wide World Web. Initially, humanist scholars were at a loss in harnessing the new platform to benefit their own research. The publishing potential of the Web however was not difficult to perceive. The Web's ease in reaching a large number of people and the sheer flexibility of the media itself proved advantageous for creating an audience for humanist scholarly research, which in turn inspired researchers in the humanities to attempt resolve accessibility problems of special collections by establishing digital facsimiles of their collections. Edward Ayers, a professor at the University of Virginia, was one the first humanities scholar to begin work on a digital thematic collection with *The Valley of the Shadow*. This was a database of primary sources from two communities, Augusta County, Virginia and Franklin County, Pennsylvania during the Civil War era. *The Valley of the Shadow* was one of the first works of historical scholarship displayed on the Web; its
open access approach attracted a wide range of users and the project became a model for other digital humanities projects.

Success of The Valley of the Shadow was shared by other digital collections such as The Rossetti Archive, The Walt Whitman Archive, and The Blake Archive, among other similar projects. Like Valley these digital collections featured electronic reproductions of archival material that centered on a writer or topic. These collections resolved a major issue many humanities scholar grappled with: accessibility. Due to the physical dispersion but dense textual relationships between Whitman's works, establishing a single source where most of his work is displayed proved a significant boon for scholars. His system of incessant revision made it especially difficult for scholars to adequately detect relationships in his work (Barney, et al., 2005).

Another area that early online digital collections introduced to humanities was, as Andrew Jewell terms it, the "responsivity" of text and the fact that through featuring a digital collection online, all users become potential editors and collaborators (Jewell, 2005). Interactive and nonlinear use of digital collections allowed users to experiment with text. Tools were developed to enable scholars to create tags, notes, or manipulate the picture or text while carrying out research. The Willa Cather Archive uses Token X, a text analysis tool that presents text as data, grouping them into columns or word clouds that highlight patterns (Jewell, 2008). McGann sought to depart from the traditional book model by creating tools that allowed for image manipulation and annotation (McGann, 1994). In short, once the interactive approach was introduced to scholars through the establishment of these early online digital archives, it made an impact on humanities research as a whole.

With some time, the complexities of sustaining a digital collection became apparent in the late nineties. Decline in interest stems from technological limitations and the labor involved in creating custom tools. Morris Eaves, an English professor who worked on the William Blake Archive, stated that at the beginning of the project he and his collaborators all had “technical optimism” towards developing the Willa Cather Archive and how that optimism quickly diminished as they struggled with communication issues and the prevailing attitude that their project was not as important as fulfilling professional obligations in the tenure track (1997).
Another limitation that curbed the spread of digital collections was copyright. Jerome McGann discusses in detail how copyright costs and the reluctance of some libraries to submit Rossetti’s work drove him to change his digitizing ambitions (1997). Rather than exhibiting digital facsimiles made from fresh full color images done by Rossetti, exhibiting photographs of Rossetti’s work taken by his friends and contemporaries during and shortly after his lifetime were shown. The Willa Cather Archive may not presently show any of Cather’s post-1922 work, which leaves out some of her major works of literature (Jewell, 2009). In both cases, the research teams developed workarounds.

The flexibility in interacting with works in the digital format and ease of access has made the digital collections a viable tool for humanities research. Next we can look at the motivations of scholars in developing early digital online collections and how this informed development.

**Purpose and Audience**

A major factor in unifying new technology and humanist interests was the Institute for the Advanced Technology in the Humanities (IATH) created at the University of Virginia in 1992. Its mission has been to “explore and develop information technology as a tool for scholarly humanities research” (“Institute for the Advanced Technology in the Humanities,” n.d.). At the time of its formation, the institute joined Jerome McGann and Edward Ayers and their projects The Rossetti Archive and *The Valley of the Shadow*. Within a few years many of the successful online archives from this period were associated with IATH and still are, such as the Blake, Dickinson, and Whitman Archives. This commonality extends to the similar reasons guiding the archives into existence; the ability to connect physically disparate collections and view original manuscripts to enhance scholarly research was a driving force in the conception of these projects.

