Integrating Collections at the Cervantes Project

Neal Audenaert, Richard Furuta, Eduardo Urbina, Jie Deng,
Carlos Monroy, Rosy Sáenz, Doris Careaga
TEES Center for the Study of Digital Libraries
Texas A&M University
College Station, TX 77843-3112, USA

cervantes@csdl.tamu.edu

ABSTRACT

Unlike many efforts that focus on supporting scholarly research by developing large-scale, general resources for a wide range of audiences, we at the Cervantes Project have chosen to focus more narrowly on developing resources in support of ongoing research about the life and works of a single author, Miguel de Cervantes Saavedra (1547-1616). This has lead to a group of hypertextual archives, tightly integrated around the narrative and thematic structure of *Don Quixote*. This project is typical of many humanities research efforts and we discuss how our experiences inform the broader challenge of developing resources to support humanities research.

Categories and Subject Descriptors

J.5 [**Arts and Humanities**]: Literature.

Keywords

Cervantes Project, humanities research, hypertext.

1. INTRODUCTION

There is growing interest in developing digital collections and tools to support scholarly research and to help "new audiences … ask new questions about new ideas" [1]. Much of the current work has focused on developing large-scale, general resources. At the Cervantes Project, we have taken a different approach, focusing more narrowly on developing resources for ongoing research about the life and works of a single author, Miguel de Cervantes Saavedra and using the narrative and thematic structure of the texts to support the automatic integration of new resources. This approach has enabled us to explore the breadth of scholarly research activities centered around a single author; largely on a single work, *Don Quixote* (*DQ*). We have found the humanities research involved to be characterized by numerous researchers conducting detailed studies of highly focused, inter-disciplinary research questions. For example, some researchers are interested in textual analysis, others on historical and biographical records, others on Cervantes' impact on music, etc. Consequently, a key question is how to provide tight interlinkages among the resources developed by these separate researchers without requiring large

amounts of follow-on customization; an unaffordably labor-intensive effort.

While we have focused on the life and works of Cervantes, the practices we have observed in this project typify many humanities research endeavors. It is common for a number of scholars to thoroughly investigate the cultural impact of a single individual. This paper presents a survey of current work in the Cervantes Project with the goal of outlining the scope of tools and practices needed to adequately support and integrate the in-depth, focused research characteristic of humanities disciplines.

2. THE CERVANTES PROJECT

The Cervantes Project is developing a suite of tools that can be grouped into five major research areas: bibliographic information, textual analysis, historical research, music, and iconography.

Bibliographies are perhaps the primary artifact for developing a scholarly "corporate" memory in numerous areas. Since 1996, we have maintained a comprehensive bibliography of scholarly publications pertaining to the Cervantes' life and work. This is published periodically in both an online and print form. We have implemented a flexible, database driven tool to manage large-scale, annotated bibliographies. This tool supports the taxonomies and multiple editors required to maintain a bibliography with thousands of records. It will soon replace the current static HTML edition of the bibliography.

Early in the project we also worked to modernize the primary resources used in traditional textual analysis—the edition and associated commentary. We developed an electronic *variorum* edition (EVE), a reader's interface (VERI), and a variant editor (MVED) [3], and are in the process of populating a collection of a scope previously unavailable to the Cervantes scholar. Currently, digital images of ten copies of the 1605 *princeps* edition, one copy of the 1615 *princeps* edition and one copy of the three volume Bowle edition (1781) are available online. Nine copies of the 1605 *princeps* edition have corresponding full text transcriptions linked to the page images and work is currently in progress to add eight copies of the 1615 *princeps* edition (with transcriptions) and about twenty-five copies of later editions. Although the presentation of the texts is centered on the printed version, we are taking advantage of the electronic medium to reshape the form of the books; for example bringing into proximity related portions of the multi-volume Bowle edition previously separated into separate books. We have also employed timelines to visualize the variants and verify the transcriptions.

To better support access to historical and biographical data, we are currently developing tools to identify dates, people, and places in a collection of approximately 1600 official documents

pertaining to Cervantes and his family [2]. Once identified, this information is used to automatically generate dense navigational links and support collection visualization. Biographies and chronologies of Cervantes and his family will be integrated with this collection, connecting events, people, facts and places to primary source materials.

We have recently begun developing a digital collection that explores the intersection of music and Cervantes. It will include detailed information about the instruments Cervantes mentions (images, audio, descriptions, etc.). It will also organize songs, dances, and other musical works inspired by *DQ* around the narrative and thematic elements of the text. This will be used to assist scholars investigating Cervantes' awareness of the musical trends of his day, the influence of that music on his writings and the subsequent interpretation of the works of Cervantes by various musicians. Specifically, it will provide them with access to playable scores from musical works written about Cervantes, discussions of themes found in the music of Cervantes' day and how those themes are reflected in his writings, historical notes, bibliographic information, and audio.

