Integrating Diverse Research in a Digital Library Focused on a Single Author

Neal Audenaert[1], Richard Furuta[1], Eduardo Urbina[2], Jie Deng[1], Carlos Monroy[1], Rosy Sáenz[2], and Doris Careaga[2]

[1] TEES Center for the Study of Digital Libraries,
& Department of Computer Science,
Texas A&M University,
College Station, TX, 77843
cervantes@csdl.tamu.edu
[2] TEES Center for the Study of Digital Libraries,
& Hispanic Studies Department,
Texas A&M University,
College Station, TX, 77843
cervantes@csdl.tamu.edu

Abstract. The works of a significant author are accompanied by a variety of artifacts ranging from the scholarly to the popular. In order to better support the needs of the scholarly community, digital libraries focused on the life and works of a particular author must be designed to assemble, integrate, and present the full scope of these artifacts. Drawing from our experiences with the Cervantes Project, we describe five intersecting domains that are common to similarly focused humanities research projects. Integrating the tools needed and the artifacts produced by each of these domains enables digital libraries to provide unique connections between diverse research communities.

1 Introduction

Like many other projects dedicated to a single author, work at the Cervantes Project initially focused on maintaining a comprehensive bibliography of scholarly research and providing access to the works of Miguel de Cervantes Saavedra (1547-1616). We made his works available via the Cervantes Project web site in a variety of editions in several versions (facsimile, old-spelling, modernized, English) along with the interfaces, hypertext links, and search engines to facilitate their use at multiple levels [13]. We have developed an electronic *variorum* edition (EVE) and are populating it with a text collection previously unavailable to the Cervantes scholar [5]. While work is ongoing in these areas, we are expanding the scope of our project to include resources to support historical and biographical research, investigations into the impact of Cervantes' cultural environment on his writings, and studies of popular and scholarly artifacts based on or inspired by Cervantes.

As we expand our focus, we are able to better categorize the breadth of scholarly research activities centered on a single author and the types of resources that digital

A. Rauber et al. (Eds.): ECDL 2005, LNCS 3652, pp. 151 – 161, 2005.

libraries need to provide to support those activities. The humanities research involved in this project is characterized by numerous researchers conducting detailed studies of highly focused, inter-disciplinary research questions. For example, some researchers are interested in illustrated editions of *Don Quixote*, others on historical and biographical records and yet others on Cervantes' impact on music. This in turn raises the question of how to provide tight interlinkages among the resources developed by various 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. The works of a significant author or, more generally, a single artist (author, painter, poet, etc.) are accompanied by a variety of artifacts ranging from the scholarly to the popular that are distributed in time from before the author was born to the present. Consequently, to fully support this research, digital libraries focused on the life and works of a particular author cannot be content merely to present the author's works. Instead, they need to be designed to assemble, integrate, and present the full scope of these artifacts in a manner that facilitates the sophisticated interpretative strategies required for scholarly research [7].

In this paper we discuss our growing understanding of the scope of humanities research practices, drawing on our current work to inform and illustrate our findings. We also describe our current efforts to employ the narrative and thematic structure of Cervantes' most notable work, *Don Quixote* (*DQ*), to provide an integrative motif that lends a natural unifying structure to the diverse artifacts in our archive.

2 The Cervantes Project

The Cervantes Project is developing a suite of tools that can be grouped into six major sub-projects: bibliographic information, textual analysis, historical research, music, *ex libris* (bookplates), and textual iconography. Figure 1 provides an overview of how the artifacts from these sub-projects fit into a timeline of Cervantes' life and writings. While each of these sub-projects is individually interesting, the primary contribution of our work stems from the fact that by integrating them, we are able to bring diverse and previously disconnected research together in a way that enhances the value of each sub-project.

2.1 Bibliographies

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 Cervantes' life and work [14]. 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.

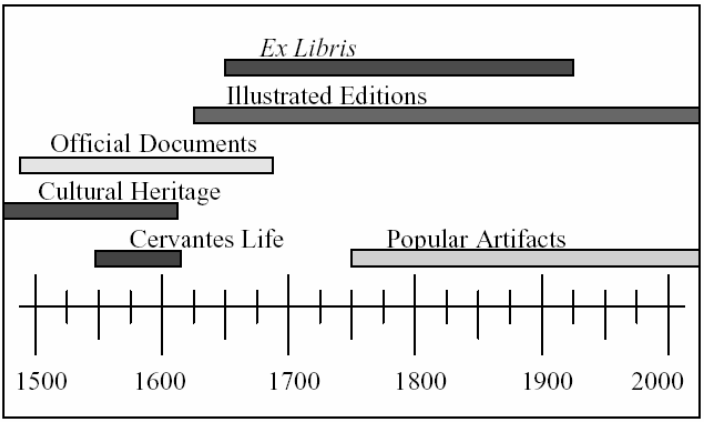

Fig. 1. A lifeline depicting the collections included in the Cervantes Project as they relate to his life

2.2 Textual Analysis

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) [5], 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 published as separate books. We have also employed timelines to visualize the variants and verify the transcriptions [10].