*The Rossetti Archive*
When Jerome McGann first envisioned a digital version of Dante Gabriel Rossetti’s poems and artworks, he built the archive with the assumption that someday all libraries and museums would be online (McGann, 2008). Thinking this way required designers to consider their infrastructure and creating the archive not “as a social space and system” (2008). He wanted the archive to address both trends in the new electronic data projects, that of hypertext and hypermedia linked for browsing and those created for "in-depth and structured search and analysis of the data" (McGann 1997). In a paper from five years after The Rossetti Archive’s inception, McGann admits this dual purpose was quite difficult to maintain, but necessary not only for scholarship but for the technological discoveries (1997).

The Rossetti Archive includes all the Pre-Raphaelite’s written and pictorial works as well as any typographical reproductions. This content decision stems from McGann’s scholarly interests in textual criticism. Comparison of differing texts is naturally promoted when many versions of the same work are represented together. McGann sees the archive as a tool to analyze critical editions and facsimiles through the visual features of text and the context, the "social intercourse of texts" (McGann, 2001).

*The Valley of the Shadow*

Edward Ayers’ conceptual organization for *The Valley of the Shadow* project began as a book idea; he was interested in the comparative narrative between two enemy communities during the Civil War. Open to the scholarship potential by merging technology in the humanities, Ayers realized his vision of comparing the historical artifacts of a Northern town and a Southern town had great potential for digital reproduction and online sharing (Ayers 2004). *The Valley of the Shadow* and Ayers' writing both recognize the research possibilities of online collections and the benefits this will bring to the humanities, should the discipline fully embrace the new technologies. The large and heterogeneous collection, the “mass of scattered records,” (Ayers, 2004) could live easily within the searchable database that was essentially the skeleton of the archive. Ayers apparently was not interested in aiming the archive at scholars necessarily; he saw the fact that the “archive could reach anywhere in the
world people could tie into the internet, a network expanding exponentially” as an amazing change for the project (“Story Behind the Valley Project,” n.d.). The Rossetti Archive and The Valley of the Shadow were the earliest attempts at collecting disparate information in one digital location, and McGann and Ayers began work with the new technologies having previously flushed out their ideas with a physical output in mind.

*William Blake Archive*

Collaboration was key in the Blake Archive, with founders from several different institutions coming together to begin the project. Already working on a print volume of Blake’s works, Joseph Viscomi from the University of North Carolina, Robert Essick from the University of California, Riverside, and Morris Eaves from the University of Rochester all came together to become the editors of the project after a conference in 1992 and meeting with Jerome McGann at the University of Virginia (Eaves, 1997). Funding from the Getty Grant Program and teaming with IATH led to planning and execution of the archive. Although the Getty was interested in funding the project especially as a learning experience to lay groundwork for other similar ventures (Eaves, 1997), the Archive did not include an “About” page on its site until 1999 detailing this information (“Plan of the Archive,” n.d.).

While the Blake Archive site now explains its history, methodology, and technical aspects, Viscomi details the archive’s inception and purpose in an article from 2002, written when the archive was in Phase Two of development. The reasons for starting the archive stem from the inadequacy of typographic reproductions of Blake’s works, and the problems with accessing the originals for study.

The nature of Blake’s integration of illustration and word does not lend itself to strictly textual editions of his works. Even when pictorially reproduced, Viscomi states that the quality was insufficient to offer insights to scholarship, or high-quality expensive facsimiles of the works were too expensive for most to own (2002). Another interesting point Viscomi makes is how usually the same plates were always reproduced, whereas Blake actually made several copies of an illustration, with slight differences owing to the hand watercoloring or reworking of an etching plate (2002). By collaborating with
institutions with William Blake collections, the Archive is able to show many versions of the same plate. This is a great advantage to researchers who might not otherwise be able to travel around to various libraries, allowing side-by-side comparison of different iterations of the same copper plate. From the beginning, the editors of the archive envisioned including as much of Blake’s output as possible, instead of the mere 20% of his corpus that was reproduced up to 1993 (Viscomi, 2002).

