The Poetess Meets IT: An Interdisciplinary Collaboration

David Woods
IT Services
Miami University
Oxford, OH USA
woodsdm2@muohio.edu

Laura Mandell
English Department
Miami University
Oxford, OH USA
mandellc@muohio.edu

ABSTRACT

Digital Humanities is an emerging area of interest for humanities scholars that involves the development and use of new technologies for conducting research and presenting scholarship. This emerging field presents both opportunities and challenges to the computing and technology fields. Among the challenges of Digital Humanities efforts is the need to address access issues including access to tools for content creation as well as the need to create user friendly tools for humanities scholars. This paper offers a case study of a collaboration between members of the English faculty and IT support group at Miami University for development of the Poetess Archive with a discussion of how access issues were addressed in this project. The suitability of this type of project for a project-based computer science course is also discussed.

1. INTRODUCTION

The field of Digital Humanities, also known as Humanities Computing, covers a wide range of activities [13, 7, 6, 9]. This makes it hard to define the field, but one useful definition is: "The field of Digital Humanities involves thinking about the impact of new technologies on thought, about ways of shaping and visualizing information, and about the relationships between code and language, programming and ideology."[15]

The field of Digital Humanities covers a large scale in terms of the complexity of the projects. Some of the more complex efforts include the Collex tool developed by the Nines project [11] and the Perseus Project [2]. The Poetess Archive [12], which is the subject of this case study, is a more modest undertaking.

For all Digital Humanities projects, access is a significant issue with two main aspects. The first is the need for access to tools and technologies that are easy for humanities scholars to learn and use in implementing digital humanities projects, but don't limit the results of the project. An alternate solution would be for the humanities scholar to work with a technology professional, but the emphasis must be on the humanities scholarship, not the technology. The second aspect is the user interface - all tools that are developed must be easy to use without limiting or biasing the questions that can be explored using the tool. This view of access also needs to make allowances for the different research and browsing techniques used by humanities scholars [8].

Digital humanities projects may also be of interest in computer science education since they offer an opportunity to combine computer science and humanities disciplines and engage computer science with core liberal education concepts [1]. These projects present an opportunity for project-based courses that allow students to use core computer science skills while also developing communication, team work and project management skills that are in demand by prospective employers [5]. Digital Humanities projects also offer the opportunity to view a project from the user's perspective and understand the importance of considering inclusiveness and accessibility when implementing technology [4].

This paper will discuss the migration of an existing digital humanities tool, the Poetess Archive, from a static web site to a dynamic, database driven site. The topics discussed include addressing access issues, the design discussions involved in redesigning the site, the technology used, the main challenges faced during this project and the potential for future growth and expansion of the Poetess Archive.

In the field of computer technology, the term "access" is usually used in terms of problems with access to technology resources or software. This paper is a case study of an effort to address a different "access" issue - the issue of access to the knowledge needed to use available hardware and software to create useful digital content.

2. THE POETESS ARCHIVE

The Poetess Archive provides an extensive bibliography and some full texts of works by and about writers working in and against the "poetess tradition," the extraordinarily popular, but much criticized, flowery poetry written in Britain and America between 1750 and 1900. In addition to bibliographic data, page image scans are available for some of the works in the archive. The Poetess Archive serves as a tool to allow scholars to explore works from and about this period and identify works that are of interest to their scholarly endeavors.

Many of the primary works in the Poetess Archive were published in literary annuals, gift books, and other types of collections that are now typically found only in private collections and rare book archives. The goals of the Poetess Archive include fostering scholarship and discussion by making these works more accessible and allowing users to perform a wide variety of searches.

The core data for the Poetess Archive consists of TEI [14] encoded XML files for each of the works in the archive. For most of the works, the data in the XML file provides only a bibliographic entry for the work, but for some works, the XML file contains the full text of the work. Scanned page images of the original works are available for some texts. XSL translations were used to produce HTML encoded files with a common appearance for all of the works.

2.1 Original Poetess Archive

The original presentation of these works was done using static HTML pages where the works were listed alphabetically in general categories, such as "Works by British Authors" and "Works by American Authors." Additional ideas for categorizing

and ordering the data had been identified but not implemented. The English faculty member, with the assistance of a few graduate students, encoded all works and maintained the Poetess web site.