2.3 Historical Research

To better support access to historical and biographical data, we are developing tools to identify dates, people, and places in a collection of approximately 1600 official documents pertaining to Cervantes and his family [11]. Once identified, this information is used to automatically generate dense navigational links and to 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.

2.4 Music

We have recently begun work with 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 collection 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 Cervantes' works 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.

2.5 Ex Libris

We have assembled a digital collection of more than 1300 *ex libris* (bookplates) inspired by or based on *DQ*. These *ex libris* are taken from collection loaned to the Cervantes Project by Doctor Gian Carlo Torre. His collection is one of the most important in the world. Now, in its digital form, we are able to offer, for the first time, easy access and reference to a modern artistic corpus that expands the iconography and visual reading of *DQ* while providing an insight into the iconic transformation of the text in the 20th century.

2.6 Textual Iconography

Finally, we have assembled an extensive collection of illustrations from various editions of *DQ*. The Cervantes Project, in collaboration with the Cushing Memorial Library of Texas A&M University, has 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). The iconography collection facilitates investigations of artists' interpretation of *DQ* throughout history, the cultural, political, and ethical factors that have influenced these interpretations, and the individual artists' unique analysis, techniques, and stylistic flavor.

We plan to enhance this collection will collation tools to facilitate access to these illustrations. These include support for 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

2.7 Narrative and Thematic Structure as an Integrative Motif

These six 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 these writings 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 these boundaries, bringing research results together in a single digital library structured by the narrative and thematic elements of *DQ*.

To bridge these boundaries, it is not enough to simply publish archives of artifacts "on-line" [2]. To adequately support the humanities researcher, the connections between these sub-projects need to be identified and the collections enhanced with tight interlinkages. One key challenge we face as we develop these resources is supporting the integration of these resources without requiring hand coding–a task that would itself be a significant undertaking.

Our approach focuses 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 (for example, page and line numbers). In support of this approach, we have developed taxonomies and controlled vocabularies and are encoding *DQ* and Cervantes' other works using TEI standards [6]. The text is naturally divided into chapters that we further sub-divide to indicate narrative units within those chapters. Each of these major and minor narrative divisions is given a short description and tagged with taxonomic categories, principal themes and dominant moods. Examples of taxonomic categories include chapter, place, direction, episode, adventure, action, and character. Examples of themes include madness, love, food, play, enchantment, knighthood, chivalry, justice, freedom, and violence. Examples of moods include parodic, burlesque, ironic, and satiric.

As this is done, other resources can be assigned metadata that identifies their relationships with the structure of the text. This approach allows us to automatically integrate new artifacts as they are added to the collections, providing links from the new artifact to previously existing resources, and from existing data to the new artifact.

3 Characterizing Scholarly Research

Our experiences with this project have encouraged us to carefully reexamine the practices that characterize scholarly research centered on a single author and the types of resources that digital libraries need to provide to adequately support those practices. The work we are currently conducting is clearly not limited to the study of the works produced by Cervantes and commentary on those works. It extends to his life in general, historical documents of his time, the contemporary cultural context in which Cervantes and his works are embedded, and the scholarly and popular artifacts that are based on or inspired by his works. Figure 2 shows a Venn diagram illustrating the relationships between these research domains. We believe that this view is generally descriptive of humanities research efforts centered on the life of a single author or, more generally, a single artist (author, painter, poet, etc.). In this section we describe each of these areas, drawing on our experiences with the Cervantes Project to illustrate the key issues involved.