This inclusive view of the Blake Archive’s contents underlines the importance of scholarship to the Viscomi, Essick, and Eaves. Viscomi states that they “designed the site primarily with scholars in mind” with the hope that it will be “not merely useful but indispensable” (2002). The focus on the breadth of the collection and research possibilities certainly follows the Blake Archive’s purpose “to evolve along lines that will achieve the greatest possible coverage of the range of Blake’s work while at the same time maintaining the greatest possible degree of scholarly coherence” (“Principles”, 2009).

Dickinson Electronic Archive

The issue of textual reproduction is also a foundational reason for creating the Dickinson Electronic Archive (DEA). As the editor of the archive notes, Emily Dickinson did not publish her poems formally in a book, but did circulate her works in correspondence (Smith, 1997). Since Dickinson wanted her poems to be read in handwritten form, the Archive allows a modern audience to do just this, as well as view the poem along with her letters. Formed out of the Dickinson Editing Collective, the digital archive is interested in presenting users with access to all of Dickinson’s writings, both poems and letters, to create new connections between the objects, a hypertext reality. With the many typographical editions of Dickinson’s poems, nuances of handwriting and the evidences that original paper exposes (stains, erasures) disappear. The DEA is putting the poems back into their original context, since “the first century of public exposure Dickinson's writings were regularized to fit the technology of print, the electronic editions … place Dickinson’s own distribution method at the center of editorial praxes” (Smith, 1997).
The Dickinson Electronic Archive hopes to reach outside of the scholarly research sphere and engage a large audience, since “electronic editions will appeal to general readers interested in Dickinson’s manuscripts, to readers interested in artistic process, to readers interested in historical documents, to readers interested in holographic records of writers, to readers interested in nineteenth-century literary culture” (Smith, 1997).

The Walt Whitman Archive

Much like the Blake and Dickinson online archives, the Walt Whitman Archive is interested in compiling the manuscripts of the poet. Ken Price, a co-director of the Whitman Archive, states that “Whitman’s work was always being revised, was always in flux, and fixed forms of print do not adequately capture his incessant revisions” (Price, 2002). Because Whitman’s materials are located across the United States, not many researchers have studied his manuscripts due to access issues (Price, 2002) and this online archive is the first chance for scholarship on this level. The archive also includes modern and contemporary Whitman criticism and some reproductions of typographic editions of his works. Beyond the scholarship intent, Price notes general users are also enjoying the site, as evidenced in comments left online (1997).

The Willa Cather Archive

Access to Willa Cather’s writings is a main goal of the Willa Cather Archive, and Andrew Jewell believes “[d]igital scholarship will play an enormous role in Cather studies because it will prove to be the best medium for the material” (2005). Not just the “unparalleled access” will create a community of research, but the materials the archive is highlighting, sometimes due to copyright issues as mentioned, are lesser-known works from early in Cather’s career (Jewell, 2005). Again the disparate nature of the writer’s works is overcome in a digital archive form, when all texts and images are searchable and can be arranged within their own context. The archive is “interested in creating opportunities for the reader to interact with the text in ways that are impractical
or impossible in print” (Jewell, 2005). Scholars and researchers are the intended audience of this online resource. Jewell states his hope to create educational resources and other projects not initially envisioned at the archive’s establishment (2005).

**Technical Histories**

Although the purposes and conceptual layout of humanities archives are not dependent on technical tools, limitations or trends, the implementation process is very informed by these aspects. A significant part of the history of these archives is therefore in their implementation. Described here are technical histories of some of the prominent early digital humanity archives. To contextualize these technologies in the broader topic of digital curation and early digital humanities archives some reflections on the tools also follow.