As additional works were added to the archive and additional ideas for categorizing and ordering the data emerged, several limitations with the static implementation became apparent. Adding works required manual editing of several HTML files and introduced opportunities for typographical errors. Adding new options for ordering the works required creating a new HTML page and manually adding all of the relevant works. In addition, the manual editing of the HTML files led to subtle differences in the "look and feel" of the different pages in the Poetess Archive.

2.2 Vision for new Poetess Archive

The redesign of the Poetess Archive had several goals. The main goal was to design a site that would make it easy for scholars and students to conduct research. Another goal was to make it simple to expand the archive – both to add additional content and additional ways of organizing and searching the content. A final goal was to ensure that the archive would require little or no on-going maintenance.

The solution of re-implementing the Poetess Archive as a dynamic, database-driven web site was readily identified. However, there was one significant hurdle that had to be overcome to implement this solution. Re-implementing the Poetess Archive web site as a dynamic, database-driven site would require more advanced technical skills. Although the English faculty member was quite technology savvy, spending significant time learning new technologies would reduce the time available to pursue literature scholarship since the skills required for developing and maintaining the dynamic web site would be significantly more advanced than the skills required for implementing the static version of the Poetess site.

3. SOLUTION

3.1 Poetess Archive Re-design

The initial re-design efforts focused on the user interface and making the Poetess Archive data accessible to potential users. This work was aided by a summer workshop sponsored by NINES [9]. This workshop provided an opportunity for the English faculty member to interact with a group of scholars working on digital humanities projects in areas closely related to the Poetess Archive. This interaction helped to identify and refine the main elements of the user interface for the revised Poetess Archive.

Implementing the revised design for the poetess Archive provided an opportunity to utilize the Research Computing Support Group that had been recently established by the Information Technology Services department at Miami University to support this type of project. A collaboration between the English faculty member department and a support specialist from the Research Computing Support Group was established to redesign and re-implement the Poetess Archive. This would allow the English faculty member to act as the client and focus on the appearance and content of the archive while the support specialist focused on the technology used to implement the archive.

The effort to implement these ideas for the new user interface and the backend database began with an initial design meeting. During this meeting, the collaborators discussed the limitations of the current Poetess Archive and reviewed ideas for the appearance and features of the new archive. Based on these discussions, the collaborators developed an initial design for the Poetess database and further refined the ideas for the appearance and features of the site.

The new interface design for the Poetess Archive site had three main groupings for the works. The first group contained all works grouped into several specific sub-categories for each author. A second grouping contained all works published in literary annuals and collections, including anthologies, miscellanies, and beauties with separate entries for the actual collections and annuals. The final group contained criticism and biographical material. During development, an advanced search option was added to facilitate more general searches.

Additional options for ordering and restricting the results for each of the major groups were identified. Ordering options included sorting by author, title, country of publication, and year of publication. For all groupings, options to restrict results based on source type (primary or secondary) and data type (full text or bibliographic only) were identified. Content restriction options specific to each main group were also noted. For example, in the criticism grouping, content could be restricted by century of publication and bibliographic content could be explicitly excluded or included.

There were some additional requirements that were common to all of the result groupings. These included hyperlinks to the XML and HTML files for each work with an indication about whether the files contained the full text of the work or only held bibliographic information. The ability to display additional hyperlinks specific to a given work was also requested.

A professional web designer assisted with the design and implementation of the visual aspects web site. Working from the requirements identified for the revised Poetess Archive, an initial design for the web site interface using a tabbed interface was developed (see Figure 1).

Figure 1 The Poetess Archive web site

The database for the Poetess Archive was designed to allow as much flexibility as possible. Two main elements of the database were identified. One element consisted of information about the creators of the works in the Poetess Archive. The term creators included not only authors, but editors, painters, and engravers with the possibility of additional types as the archive grew.

The other main data element was the actual works, which included prose, poetry, and illustrations. In addition, since the Poetess Archive included many works that were published in collections, the works data included tables of contents, tables of illustrations, inscription pages, title pages, dedications, etc. for these works. A cross reference table was used to link creators with works with additional information about how creators were involved with a particular work, such as editor, author, or engraver.

The English faculty member reviewed the initial database design to ensure that it would be able to hold all of the required information about the creators and works. She also identified constraints that would be applied to specific database fields. Several iterations of the refinement and review process were completed before development was started. Later iterations of the design process included a prototype database with a limited set of data that was used to validate the results of the planned queries. During the review process, a number of data validation requirements were identified for implementation in the data loading process.