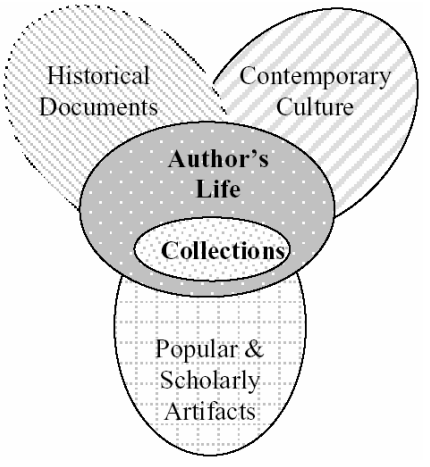

Fig. 2. A Venn diagram illustrating the relationships between the five domains of humanities research focused on a single author: collections of the author's works, the author's life, historical documents, contemporary culture and derivative popular and scholarly artifacts

3.1 Collections

Assembling and presenting a collection of an author's work is central to any project purporting to be a digital library that supports research about a single author. McGann has clearly demonstrated the potential advantages digital resources have for studying texts over traditional print based approaches and has presented a vision for digitally based humanities research [8]. Numerous projects have taken up his vision, including the Rossetti Archive [16], the Canterbury Tales [12], and the Picasso Project [15] among others. While this is the smallest and most well defined domain in a humanities research project, it is by no means simple. Tools are needed to support textual criticism of older texts. Difficult decisions must be made: How should the texts be presented visually? Should the original structure of the book be retained? Which texts should be given primacy when revisions have been made by the author (e.g., should the last version of the text published during the author's lifetime be given primacy, or are all editions of equal value)? Are individual copies important in their own right or only as they serve as an exemplar of an edition? Is it sufficient to provide a textual representation of the original text or are facsimiles of the original pages needed [4]?

Within the Cervantes Project we have found that these decisions need to be carefully evaluated not just on a project-by-project basis, but also within each project on a more fine-grained basis focused on the nature of specific research questions. For the majority of Cervantes' works (e.g. *Novelas ejemplares*, *La Galatea*, etc.) we have been content to provide textual transcriptions of the work in a variety of editions (old-spelling, modernized, English translations). This reflects their relatively decreased prominence in the corpus of Cervantes' work. On the other hand, we have paid considerably more attention to these questions with respect to developing resources for *DQ*. Our facsimile and critical editions form the most significant part of our textual archive of Cervantes' work. For the most part, we have retained the original structure

of the book in presenting these editions, but in the Bowle edition we recast that structure in order to align the commentary section in the third volume with the corresponding pages in the first two volumes. Breaking the structure of the book, for example to bring two different editions into proximity for comparison purposes or to integrate images and commentary that were not originally part of the text, is likely to become a more prominent feature of our approach. Depending on features of a particular edition, we have made different decisions about the primacy of copies. For the *princeps* edition, each extant copy is different and hence important. For the Bowle edition, one copy is sufficient to serve as an exemplar of the edition as an abstract entity. In deciding between textual vs. facsimile representation of the text, our answer has again depended on specific research questions. For the task of textual criticism, both full text transcripts and images of the original pages are required to adequately support the analytical tasks involved. For many of the later editions in which variants between the text of the edition and the "original" are less important, we have been content to provide only a facsimile edition. This reflects scholarly interest in the form of the publication as it changed from edition to edition rather than in the less significant changes in the content of the text.

3.2 The Author's Life

The works of an author are encompassed within the broader context of his life and nearly all projects focused on a single author make an effort to present at least a minimal amount of information about the author's life. Access to detailed information about an author's life is important to scholarly work for two broad purposes. First, a biographical understanding of the author may lead to insights into the motivations for and perspectives influencing his writings. Second, learning more about the author's life is interesting in its own right. It is the subject both of scholarly research and popular interest (as indicated by the existence of television stations such as the Biography channel). Integrating biographical information into a scholarly archive enhances reader' understanding promoting a dialog between the focused studies of the author's works and the biographical information.

Traditional scholarly tools for studying an individual's life are well recognized and include biographies and chronologies. These are often supported by paintings, maps. and other materials. These are secondary resources, however, and are limited in their ability to support the discovery of new information. Similar to the way in which McGann critiqued the copy-text approach to textual criticism for selecting an authoritative central text and describing its differences with other, marginalized texts [9], a biography can be seen as establishing a centered, "authoritative" narrative of the author's life, relegating other possibly significant elements to the margins. While this may be useful for many purposes, one of the key advantages of developing a digital archive is to assist scholars (and the public) in finding their own way through the available information developing and utilizing their own unique interpretive perspectives [3]. Set in the context of a digital library, research into the life of the author can benefit not only from the explicitly biographical information commonly provided, but also from the other information in the archives supporting historical research and information about the contemporary cultural context of the author. In this environment, biographies can serve as annotated trails and their authors as trailblazers in

ways similar to those envisioned by Bush [1]. This attitude reflects a different perspective on this domain of scholarly research and affords new possibilities for supporting research without constraining scholars to the single interpretation presented by the biographer.