*The Rossetti Archive*

Insights on the technical history of the Rossetti Archive come from Jerome McGann. SGML was used to structure the text, while facsimiles of primary documents would be display digitally. As McGann states, the decision to use SGML was not an easy one. Implementers were well aware of the limitation of its markup design. This limitation is succinctly put by McGann: “SGML processors have no aptitude for . . . markup of textual features that are concurrent but logically distinct” (1999). This led to an extension of the TEI (the standard was then encoded with SGML) that could deal with some of the features found in Rossetti’s manuscripts and graphic works. Nevertheless this extension did not resolve the fundamental limitation. SGML was chosen however because it was a standard, and developing a language to handle concurrent but logically distinct features was not the goal of the Rossetti Archive. Along with this DTD, termed the Rossetti Archive Commentary (RAC), three other DTDs were made for other critical areas of the archive: the Rossetti Archive Work (RAW), the Rossetti Archive Document (RAD), and the Rossetti Archive Picture (RAP). None of these three were based off TEI but instead entirely custom designed, a decision
McGann later regarded as misguided (Pitt & Unsworth, 1998).

Along with this effort a Java tool was developed to categorize and annotate images. This was Inote (http://www.iath.virginia.edu/inote/). McGann describes the features of the tool, noting critically that while it can provide an annotative overlay to images, such that users can read text (linked or no) specific to certain parts of an image, it does not conform to the same DTD (or any DTD) as do the texts of the archive. Therefore at the time of this writing SGML texts could not be mapped to images of those texts (1999). Inote was also used by the Blake Archive to annotate and markup its images (Viscomi, 2002). Here the results were more synchronous with the SGML texts because of the more textual nature of Blake’s plates (McGann, 1999).

*The Blake Archive*

As described on the archive’s own history page, the arrival of SGML and TEI, and the stability they represented in the face of persistent technological change, was a turning point in the conception of the Blake Archive. The first year of project work (1995) saw scanning of images and SGML markup of images. Work on the Blake DTD (BAD) began in 1996, this DTD saw its first official version completed that year. This DTD encodes all works in the archive and emphasizes description of Blake’s works as physical artifacts. At its inception it was a DTD expressed in SGML, built entirely from scratch. The archive’s plan document states an ambition to continue developing this DTD such that it can be the “standard for encoded electronic images that the TEI is for encoded electronic texts” ("Plan for Archive," n.d.). The Blake Archive uses TEI strictly for its textual material such as bibliographies, collection lists and *Erdman’s Complete Poetry and Prose of William Blake*. Like other archives examined here, managers of the project are presently working on a migration from SGML conforming DTDs to DTDs conforming to XML, including the extensive BAD DTD. Along with this change comes a switch in search engine technology, from DynaWeb to some engine capable of indexing and performing full text searches on XML documents.

Although the archive’s records are discoverable site such Romanticism on the Net and Romantic Circle, there is no mention of records exported to other repositories
for discovery there.

*The Valley of the Shadow*

Conceptually and technically, The Valley of the Shadow began as a networked database through which scholars could access dispersed texts regarding the Civil War. Planning had begun as early as 1991 when IBM offered to codevelop projects with UVa that would use networked computers in new applications (Alexander, 2001).

Since the archive was being actively built even before the first Web browser, Mosaic, was released in April 1993, it is safe to assume its original rendering was in straightforward HTML. Newspaper content was originally encoded in SGML in the summer of 1993 (“Story Behind the Valley,” n.d.). It was only after the team was introduced to the Mosaic browser and the Web that the idea of publishing material incrementally as it was encoded was conceived. This shortened the projected period to reach a usable site, and as Ayers described it the site initially appeared as a research report on the IATH website in 1993: “long stretches of text with a few grainy images, large red icons, numbers in blue, and brackets for hyperlinks” (2004, p. B24).

