The Cervantes Project: Steps to a Customizable and Interlinked On-line Electronic Variorum Edition Supporting Scholarship

Richard Furuta, Siddarth S. Kalasapur, Rajiv Kochumman, Eduardo Urbina, Ricardo Vivancos-Pérez*

Center for the Study of Digital Libraries
Texas A&M University
College Station, TX 77843-3112, USA
{furuta, ssk9770, rajiv, e-urbina, rv} @csdl.tamu.edu

Abstract. The Cervantes Project, housed under the auspices of the Center for the Study of Digital Libraries at Texas A&M University, aims to provide a comprehensive on-line research and reference site on the life and works of the author Miguel de Cervantes Saavedra (1547-1616). This activity is a joint collaboration among researchers in the Department of Computer Science and the Department of Modern and Classical Languages, Texas A&M University. This paper outlines the work being conducted by the project, focusing on the creation of an Electronic Variorum Edition of Cervantes' *Don Quixote*.

1 Introduction

Miguel de Cervantes Saavedra (1547-1616), one of the world's greatest and most influential authors, is generally-recognized as a central figure in Hispanic literature and culture. His best-known work, *Don Quixote*, first published in two parts in 1605 and 1615, has been called the first modern novel and has been translated into more languages than any other literary text except the *Bible*.

The Cervantes Project, initiated in 1995, has as goal the creation of a comprehensive Web-accessible reference and research site dedicated to the study of Cervantes' works and life. To this end, the Cervantes Project is supporting a number of different components: the Cervantes Digital Library (CDL), an electronic collection of Cervantes' plays, novels, and other writings;[1] the Cervantes Digital Archive of Images (CDAI), an online archive of photographic images related to the

---

* Authors are listed in alphabetical order. Richard Furuta, Siddarth S. Kalasapur, and Rajiv Kochumman also are affiliated with the Department of Computer Science. Eduardo Urbina and Ricardo Vivancos-Pérez also are affiliated with the Department of Modern and Classical Languages. The project's Web pages are at http://www.csdl.tamu.edu/cervantes/

[1] The Cervantes Project and the CDL are unrelated to the University of Alicante's similarly-named "Biblioteca Virtual Miguel de Cervantes," which is building a general electronic library of Hispanic literature.

life and works of Cervantes; and the Cervantes International Bibliography Online (CIBO), a comprehensive annotated bibliography on the studies, works and life of Cervantes. All of these components are constantly growing, with new records being added to the bibliography on an annual basis, in parallel with the publication of the "*Anuario Bibliográfico Cervantino"*[2], the annual Cervantes bibliography.

A current focus of the project is the creation of an Electronic Variorum Edition (EVE) of *Don Quixote*. To this end we are developing computer-based tools that support the tasks of creating and accessing the EVE. These are a Multi-Variant Document Editor (MVED), which aids scholars in detecting and evaluating the differences among the various versions of *Don Quixote*, and the Virtual Edition Reader's Interface (VERI), which enables users to view texts along with the comments and emendations made by the scholars. Both these tools will be described in more detail subsequently.

2 Electronic Variorum Editions

The digital environment allows the conception of new document forms in support of scholarship. Once such form is the Electronic Variorum Edition (EVE)—an electronic edition containing all existing editions of a text, annotation of the variances present among the editions to allow for their comparison, derivative editions, generated as the result of scholarly analysis of the variances and bearing supporting reasoning, and scholarly commentary by expert editors that illuminates elements of the texts and of the comparisons among editions. The reader of the EVE should be able to customize the text presentation, perhaps selecting different interpretations for different applications, as well as annotate the results. Furthermore, all components in the EVE should be interlinked, allowing easy traversal among the representations.

The EVE is a general representation, allowing the specification of collections that parallel the traditional notions of Documentary Edition, Critical Edition, and Variorum Edition.[3] The application of these distinctions within the EVE is discussed in this section.

---

[2] The "*Anuario Bibliográfico Cervantino"* is published by the Universidad de las Islas Baleares (Spain). Three volumes (1996, 1998, and 1999) are available at present.