The Cervantes Project has long provided access to the traditional tools of biographical research (limited by ever-present copyright restrictions). We are beginning to augment this approach with the extensive collection of primary source materials drawn from the official documents collected by Sliwa [11] as well as the results of research into Cervantes' awareness of contemporary musical themes. While these two collections are distinct sub-projects, we are working to integrate them into the overall structure of our archive in order to enable researchers and the public to gain a better understanding of Cervantes' life.

3.3 Historical Documents

Both the author's life and works are embedded in, though not encompassed by, a broader historical context. Providing resources to inform and clarify scholars' understanding of this historical context enhances their understanding of, and engagement with, the author. From a different perspective, exploring the interactions between a particular artist and the historical context in which he and his work are embedded can help scholars better understand significant facets about that historical context. To support research efforts in this domain, digital libraries need to provide access to relevant historical documents and research results. This opens the challenging question of what historical resources ought to be included in a library. While many resources may be potentially helpful, including too much information could detract from the library by unnecessarily diverting it from its ideological focus. One alternative would be to develop supporting collections. For example, one could imagine a digital library focused on collecting historical resources for Spain and the western Mediterranean between the 15th and 17th centuries that could support the Cervantes Project as well as a number of other similar humanities archives.

From a more practical standpoint, this problem is often much simpler than the above considerations. Few projects focused on the life and work of a single author have access to collections of the scope of, for example, the Perseus Project's London collection [3] or the resources to digitize it. Thus, the primary considerations become identifying appropriately sized collections that are be available and closely related to the author's life and works. Once a suitable collection is identified, further work will be needed to effectively determine how this collection should be incorporated into the body of artifacts maintained in the digital library. When compared to artifacts collected with the narrower objective of supporting research about the author's life and work, this domain is much more open-ended. The process of determining what artifacts and collections will be most beneficial is heavily dependent on the idiosyncratic needs of particular research communities and on the current availability of materials.

Within the Cervantes Project, we have provided access to the collection of 1600 official documents pertaining to the life of Cervantes and his family as mentioned earlier. The availability and clear relevance of this collection provide an obvious answer to the question of whether or not it merited incorporation into our digital library. Our efforts at integrating it into the larger context of our archive are currently focused on

developing tools to support named entity identification and the identification of other potentially relevant information (the price of eggs in Madrid). As this is done, we will be able to integrate these primary source records better with the biographical information already present in our archive. Questions about how to integrate this information with the texts of Cervantes' remain fruitful areas for future investigation.

3.4 Contemporary Culture

A closely related line of research investigates the nature of the author's engagement with the contemporary cultural context in which he and his works are embedded. To what extent was an author aware of a given cultural theme and how did this awareness affect his writings? What impact did the author have on his contemporaries? How does an author's work inform other questions about his culture? Incorporating cultural artifacts and resources in a single author digital library also assists readers by providing access to information about the author's embedding culture that may be outside the scope of the reader's expertise. Like collections of historical resources, the scope of work that could be incorporated to facilitate an awareness of the author's embedding culture is open ended. Examples of cultural elements that could be integrated include literary, poetic, culinary, societal, dramatic, and musical culture. Again, which resources warrant inclusion is heavily dependent on the characteristics of the author's life and work as well as the pressing issues currently being discussed in the scholarly community.

Our work in developing a music archive related to Cervantes has a significant component that deals with the cultural elements of Cervantes' day. We are incorporating information that connects the texts of Cervantes to information about the musical instruments they mention. Since many of these instruments are not familiar to modern readers or have since significantly altered in form, the connections will improve the depth of our understanding of the texts. From a more scholarly perspective, this collection will describe how major themes and topics found in the music of Cervantes day may be reflected in his writings. This facilitates the exploration of Cervantes' awareness of this aspect of his culture and offers possible insights into his intentions.

3.5 Popular and Scholarly Artifacts

Conceptually, assembling, analyzing, and presenting the body of popular and scholarly artifacts based on or inspired by an author's work is the largest domain for supporting scholarly research. These derivative artifacts lie on a continuum from the results of scholarly research at one end to popular trinkets at the other. At the scholarly end, these artifacts include critical editions, diagrams, and scholarly writings about all aspects of the author's life and works. One clear example of research focused on these artifacts is annotated bibliographies. At the other end of the spectrum are popular artifacts, including souvenirs, trading cards, toys, wrappers, and posters. Between these two extremes lies a tremendously diverse set of artifacts that are unified by their common derivative status.