After the site had been published a CD-ROM version of the project that featured richer media and quicker response time was produced. This was *Valley of the Shadow: Two Communities in the American Civil War. Volume I: The Eve of War*, published by Norton in 2000. The book and CD-ROM package was well received. Reviewers did not considered the item a frivolous purchase despite a freely accessible web site, partially because of new content pertaining to the CD-ROM’s coverage area (antecedents to the war), and partially because of ease of use compared to Internet capacity at the time, i.e. no lag as a result of traffic (Brown, 2001; Kornblith, 2001; Lowe, 2003). However, the CD-ROM would point to resources only located on the site. In addition the site itself, at least in 2001, was not able to search across multiple databases or record areas with a single search query (Kornblith, 2001). The presence of a CD-ROM and some of the problems with the site attest to weaknesses in the Internet’s infrastructure at the time: low bandwidth, narrow access, and an inability to perform a single comprehensive search across the archive’s multiple record domains.
In 2002 a new project manager oversaw the site. At this point texts began the slow transition and re-encoding from SGML to XML, which finally allowed full text searching across multiple domains. Maps were encoding with Geographic Information Systems technology (GIS). During this phase extensive new materials were continually added to the archive ("Story Behind the Valley").

A final source can give us information on the technical nature of The Valley of the Shadow archive as it stood upon its completion in 2007 (“VCDH Celebrates,” n.d.). This is the Valley of the Shadow SDS Overview, a report by the Sustaining Digital Scholarship project, part of the umbrella Digital Curation Services at the University of Virginia Library. This ongoing project was aimed at moving The Valley of the Shadow onto the Library's content management environment for permanent retention. There were three main file types to process: HTML, XML and image. Processing including updating XML files to conformance, converting pages to XHTML, rescanning with better tools, and moving data tables from a PostgreSQL format to an XML format for later import into the Library’s own search technology, a Solr search engine running on an Apache Lucene instance. Among other reconstruction efforts, the site was reconfigured to be delivered through Apache Cocoon. This server software allowed the archive to compartmentalize operations based on function or content. Page rendering was largely switched from straightforward HTML to HTML through XML with Extensible Stylesheet Language Transformations (XSLT). These stylesheets were aimed at maintaining the site’s look and feel along with intellectual and relational properties ("Valley of the Shadow Technical Overview,” n.d.).

Regarding the site’s history, Ayers has said, “we kept tearing down the Valley Project and rebuilding it” (2004). The archive began on a simple technical foundation of HTML files and SGML marked-up content, but subsequent revisions added images, maps and new databases of material leading to some 12,000 files. The final project was a mix of HTML and XML, with scans that needed updating and an outmoded server architecture that had to be dropped so the site could be reconfigured with the Library’s server configuration. Despite the humble beginnings, the news piece announcing the completion of the project provides a good statistic to gauge the archive’s usage: on average, 450,000 unique visitors every day (“VCDH Celebrates”).
The Willa Cather Archive

This section cites an interview with Willa Cather Archive editor Andrew Jewell, conducted November 11, 2009.

The archive began as an individual project in the mid 1990s by an employee of the University of Nebraska Press. The site consisted of HTML files containing the text of some of the Cather’s works, along with some scholarly works on Cather the Press had published. The site was maintained by the University of Nebraska Press until 2000 or 2001, when the English department at the University of Nebraska-Lincoln became interested in developing the site into a more expansive project. Additional content was added to the site, but after an initial wave of activity the project languished. As described, the site at this point was “a fairly minimal site. It didn’t have a lot of original content, it just had electronic texts among other things.” In 2000 Kenneth Price, who had been working on the Walt Whitman Archive, was brought on to help move the project forward, and at this point the first XML texts and an initial stylesheet were introduced. When the Center for Digital Research in the Humanities at the University of Nebraska–Lincoln was initiated in 2004, there was renewed interest in the project. This is when the current editor came on board.

During this time the archive was heavily preoccupied moving its dominantly HTML markup to a fully XML-conforming TEI P4 (public release version 4) and styled with XSLT. TEI P4 had been released in June 2002 and was the first release to implement full XML support. Besides adopting a standards-based approach to markup, this revision offered the first full separation of content and form. The archive is considering TEI P5, released November 2007, but as Dr. Jewell has explained, it is hoped that for the time being the present implementation will work for the all archive’s needs. Included in this revision process was a move to the Apache Cocoon web application framework.