[3] Definitions, adopted and adapted from [11], for some of the terms we used are contained in this footnote. *Critical edition*: A scholarly edition that presents a text constructed by adopting readings from one or more documents and by correcting readings determined to be errors, and accompanied by apparatus explaining editorial principles and procedures, lists of emendations, and a historical collation of the text. *Documentary edition:* Also known as diplomatic edition; a scholarly edition that presents the text of a particular document without emendation. It includes an apparatus describing the document, the basis for its selection, the principles of transcription used, and a list of variants found in other documents. *Variorum edition*: A scholarly edition in which a base text (not necessarily critically edited) is annotated with a record of critical and textual commentary on particular passages, of editors' emendations, or of variant readings present in other texts. A critical variorum edition includes primarily critical commentary; a textual variorum edition primarily reports textual variation.

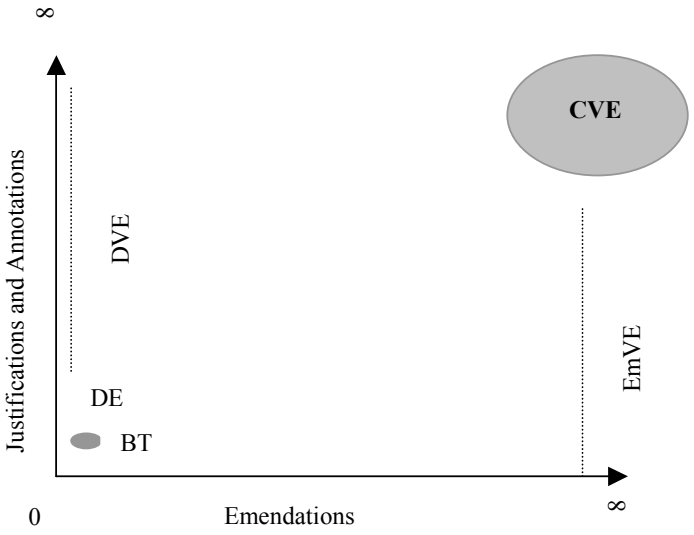

Figure 1. The relationship between Base Text (BT), Critical Variorum Edition (CVE), Documentary Variorum Edition (DVE), and Emended Variorum Edition (EmVE).

Any commentary made by an editor or a scholar that either describes the text or a variant, but does not make any changes to the text is called an annotation. On the other hand, the editorial alteration of the text to adopt readings from other documentary texts or to adopt readings not present in any document but arrived at through editorial conjecture is called an emendation. Thus while an annotation serves to describe or elaborate the underlying text, an emendation changes the same.

A direct transcription of an edition, without any modifications, is a Documentary Edition (DE). In terms of annotations and emendations, none exist as far as documentary editions are concerned. We identify one of the DEs to serve as a *base text* (BT). The base text is used as the source or reference to compare with other texts during the collation process, as will be elaborated later. During the process of editing the base text, or while comparing with other texts, scholars may annotate the base text, or make changes to the same, along with a justification for the corrections. An electronic edition that comprises of the base text, with changes made to the same is called an Emended Variorum Edition (EmVE). Cross-linking of a base text with other documentary editions and association of annotation constitutes a Documentary Variorum Edition (DVE). Finally, a collection of a base text, emendations of the base text, along with justifications, and annotations to either the text or the emendations, constitutes a Critical Variorum Edition (CVE). The Critical Variorum Edition, which includes all emendations and annotations, represents a scholar's complete interpretation of the underlying text. Figure 1 illustrates the relationships among these terms.

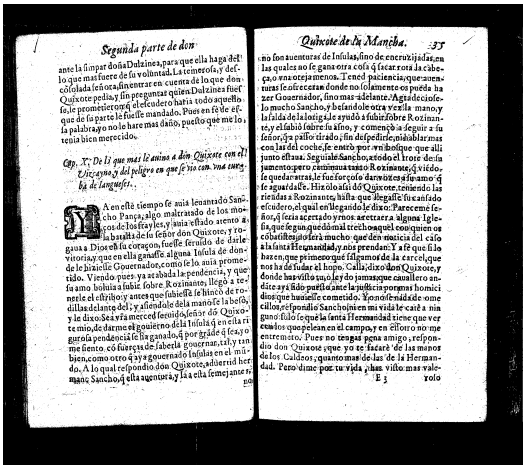

Figure 2. A representative page.