A few examples from our work within the Cervantes Project will help illustrate some of the possibilities of this class of scholarly research as well as the type of support that a digital library can provide. Our music collection, in addition to providing

resources about music during Cervantes' life, also places a major emphasis on collecting and analyzing the musical compositions that have been based on his works. This resource places the research generated by music scholars into the context of the works on which that music is based. It also provides a unique perspective on the texts of Cervantes by bringing artifacts from this unique interpretive media into proximity to the texts they interpret. Our two collections that focus on artistic elements that been added to the text also fall into this area. The textual iconography project overlaps with research about the text, but since the illustrated editions are all artistic interpretations of *DQ* it is more strongly identified with this area of research. (Contrast this with the illustrated novels of Dickens and Thackeray in which the illustrations were a part of the original published work). The artifacts of the *ex libris* collection are clearly an example of artifacts worthy of scholarly inquiry that are derived from the author's work.

These three examples serve to indicate the breadth and open-ended nature of the scholarly research involved in this domain and of the derivative artifacts that could productively serve in a digital library. This area is the most characteristic of research practice.

4 Conclusions

While our work at the Cervantes Project 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. In carefully reexamining scholarly practices, we have identified five intersecting domains of that are common across similarly focused humanities research projects: the textual analysis, biographical studies, historical context, contemporary cultural context, and derivative popular and scholarly artifacts. Work in each of these areas is characterized by detailed, thorough investigations with a relatively narrow focus, the engagement of a broad range of humanities disciplines (for example, art, music, publishing, literature, sociology, etc.), large bodies of secondary work developed over long time spans (requiring bibliographies and other tertiary scholarly works), and the need to integrate primary and secondary materials. By integrating the tools needed for and the artifacts produced by each of these five major research domains, a digital library focused on scholarly research pertaining to a single individual is able to support unique connections between diverse and otherwise disconnected research communities.

References

1. Bush, V., "As We May Think", Atlantic Monthly (July 1945). 101-108.
2. Crane, G., et al. "Drudgery and Deep Thought," In Communications of the ACM, Vol. 44, Issue 5. ACM Press, New York (May 2001). 35-40.
3. Crane, G., Clifford E. Wulfman, and David A. Smith, "Building a Hypertextual Digital Library in the Humanities: A Case Study on London", In Joint Conference on Digital Libraries, JCDL01, (Roanoke, Virginia, June 2001). ACM Press, New York (2001). 426-434.
4. Flanders, J. "Trusting the Electronic Edition." In Computers and the Humanities, Vol. 31. Kluwer, The Netherlands. (1998). 301-310.

5. Furuta, R., 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: First Springer, 2001. 71-82.
6. Kochumman, R. et al. "Tools for a new Generation of Scholarly Edition Unified by a TEI-based Interchange Format." In Joint Conference on Digital Libraries, JCDL04, (Tuscon, Arizona, June 2004). ACM Press, New York (2004). 368-369.
7. Lynch, C., "Digital Collections, Digital Libraries and the Digitization of Cultural Heritage Information," First Monday, Vol. 7, Issue 5. (May 6, 2002).
8. McGann, J. "The Rossetti Archive and Image-Based Electronic Editing." In Finneran, R. J. (ed.): The Literary Text in the Digital Age. University of Michigan, Ann Arbor, MI (1996) 145-183
9. McGann, J. "The Rationale of Hypertext." In Sutherland, K. (ed.): Electronic Text: Investigations in Method and Theory. Oxford UP, New York, (1997) 19-46.
10. Monroy, C., et al. "Interactive Timeline Viewer (ItLv): A Tool to Visualize Variants Among Documents," In Visual Interfaces to Digital Libraries, Lecture Notes in Computer Science, Vol. 2539. Springer-Verlag, Berlin Heidelberg New York (2002) 33-49.
11. Sliwa, K.. Documentos Cervantinos: Nueva recopilación; lista e índices. New York: Peter Lang, 2000.
12. "The Canterbury Tales Project." De Montfort University, Leicester, England. http://www.cta.dmu.ac.uk/projects/ctp/index.html. Accessed on Feb 7, 2005.
13. "The Cervantes Project." Center for the Study of Digital Libraries, Texas A&M University. http://csdl.tamu.edu/cervantes. Accessed on Feb 7, 2005.
14. "The Cervantes International Bibliography Online (CIBO)." Center for the Study of Digital Libraries, Texas A&M University. http://csdl.tamu.edu/cervantes. Accessed on Feb 7, 2005.
15. "The Picasso Project", Hispanic Studies Department, Texas A&M University.. http://www.tamu.edu/mocl/picasso/. Accessed on Feb 7, 2005.
16. "The Rossetti Archive." The Institute for Advanced Technologies in the Humanities, University of Virginia. http://www.rossettiarchive.org/. Accessed on Feb 7, 2005.