The archive is discoverable on the Networked Infrastructure for Nineteenth-century Electronic Scholarship (NINES) at UVa through Dublin Core records pointing to local resources on UNL servers. These DC records have been generated and submitted
in two waves, one in 2006-2007 and another recently this year. The archive has thousands of records discoverable on NINES. Future plans may consist of integrating the archive’s content with UNL’s own library catalog, Encore. From this location the records could be discoverable on OCLC.

The archive’s main content lives on one server, but underlying databases, such as a MySQL database that manages the annotations for the journalism section, are spread across several servers. The image gallery is managed by OCLC’s CONTENTdm software, and this content is located on a separate server as well.

Both AWStats and Google Analytics are used to measure and evaluate usage of the archive. Dr. Jewell has found both tools valuable, noting the benefit in observing that “this year there was about 80,000 people who came to this site. That’s a significance number for something like the Cather archive, of that content, that’s important. And being able to track Google Analytics, that they’re coming from all over the world, that several people from China or Japan or France or whoever is coming to this site and using it for an extended period.” Statistics that can point this out help defend the project and bolster project morale and purpose.

Because of the present copyright laws, all of Cather’s published work in and before 1922 are in the public domain, all work after 1922 are copyrighted. Post-1922 works are therefore only shown in relevant excerpts, perhaps twenty to thirty words, when a user queries a post-1922 text.

To characterize the Willa Cather Archive’s progression, we can say that there has been comprehensive push to standards-based transcription and metadata (TEI and DC), increasing use of middleware or server software (Apache Cocoon and CONTENTdm), increasing export of records for discovery in federated searches (NINES, perhaps Encore and OCLC), and adoption of the best practice of separating content and presentation (A. Jewell, personal communication, November 11, 2009).

The Walt Whitman Archive

Kenneth Price, editor of the archive along with Ed Folsom, recounts in a 2005 Mickle Street Review article, cited here from Walt Whitman Archive itself, that the site
began in earnest in 1995, at the College of William & Mary, and that the result was “an HTML site, what I'd now call a prototype for a serious site, though at the time we didn't think of it as a practice run” (2005). Most work was done by himself and a graduate student, both with “little technical expertise, little guidance, and very little access to the network—the English department at William & Mary had only one room, a kind of glorified study carrel that had network access” (2005). Regarding the use of HTML markup, it was not apparent to them, as it was not to the majority of web developers and users, that this rendering language for browsers would be insufficient for scholarly purposes, and indeed ill-suited for most content-explicit markup.

The project was quickly approached by Primary Source Media in 1996. The publisher was interested in marketing a CD-ROM version of the site, with more media and content currently available. Price agreed, with the result of both Major Authors on CD-ROM: Walt Whitman and Primary Source Media’s donation of the copyright texts of Leaves of Grass. Like The Valley of the Shadow, a thematic digital collection generated interest as a CD-ROM largely because of the limitations of bandwidth and access.

In 2000 it was decided to add only “structured data to the site and to gradually redo older HTML parts of the site, as time allowed” (Price, 2005). This second phase resulted in a site redesign that also improved navigation, along with the first batch of XML encoded texts. At this point Dr. Price moved to UNL to assist with the Willa Cather Archive.

The archive encodes its finding aid in the Encoded Archival Description (EAD). Two key challenges in deriving a unified finding aid in EAD were the differing aids furnished to the archive by other institutions and the idiosyncratic nature of Whitman’s creative production. When a finding aid from another institution is furnished in either paper or HTML form, the aid is converted into EAD. Finding aids already in EAD had to be revised according to the archive’s own implementation of the standard. In resolving the complexity of Whitman’s output, the implementers constructed two identifiers for items in the archive: a unique string of characters corresponding to a manuscript, and a work ID constituting an abstract work to which various published and unpublished instances are related (Barney et al., 2005).