3 An EVE of *Don Quixote*

We are preparing an Electronic Variorum Edition of key editions of *Don Quixote*, specifically of the significant editions that were published during Cervantes' lifetime. Since no manuscripts are known to survive, the available textual resources for *Don Quixote* begin with the first published editions of its two volumes (these are called the *princeps*), which appeared in Madrid in 1605 and 1615. The significant editions for Volume 1, then, are Madrid 1605 (*princeps*), Madrid 1605 (2nd edition), Valencia 1605, Brussels 1607, and Madrid 1608 (3rd edition). For Volume 2, the significant editions are Madrid 1615 (*princeps*) and Brussels 1616. Additionally, a combined edition (Madrid 1637), containing both volumes, is included in our collection. Although this edition appeared after Cervantes' death, scholarly consensus (e.g., see [9]) indicates that its contents provide additional insight into the early development of the work. Each edition is about 700 pages long.

We have obtained multiple microfilmed copies of these editions. Because changes may be made during a print run, established scholarly practice is to consult multiple copies of an edition. We currently possess, and have digitized, six copies of each of the two *princeps* and at least two copies of each of the other editions. We hold microfilmed copies of two other editions (Lisbon 1605 and Madrid 1647), which we are not including in our initial efforts since scholarly analysis indicates that they are of limited significance. The Lisbon 1605 edition is a pirated reproduction of the Madrid 1605 and the Madrid 1647 combined edition adds little additional insight beyond that provided by Madrid 1637.

Figure 2 shows a typical page from our collection. Note that numerous artifacts exist on the image. Some artifacts are easy to handle, for example the dark region around the manuscript and the skewing of the image. Others are more difficult, for example the dark spots in the image's text area. Removing artifacts in the latter

category will take great care in processing as improper adjustment may result in changes to the text's semantics (see, for example, Donaldson's description of semantic differences introduced by interpretation of an ambiguously printed character as "f" or "s" in a Shakespeare text [1]). Processing of digitized microfilms raises questions that still remain unanswered. One clear implication, however, is that we will need to make images available in both their original and also their processed form.

Our use of microfilmed sources represents a compromise position. On the one hand, the image quality of the microfilms is not as good as would be obtainable from customized digitizing. On the other hand, microfilm copies often are already available and do not subject the original volumes to the potential of damage—a risk that many of the editions' owners are unwilling to take. The 1605 *princeps*, for example, is quite rare, with only 18 copies known to exist. Even with image compromises, our combined archive already is of great significance to the Cervantes scholar and far more extensive than that formerly available.

Given a collection of digitized editions, we must support several tasks in creating an EVE. The first is to interlink the editions and their different representations (textual, original image, and modified image). A second is support for the creation of a unified text. This requires first that the scholar selects the base text. Given the base text, it is then necessary to detect differences between it and the other texts in the collection, to enable the scholar's emendation of the base text (along with justification), and to allow the association of general commentary. This set of functions is supported by the MVED, which will be described in the next section.

Given the participation of one or more scholars serving as editors, we also must provide a means for readers to examine the texts and commentary. This is provided by the Reader's Interface, which will be discussed later in the paper

The initial implementations of the MVED and the VERI were completed by Shueh-Cheng Hu, and were reported earlier [5]. In the following sections, we review the interfaces and the modifications that we have made to them as we gained experience with their use.

3.1 MVED

The MVED (the Multi Variant Document Editor) is a software tool intended for use by scholars to aid them in collating different editions of the same text. It helps scholars in identifying, analyzing, and editing variances between a base text and different editions of the same.

3.1.1 Motivation behind the MVED

Access to old documents is rare and restricted. To prevent damage to these rare documents, and also to make them widely accessible, facsimile copies are taken and textual transcriptions are made. Scholars then need an interface they can use to relate the actual image and its textual transcription, and the MVED provides this facility.