Finally, we are assembling an extensive collection of illustrations from various editions of *DQ* and *ex libris* (bookplates) inspired by or based on *DQ*. We have collected more than 1300 *ex libris*. For our textual iconography collection, we have acquired nearly 400 copies of illustrated editions of *DQ* published between 1620 and 2004. Currently, we have digitized more than 4000 images from 74 of the most significant of these editions. These illustrations are encoded with detailed metadata information pertaining to both their artistic features (e.g. artist, date, size, style, texture) and their literary context (e.g. thematic and narrative elements, characters). We are developing a number of different collation strategies to facilitate access to these illustrations. These include book-based collations that allow the illustrations to be placed in their original physical, narrative or thematic context, natural collations that group illustrations by author, style, size, etc., and custom collations created or tailored by individuals. The iconography collection facilitates investigations of artists' interpretation of *DQ* throughout history, the cultural, political, and ethical factors that have influenced these interpretations, and individual artists unique analysis, techniques, and stylistic flavor.

These five lines of research come from distinct scholarly traditions and offer diverse perspectives. They are united, however, in their common goal to better understand Cervantes' writings and the impact of those writing on the human experience. Historically, the unity of this research has often been lost to the pragmatic difficulties of identifying and accessing the relevant research across the boundaries of academic disciplines. The digital resources we are developing help bridge those boundaries, bringing research results together in a single digital library structured by the narrative and thematic elements of *DQ*.

3. DISCUSSION
Each of these collections represents independent work by individual Cervantes scholars. The electronic medium affords the possibility to interlink these individual efforts, creating a resource that enables investigations to be enhanced by new understandings of previously isolated areas. A key question is how to support this effort without requiring hand coding—a task that itself would be a significant undertaking.

The approach we have taken is to focus on identifying and tagging the narrative and thematic elements in the texts themselves, rather than relying on the more traditional positional ties to printed volumes (e.g. page and line numbers). We have developed taxonomies and controlled vocabularies to support this task and have begun the process of encoding *DQ* and other works using TEI standards. As this is done, other resources will be assigned metadata that identifies their relationships with the identified structure of the text. Additionally, we are working to develop more sophisticated logical representations of the materials in our collections so that additional tools can be layered on top of existing data structures to quickly meet new research requirements. Once this is done, it will be possible to identify how a new research result interlinks into the collection via its metadata and automatically integrate it into the digital library.

Second, we want to begin to explore how the lessons learned in the Cervantes Project might be applied more generally to work being done with other authors. We expect that some elements (e.g. bibliographies) will be applicable to a wide range of projects while others (e.g. the music collection) will be less so. To what extent can our tools and techniques be generalized to support other projects? How might multiple such projects be combined to enhance each other?

4. CONCLUSIONS
While the work outlined here by no means exhausts the scope of scholarly research that may be motivated by a single author's work, it does begin to suggest the breadth and interdisciplinary nature of that research. This type of work is characterized by detailed, thorough investigations with a relatively narrow focus, the engagement of a broad range of humanities disciplines (e.g. art, music, publishing, literature, sociology, etc.), large bodies of secondary work developed over (potentially) hundreds of years (requiring bibliographies and other tertiary scholarly works) and the need to integrate primary and secondary material. Many large collections (e.g. Perseus Project) are built from resources that lack a strong, natural structuring motif. Instead, structure is imposed as a result of choices made in developing visualizations and linking services (e.g. spatial-temporal visualizations). In the Cervantes Project, both our documents and our research needs are intrinsically structured by the narrative and thematic elements of the texts. This structure brings unity to a variety of research directions and integrates otherwise disconnected work.

5. REFERENCES
[1] Crane, G., Wulfman, C. Towards a cultural heritage digital library. In *Proceedings of Joint Conference on Digital Libraries (JCDL03) (*Houston, TX, 2003). IEEE Computer Society, Washington, DC, USA, 2003, 75-86.

[2] Krzysztof Sliwa. *Documentos Cervantinos: Nueva recopilación; lista e índices*. New York: Peter Lang, 2000.

[3] Furuta, R., Urbina, E., et al. The Cervantes Project: Steps to a Customizable and Interlinked On-Line Electronic *Variorum* Edition Supporting Scholarship. In *European Conference on Digital Libraries, ECDL2001*. (Darmstadt, Germany, September 2001). Berlin: Springer, 2001. 71-82.