3.2 Re-implementation

Once the initial design for the web interface and database were completed, decisions about the implementation environment were made. The server that was identified to host the new version of the Poetess Archive was a general use Sun Solaris server with an Apache web server (www.apache.org), Oracle database server (www.oracle.com), and PHP scripting language (www.php.net). To allow for the possibility that the Poetess Archive would be moved to a different server and/or database in the future, the PEAR DB package for PHP [3] was used as a database abstraction layer and use of Oracle specific database features were avoided.

The work to implement the revisions to the Poetess Archive was completed over the course of a few months. As the site was developed, some additional improvements such as the "Advanced Search" option were also identified and implemented. While the web site revisions were being implemented, a large number of additional works were identified and coded for inclusion in the Poetess Archive. When this additional data was load into the Archive, additional usability and performance issues were identified and resolved.

4. RESULTS

4.1 Reflections on the Project

4.1.1 Humanities Viewpoint

Collaborating with a database designer requires an amazing amount of patience and persistence on the part of the database designer. Because English professors do not think in terms of data organization, it was only when I tried to ask myself certain research questions that I realized I needed to add certain fields to the database or change the kinds of answers that constituted legitimate data.

The collaboration worked together rather differently than most of those who undertake a database. Instead of working directly from scanned or typed documents, we began with encoding the documents using the XML application created by the Text Encoding Initiative [14]: a group of Humanities scholars who are trying to standardize the way library-quality digital texts are created so that we can help each other with continuously upgrading them as new applications, standards, and modes of presentation become available. I would advocate this workflow system to any Humanities scholar creating a database because it allowed me to create XSL programs that could pull information out of the texts in different ways every time we decided we needed new fields.

The challenge for the technology specialist was that, whenever he found mistakes in my data, he had to wait for me to change my programs and rerun them, rather than me manually correcting the mistakes. That takes longer but in the end insures database integrity. It would have been easy to simply correct data input using PHP, but we had to force ourselves to keep the data entry on my side of the work-flow divide so that my data-extracting programs could be complete and effective as well as uniform site-wide. Without the technology specialist continuous checking my data and telling me exactly what mistakes he had found, I never would have been able to properly, consistently encode my documents and write programs for data extraction.

4.1.2 Technical Viewpoint

The main technical challenges were working with string data and maintaining acceptable performance as additional data was added to the Poetess Archive. One example of an issue with the data was ordering data by place of publication. The required categories were "Works published in the United Kingdom," "Works published in the United States," "Transatlantic Works" (works by American creators published in the UK and vice-versa), and "Other." The bibliographic data typically had the city and state, for example "New York, NY" or "London" and could be easily classified as US, UK, or Other by a human, but required additional data for computer classification. The designation of works as "Transatlantic" was left to the English faculty member.

Sort ordering of works by title shows one example of the issues faced in building domain specific knowledge from the field of literature into the Poetess Archive. The desired sort order required that leading articles such as "The", "An", and "A" should be ignored. Once this domain specific requirement was identified, it was easily implemented.

Several performance and usability issues were identified as the volume of data in the Poetess Archive increased. Extensive analysis and monitoring of the database was done to identify any potential bottlenecks. Significant changes including caching of the largest pages were implemented to improve performance and usability.

4.2 Lessons Learned

The biggest challenges that were faced in this project involved communication. Both of the disciplines involved in this project, Literature and Computer Science, have their own concepts and vocabulary that are not familiar to persons outside of the field. To minimize the chances of confusion, face-to-face meetings were used for project discussions when possible, and specific examples were frequently used when discussing project details. In all aspects of the project, technical jargon was avoided as much as possible, and when technical terms had to be used, the project members made sure they were clearly defined and understood.

Another lesson learned in the project is that is was possible to develop a web site that was flexible enough to meet the needs of

the literature scholar. The main challenge was to resist the urge to tightly constrain the database and query designs. Doing this would have simplified the technical aspects of the site, but with the undesired result of restricting the scholarly exploration potential of the site.

4.3 Suitability for Course Project

Based on our experience with this project, we feel that this type of project could be completed as a project assignment in a computer science course or a course cross-listed in computer science and a humanities field. The technical skills required include web site design, database design, and web programming which would be a good fit for an upper level project-focused course or a capstone project.