Also, there might be multiple versions of a single document, with no sure mechanism to detect the original. In some cases like *Don Quixote*, the original itself might be lost. In such cases, scholars require a mechanism to compare the many

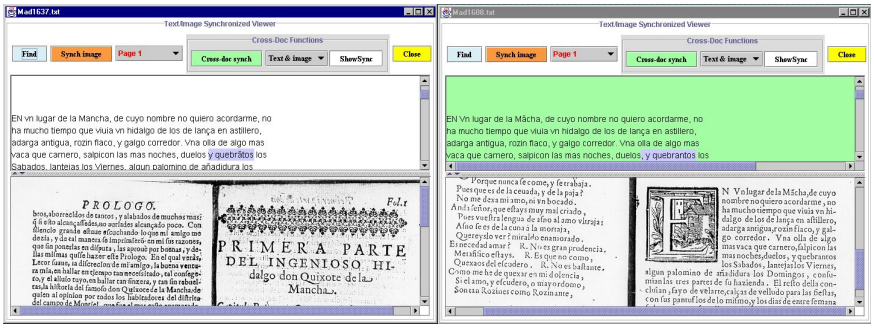

Figure 3. A collation session in progress, with the images and the corresponding texts displayed by the dual-form *document viewe*r of the MVED. The base text, on the right of the figure, is shown in with a uniquely colored background.

versions of the same document, and make emendations to the underlying text, with an aim to develop their understanding of the document. The MVED provides the facility to compare and edit multiple documents simultaneously, with provisions to associate annotations and emendations.

The MVED was thus developed with the intention of providing an environment for creating and presenting electronic variorum editions originating from document facsimiles stored in microfilms. Starting from a base text, the scholar can make emendations to the text, and create an Emended Variorum Edition. The scholar can also annotate the text, adding valuable commentary, without actually emending it, thereby creating a Documentary Variorum Edition. A combination of emendations, annotations and justifications yields what is called a Critical Variorum Edition. Readers can then read the newly created editions, using the VERI.

3.1.2 MVED Data Entities

MVED uses data entities like the textual image and its plain-text transcription, and creates additional data entities like editing records during the process of collation. It is necessary to maintain a relationship between these data entities, for example between the text portion and its annotation. The centralized data entity management framework within the MVED achieves just this, maintaining a tight coupling between the steps in the collation process. All data entities created within a collation session can be viewed within the *data entity browsing interface*, a component of the MVED that enables editors to view the existing entities and the relationships among them.

3.1.3 Components of the MVED

The MVED has a set of tools to aid the scholar in the collation process. It has a dual-form *document viewer* that displays both an image and its textual transcription (see Figure 3). The document viewer synchronizes the text and the corresponding image portion.

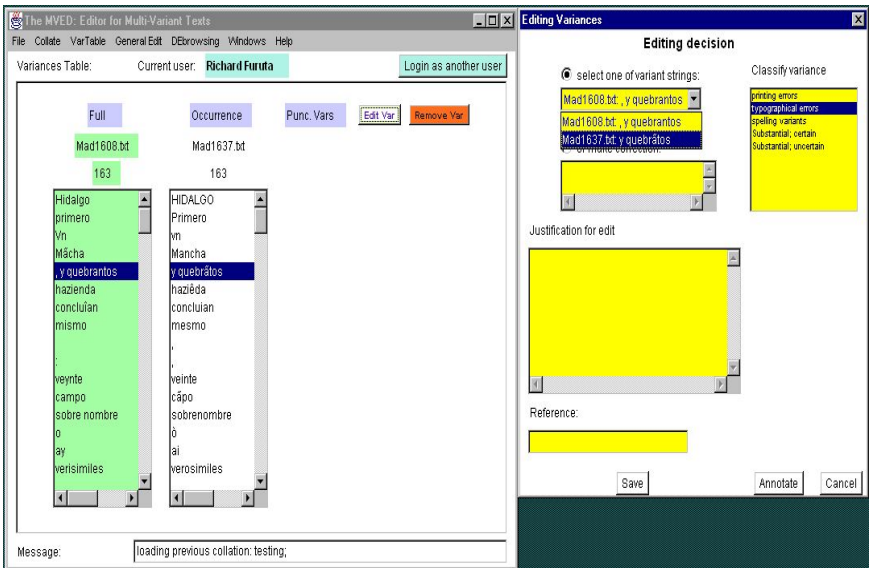

Figure 4. The left portion of the figure shows the list of variants resulting from Figure 3's collation. The scholar then uses the interface on the right to classify the variances and to specify emendations.

Synchronization between two texts under collation is achieved by means of the *text synchronizer*. This enables scholars to compare two or more texts simultaneously, so that variances can be detected and corrected.