The encoding of the site is presently defined by a Whitman DTD that extends the
TEI P4 standard. This extension deals with constructs unique to Walt Whitman’s corpus, such as densely interrelated manuscripts and his unique history of drafts, letters, trials, notebooks, and jottings. Metadata Encoding and Transmission Standard (METS) is used as a high-level organizing scheme that relates the archive’s various digital objects, such as an EAD record, a manuscript record, and a manuscript TIFF, to one another (“Technical Summary,” n.d.).

Apache server software disseminates all the archive’s documents from a single server at UNL, and separate daemons are used to compartmentalize different applications and processes. For instance an XML indexing software, Tamino, runs on its own daemon. A separate application runs XSLT conversions of XML for HTML viewing. PHP is used with relational databases (MySQL) to generate pages for “three major sections of the Archive: the Walt Whitman Encyclopedia entries, the Whitman Image Gallery, and the Bibliography of Scholarship” (“Technical Summary”).

We can see a similar arc between the Walt Whitman Archive and the Willa Cather Archive: a natural move from simple HTML rendering to a comprehensive use of XML on all levels (from content encoding in TEI, to record and finding aid encoding in EAD, to entity relationship encoding in METS), accompanying this revision a distinction between content and design (occurring again with the use of PHP and MySQL), multiple underlying relational databases, and compartmentalized middleware to manage different server tasks.

_Dickinson Electronic Archive_

As stated in a plan for the archive, the project had intended from the start to use XML-conforming TEI to structure their records. This DTD would be, like other archives seen here, an extension of TEI and thus constitute a unique grammar for the element set. Similar to the Blake Archive, TEI is under equipped to address physical properties: “As TEI has an inherent bias toward encoding logical and coherent structure, we have modified the TEI in order to privilege instead the physical features of the manuscript, while still allowing for the deep searching and sorting capabilities facilitated in encoding logical structural elements" (Smith, n.d.). Also stated in this document is a plan to move
HTML rendered pages into a XML/XSLT rendering. In line with these changes is projected switch to Tamino to compartmentalize the rendering and searching of records.

Summary of Tools

In the technical history of early digital humanities archives, there is a predictable migration away from straightforward HTML rendering to a final rendering conceived through XML files processed by XSLT conversions. This is seen in Willa Cather, Walt Whitman, and the Valley of the Shadow archives. A similar process is also seen in the SDS Final Report (2003). This report describes the Sustaining Digital Scholarship group’s work on the Salisbury Project at UVa, an archive of images and educative content for the architecture of the Salisbury Cathedral. The goal was to move the archive to the UV Library’s own repository for permanent keeping. In this case EAD files and DynaWeb stylesheets were converted into XSL stylesheets before placement in the repository. A second project in the report covers the migration of the Rossetti Archive into the Library’s archive. Here a high level METS expression was devised to describe relationships among the archive’s interrelated works. This allowed XML files to be generated from one XSL stylesheet, and these XML files successfully defined all the relationships and characteristics of the objects in the archive (SDS Final Report, 2003).

Early SGML transcriptions of textual content are common to early archives as well. This is found in the Rossetti, Emily Dickinson (“Detailed Description,” 1997), Blake and The Valley of the Shadow archives. These markups were usually based on TEI, but in every case significant extensions were made to the standard to accommodate features unique to the collection; in other cases wholly new DTDs were developed. Other trends seen in these archives is a marked move to separate content from form wherever possible, increasing server software sophistication, and a very gradual exportation of metadata records to other repositories.
Conclusion

With the rapid pace of technical and networking advancement, digital humanities archives have seen enough history to begin considering the very long term curation of both their own records and the ongoing record of their existence. These relatively new priorities are especially pressing for archives which are no longer accumulating materials such as The Valley of the Shadow, The Rossetti Archive, and the Salisbury Archive. When these archives hit “retirement” what characteristics in them need to be preserved besides the archival content? Sustaining Digital Scholarship has engaged this question, but the problem is ongoing and invariably unique to every project. The technical histories show us that there is already a record stored in these archives; this is the record of their organization and implementation through time. Preservation and curation of this technical history is key to their total history.
References


Palgrave.