In addition to giving students an opportunity to apply skills learned in CS courses, this type of project would also provide an opportunity to build troubleshooting, communication, project management, and team participation skills. Working on a project involving a humanities discipline would present challenges with learning discipline specific vocabulary and understanding accessibility and usability requirements from the user's perspective.

Work on digital humanities projects would also give students experience with another common real-world situation by working with a client to develop and refine functional requirements and turning these into technical requirements that would meet the client's needs. As part of this process, the student project team could use rapid prototyping techniques to allow the client to continuously validate that functional requirements were being met.

The need for domain specific knowledge of a humanities field also makes Digital Humanities projects an excellent way of integrating technically focused CS curriculum with aspects of the core liberal art curriculum. Interdisciplinary Digital Humanities projects also provide an opportunity to expose humanities students to aspects of modern computing technology. This could give humanities students a better understanding of the technology they encounter in their everyday lives and could also improve their employment prospects or open them to new fields of study.

4.4 Results

The revised Poetess Archive web site has seen steady traffic since it was made public. The database currently contains data on over 4100 works and over 900 creators. The site has been publicized by the English faculty member in several seminars and through word of mouth. The site has also been submitted to the Collex tool developed by the NINES (Networked Interface for Nineteenth-Century Electronic Scholarship) project [2], which has generated site traffic to specific works in the Poetess Archive.

Statistics on site traffic have not been collected, but a review of the web server logs show steady traffic to the site. As expected, visitors have been from educational institutions in the English-speaking worlds have been logged. In addition, a number of visitors from sites outside the English-speaking world including Japan, France, and Poland have been observed.

5. FUTURE PLANS

In addition to expanding the number of works contained in the Poetess Archive, we also plan for several other enhancements. As more scholars utilize the Poetess Archive and identify needs, we will store additional metadata about the works to allow more search options. We are also investigating designing a visualization tool linked to the database that would allow scholars to visualize and interact with the data in dynamic 3-D models. An online journal, the *Poetess Archive Journal*, has been launched and tools for online discussion and review of the articles in this journal are being investigated.

6. ACKNOWLEDGMENTS

We would like to acknowledge the support of Miami University Information Technology Services.

7. REFERENCES

[1] Association of American Colleges and Universities, Statement on Liberal Learning. http://www.aacu.org/About/statements/liberal_learning.cfm Accessed 12/19/2006.

[2] Collex (COLLection and Exhibition) mechanism for the semantic web, NINES (Network Interface for Nineteenth-century Electronic Scholarship) http://www.nines.org/collex

[3] DB Package from PEAR (PHP Extension and Application Repository) http://pear.php.net/package/DB Accessed 12/20/2006.

[4] Howard, Elizabeth V. Promoting Communication and Inclusiveness in the IT Classroom. In *Proceedings of the 6th Conference on Information Technology Education (SIGITE '05)*, pages 311 – 317. Newark, NJ, USA, 2005.

[5] IT Industry and Education Forum. (2005). Miami University Hamilton

[6] Jessop, Martyn, Computing or humanities? *Ubiquit,y 5*, 41 (December 23, 2004), 1-1.

[7] Katz, Stanley N., Why Technology Matters: the humanities in the twenty-first century. *Interdisciplinary Science Reviews, 30*, 2 (June 2005), 105-118.

[8] Massy-Burzio, Virginia. The Rush to Technology: A View from the Humanists. *Library Trends, 47*, 4 (Spring 1999), 620-639.

[9] McCarty, Willard. *Humanities Computing*. Palgrave MacMillian, New York, NY, 2005.

[10] NINES Summer Workshops - http://www.nines.org/join/workshop.html Accessed 12/14/2006.

[11] The Perseus Digital Library – http://www.perseus.tufts.edu

[12] The Poetess Archive – http://unixgen.muohio.edu/~poetess/

[13] Schreibman, S., Siemens, R., and Unsworth, J. (Ed.). *A Companion to Digital Humanities.* Blackwell Pub., Malden, MA. 2004.

[14] The Text Encoding Initiative – http://www.tei-c.org

[15] http://www.units.muohio.edu/technologyandhumanities/field.htm Accessed 11/20/2006.