Another tool called the *collator* helps in identifying the variances among the collated texts. This tool automatically identifies variances between the base text, and the other texts (called comparison texts) under collation, and presents the list of variances to the scholars, who can then make decisions on the validity of variants. This is illustrated in Figure 4, which shows automatically generated variances from all texts involved in the collation. The editing scholar can see the variances from different versions simultaneously in "full" form, or can view them in compact form, using an expandable tree structure. The MVED provides additional modifications to the display, for example detection (or not) of punctuation variances and case-sensitivity, so that the scholars can view variances in a flexible format.

At any time during the collation process, a comprehensive summary of the collation session in progress is available for the scholar, who can view the same using the earlier-described data entity browsing interface.

3.1.4 Annotations

Annotations are an important part of the new editions; annotations made by scholars are especially valuable because they give the reader an insight into the text. The MVED provides a facility for the user to attach annotations to a variance, or to a selected text portion, or to an emendation to the text. An annotation can be classified

as historical, geographical, cultural, biographical, linguistic, or literary, based on its significance, or may be attached to a base text or a variant without any classification (generally unclassified annotations can be avoided, since the classification categories can be extended by the scholar if necessary). Thus the MVED allows the editing scholar to not only annotate the base text itself, but also to annotate his own emendations of the text.

3.2 VERI

The Virtual Edition Reader's Interface (VERI) is a WWW front end for the data sets produced by the MVED, so that any user who wishes to view the results of collations can do so. It is instructive to compare the view of the information space as shown in Figure 1, as perceived within the MVED and as presented in the VERI. The goal of the editors using the MVED is to produce a Critical Variorum Edition. In accomplishing this task, an editor begins in the lower left hand corner of Figure 1, and associates annotations, emendations, categorizations and justifications while approaching the CVE in the upper right hand corner of the Figure. The reader, on the other hand, has goals that may favor examination of the CVE, of an EmVE, or of an intermediate state. Consequently, the VERI should support flexible traversal of the space represented in the Figure.

Our initial model of the VERI was intended to do just this, by providing a "category-centric" selection mechanism. In that version of the VERI, each of the categories of editing decisions made by an editor (or editors) could be manipulated separately. However, our evaluation of the interface demonstrated that exercising some degrees of flexibility result in semantically-inconsistent representations, as an editor's actions were generally unified rather than discrete. Consequently by default, we now support an "editor-centric" model for customizations in which the reader's initial selections track that of an individual editor selected by the reader. We also continue to explore representations that will enable the semantically-consistent comparison of editions created by different editors, in order to support development of a reader's insight into the methodologies employed by the different editors.

We continue to find additional dimensions for customizations. The international appeal of *Don Quixote* means that we must plan to support Spanish-only readers and scholars, as well as bilingual Spanish/English readers and scholars. In this context, selection based on the language used in annotations becomes a useful customization option.

We also need to incorporate a flexible security model into the VERI. Although *Don Quixote* is not subject to copyright protection, our digitized page images are subject to licensing agreements with the edition's holders. Access allowed to the images will need to respect the differing conditions of those agreements.

4 Electronic Documentary Editions

No commonly-available commercial OCR program currently is capable of producing reliable machine-readable files for seventeenth century texts such as *Don Quixote*. To

tackle an otherwise very laborious and time-consuming task, we have adopted a compromise approach to "reconstruct" the different editions by taking as our initial base text the Schevill and Bonilla's (Madrid, 1914-1941) electronic text, converted and edited by Jehle [10] and already in our Cervantes Digital Library. This initial text is then collated manually with the digitized images of the different copies and editions of *Don Quixote*, and variants are recorded and introduced into the text to produce multiple electronic documentary editions. Similarly, manual collation of the non-*princeps* editions with the electronic version produces electronic documentary editions for them as well.

Three set of manual collations are required: 1) the preliminary collation of the Spanish National Library's unique copy of the 1605 *princeps* edition; 2) the *princeps* collation, using the documentary edition developed in the first step along with five other *princeps* copies obtained in microfilm from libraries around the world; and 3) the collation of the *princeps* text produced by the previous collation with the other early editions published between 1605 and 1637 in order to produce documentary texts of each one of them.

The preliminary collation aims at producing an "old-spelling" documentary edition in electronic text format of the National Library of Spain's copy of the *princeps*, by removing all the changes and emendations introduced by Schevill and Bonilla. The results of this collation are then compared with Flores' text [2] as a control, and the final version is saved in electronic format to be used as the base text for the MVED, and to be manually collated against the other copies of the *princeps* to produce their electronic text representations. Finally, the MVED-produced unified text from the second collation is used as base text for MVED collation against other non-*princeps* editions

All manual collations are done chapter by chapter, and are checked at least thrice. Once the preliminary collation has produced an electronic documentary text for the National Library copy, we can easily reconstruct the other five copies by identifying the variants and, therefore, creating a documentary edition of all copies of the 1605 *princeps*. So far, this process has been completed for 42 of the 52 chapters of *Don Quixote.*

The final manual collation involves the production of multiple electronic texts of the significant editions identified earlier in this paper. The electronic texts resulting from these different sets of collations are introduced into the MVED and form the basis for the critical variorum edition, which can then be accessed through the reader's interface.

4.1 Observations about Source Materials

Many problems arise regarding spelling and other variants among the different copies. In the preliminary collation we are dealing with an old-spelling emended text in which Schevill and Bonilla also incorporated their own punctuation parameters. For example, they use different capitalization criteria and semicolons, a punctuation mark that was not in vogue in 1605 and 1615, when *Don Quixote* was originally published. Since our purpose is to produce a documentary edition—an edition that preserves all the characteristics and errors of the original, we must encode marks and symbols

which are no longer used, such as the long intervocalic ‘s’ and the abbreviation of certain vowels and consonants represented by a tilde ‘~’.

Another interesting observation is the presence of *press variants*. These are textual variants that arise during a press run, such as those caused by a stop press correction. For instance, it is not unusual to find letters that have been printed twice or upside down type.

Finally, difficulties arise concerning the use of accents, since three different kinds of accent marks were issued by the composers: the acute accent ‘á’, the grave accent ‘à’, and the circumflex accent ‘â’. There are instances where such accents seem to appear, but are in fact caused when facsimile copies of microfilms were produced. In other words, the original text does not contain these accents, but imperfections in the reproduction process or subsequent aging of the microfilm made it seem that they were present. Apparent accents may also be caused if the type used during printing was broken, or worn, or if for some reason, the inking unit was not working properly.

5 Discussion and Conclusions

The past few years have seen a blossoming of research efforts devoted to producing state-of-the-art network-accessible digital libraries over humanities-based materials [7]. Just a few of the many notable examples include the Perseus Project[4], components of the Library of Congress’ American Memory project[5], the Shakespeare Electronic Archive[6], and the Canterbury Tales Project[7]. These activities are historically-grounded in the much longer-lived decades-long activities directed towards humanities computing. Taken as a whole, these activities provide compelling evidence of the strength of humanities applications as foundation for novel research in digital libraries.

The Cervantes Project continues to teach us valuable lessons, both technical and organizational. From a technical perspective the topic domain of early modern Spanish texts frequently illustrates the design assumptions of the tools that we use. For example, searching utilities frequently are character-set neutral (we use the MG system [8, 11], but expect that these observations apply more generally). Because of our multilingual reader population, we are considering “folding” accented and non-accented characters together in search queries. This is firstly because English speakers often do not distinguish the accents and leave them out when typing. In addition, old Spanish includes accent marks and abbreviations unused, and consequently unexpected, by modern Spanish speakers, as suggested in the previous section’s discussion. Often, the solution adopted here is to “modernize” the spelling, but this approach seems unattractive to us, as it discards information that may be of use to the serious scholar.

The world-wide popularity of Cervantes and of *Don Quixote* requires that we provide our materials in many forms and in multiple languages. Earlier we discussed

[4] http://www.perseus.tufts.edu/
[5] http://memory.loc.gov/
[6] http://shea.mit.edu/
[7] http://www.cta.dmu.ac.uk/projects/ctp/

a scheme for tracking changes in parallel versions of a Web site collection [4]. We also are investigating approaches towards specifying different "cuts" of our collection for different classes of readers (e.g., high school students, university students, university researchers, and the general public).

From an organizational standpoint, we continue to be interested in identifying ways to help achieve successful interdisciplinary collaboration. The easy observation is that a successful collaborative project requires an application that is meaningful and challenging to all participants when viewed from the context of their separate disciplines. Additionally, the participants must invest significant effort in understanding the others' areas and assumptions, and must remain open to reexamination of their own area's seemingly pre-established tenets.

The best methodology (perhaps the only successful methodology) we have been able to find for fostering cross-discipline understanding is a commitment to frequent and often lengthy whole-group meetings. Even so, after more than six years of collaboration we still encounter issues that we discover, after lengthy discussion, reflect differences in our own terminology and expectations, rather than fundamental differences in viewpoint. A recent multi-week discussion, for example, centered around differing definitions of terms such as "edition" and "annotation"—terms for which we had assumed we had a common understanding.

We continue to find rewards in the Cervantes Project. From the point of view of the Humanities, the materials that we are collecting and the tools that we are creating hold promise of making significantly broader resources available to scholars of Cervantes. This in turn has potential for modifying the ways in which research and education take place. From the point of view of Computer Science, the project provides a concrete testbed that supports the development and evaluation of tools and representations of a richly structured information collection. This provides the basis for investigation of Digital Libraries techniques with the potential for quite broad applicability.

Acknowledgements

We acknowledge Shueh-Cheng Hu's contributions to the project, completed in conjunction with his dissertation research [6]. This material is based upon work sponsored by the National Science Foundation under Grant No. IIS-0081420. Support for this work was provided (in part) by the Interdisciplinary Research Initiatives Program, administered by the Office of the Vice President for Research, Texas A&M University.

References

1. Peter S. Donaldson, "Digital Archive as Expanded Text: Shakespeare and Electronic Textuality." In Katheryn Sutherland, editor, *Electronic text: Investigations in Method and Theory*. Oxford Clarendon Press, 1997, pp 173-197.

2. *Don Quixote de la Mancha. An old-spelling control edition based on the first editions of Parts I and II*. Prepared by R. M. Flores. Vancouver: University of British Columbia Press, 1988, 2 Vols.
3. Richard Furuta, Shueh-Cheng Hu, Siddarth Kalaspur, Rajiv Kochumman, Eduardo Urbina, and Ricardo Vivancos-Pérez, "Towards an Electronic Variorum Edition of Don Quixote", JCDL 2001: *The Joint ACM-IEEE Conference on Digital Libraries*, June 2001, to appear.
4. Shueh-Cheng Hu and Richard Furuta. "A Tool for Maintaining Multi-variant Hypertext Documents," in Roger D. Herch, Jacques André, and Heather Brown, editors, *Electronic Publishing, Artistic Imaging, and Digital Typography (7th International Conference on Electronic Publishing, EP'98, held jointly with the 4th International Conference on Raster Imaging and Digital Typography, RIDT'98),* Springer 1998 (Lecture Notes in Computer Science #1375), pp. 525-536.
5. Shueh-Cheng Hu, Richard Furuta, and Eduardo Urbina, "An Electronic Edition of Don Quixote for Humanities Scholars". *Document Numérique*, 3(1–2), June 1999, pp. 75–91.
6. Shueh-Cheng Hu, *Towards an Electronic Variorum Edition Originating from Available-Quality Document Facsimiles*, Ph.D. dissertation, Texas A&M University, College Station, Texas, December 2000.
7. Cruz Yolanda Lugo Ibarra, *Don Quixote in the digital age: An analysis of traditional editorial practices and current electronic editions*, M.A. thesis, Texas A&M University, College Station, Texas, December 1999.
8. New Zealand Digital Library homepage at http://www.nzdl.org/
9. Francisco Rico, "Historia del texto," Prólogo, *Don Quijote de la Mancha*, Francisco Rico, dir., 2 vols. (Biblioteca Clásica 50, y Vol. Complementario). Barcelona: Instituto Cervantes-Crítica, 1998. 1: cxcii-ccxlii.
10. http://www.csdl.tamu.edu/cervantes/english/ctxt/sb/. Digital texts converted by Eduardo Urbina and Fred Jehle.
11. William Proctor Williams and Craig S. Abbott, *An Introduction to Bibliographical and Textual Studies*, Third Edition. Modern Language Association of America, 1999, pp. 69–125.
12. I. H. Witten, A. Moffat, and T.C. Bell, *Managing gigabytes: compressing and indexing documents and images***,** Morgan Kaufmann, San Francisco, CA, 1999.