Tools for a New Generation of Scholarly Edition Unified by a TEI-based Interchange Format

Rajiv Kochumman, Carlos Monroy, Jie Deng, Richard Furuta, and Eduardo Urbina
TEES Center for the Study of the Digital Libraries
Texas A&M University
College Station, TX 77843-3112, USA
(979) 845-3839
cervantes@csdl.tamu.edu

ABSTRACT

We report on experience gained from our ongoing multi-year project to produce an Electronic Variorum Edition of Cervantes' *Don Quixote de la Mancha*. Initially designed around a custom database representation, the project's evolution has lead to the adoption of a TEI-based format for information interchange among the project's major components. We discuss the mechanics of this approach and its benefits.

Categories and Subject Descriptors

I.7.2 [**Document and Text Processing**]: Document Preparation - format and notation, hypertext/hypermedia, markup languages.
J.5 [**Arts and Humanities**]: Literature.

General Terms

Design, Standardization.

Keywords

Cervantes Project, TEI, text encoding.

1. INTRODUCTION

Today's technology makes feasible scholarly editions of a scale previously unrealistic to achieve. Printed editions restricted the amount of material that could be added for practical and financial reasons. Digital editions promise to catalyze fundamental changes in scholarly practices as they are not subject to these restrictions and can include complete facsimiles of original sources, permit readers to tailor the form of display of the text and the level of detail of commentary, enable sophisticated automated analysis and visualization of textual elements, and in general can place the reader on an equal scholarly footing with the editor.

Previously [4], we have described the tools developed by our Cervantes Project to create an Electronic *Variorum* Edition (EVE) of *Don Quixote* based on facsimiles of the early significant editions, published between 1605 and 1637. In the process of carrying out this project, we have digitized and transcribed microfilmed copies of over 40 copies of the editions, including 9 each of the rare initial printings of the book's two parts. Our collection represents a resource that far exceeds that previously available to any Cervantes scholar.

To create the EVE, we created two customized tools: a stand-alone application called the MVED (Multi-Variant Editor for Documents) and a Web-based viewer called the VERI (Virtual Edition Reader's Interface). The MVED is used to electronically collate editions of the text. The result of the collation is a set of textual variants linked with their corresponding textual transcriptions and synchronized with the facsimile images. An editor can then classify the variants, emend and annotate the texts, and thereby create an EVE. The VERI allows readers to navigate and browse through the texts, images, and compose virtual editions from the EVE.

Since the MVED and the VERI were developed for the purpose of creating this specific EVE, the functions included in the MVED were developed from the task's viewpoint. In addition to viewing aligned text and image, the MVED allows the editor to carry out three basic operations on the text, using the variants identified by the collation as a guide:

Emendation: when an editor is certain that a variant is the result of an error in the transcription or the facsimile image is not very clear, he or she emends the text.

Correction: for each variant, the editor selects one and thus corrects the text. Corrections are classified into five categories: a) printing errors, b) typographical errors, c) spelling variants, d) certain substantive (resolved) variants, and e) uncertain substantive (unresolved) variants.

Annotation: editors can annotate the texts in three different ways: a) annotate free text, b) annotate the process of correcting text, and c) annotate the process of emending text. Since all three can be done on the same segment of the text, multiple annotations on the same segment are allowed. Annotations are classified into four categories: a) historical, b) geographical, c) cultural, and d) other.

The MVED's operations are stored in a database and do not modify the original text.

2. INTERMEDIATE FORMAT

In our original implementation, the information representation used by the MVED and VERI was based on their custom-designed database—i.e., the VERI read the MVED-created database. To increase the potential applications of the editions prepared in the MVED and the potential sources of materials to be

JCDL'04 June 7-11, 2004, Tucson, Arizona, USA.
Copyright 2004 ACM 1-58113-832-6/04/0006…$5.00.

displayed in the VERI, we decided to adopt an externally visible intermediate representation using the TEI encoding [7]. The applicability of documents encoded in the TEI for visualization and analysis is well-known; see for example [1], [2], and [3].

We convert our internal database representation to TEI using a tool we named Text2TEI. There are 5 levels of encoding within the TEI [8]. We proceed level by level till all the levels are covered. To validate TEI compliance, we used commonly available TEI decoders [9] to decode our document.

Level 1 of the TEI differentiates the text into header and body. This is a structural division. For our texts, we insert a TEI header, and the rest of the text follows in the body portion. This is the minimal encoding needed for a text to be TEI-compliant.

Level 2 includes chapter headings, and book headings. This enhances the structural division derived from Level 1. In the case of *Don Quixote*, the texts itself provides an easy pattern for recognizing these headings.

Level 3 adds annotations to the text. We make use of the results of the collation, and pick up the annotations from the database repository. The TEI tag '<note>' is used for this purpose. This tag has attributes such as 'type' to classify the annotation, and 'resp' to specify the person responsible for the annotation.

Level 4 allows us to insert corrections into the text. As with the annotations in Level 3, the corrections are picked up from the database repository where they exist following a collation. Although this level is aimed primarily at correcting typographical errors, we use it to correct the other types of errors are well. The '<corr>' tag is used in corrections. To annotate the process of correction, we simply use a '<note>' tag with the correction itself. Note that the correction tag in TEI does not provide the functionality to annotate the correction itself.

Level 5: The highest level of the TEI specifies that the corrections applied in the TEI document go beyond simple content analysis, and include scholarly knowledge of the text as well. This, we contend, is already achieved at levels 3 and 4. The annotations and corrections present in the database are the results of a collation done by Cervantes scholars. It reflects their knowledge of the DQ texts, and of ancient Spanish literature. The critical commentary provided within the collation goes beyond basic content analysis. Hence these reside at Level 5 compliance of the TEI guidelines.

We also are using level 5 encodings to represent a set of elements representing the various episodes and adventures in the *Quixote*. While these elements are generally accepted by Cervantes experts, a comprehensive taxonomy is being created for the first time by our project. The taxonomy is essential due to the extraordinary popularity of the *Quixote*—for example, in addition to editions in almost every imaginable language, over 80 operas and 100 films have been based on the *Quixote*. Hypertextual connections among these many representations cannot be achieved without the presence of a taxonomy such as ours.

A new activity in the project is enabled by the taxonomy. Although the original versions of the *Quixote* were not illustrated, the later depictions were a major contributing element to both the canonization of the novel and to the iconic transformations of its principal character. This rich critical and artistic tradition remains largely unknown due in great part to the rare and inaccessible nature of the editions in which the thousands of illustrations have appeared that constitute the work's visual narrative and interpretation.

At present, the project and our university's Cushing Memorial Library have acquired over 245 editions of the *Quixote* published since 1620 [5][6]. The collection comprises over 566 volumes and is concentrated in 18th and 19th century English, French, and Spanish illustrated editions. The estimated 8,000 images in these volumes, coupled with the encodings described here, will form the basis for a new focus on the textual iconography of the *Quixote*.

3. CONCLUSION

We believe that our efforts will help to shape the scholarly practices of the next generation of humanities researchers. The inclusion of an openly accessible intermediate interchange format is a key component in this effort, as digital representation affords widespread dissemination, interlinking, and analysis of scholarly works *and* of the volumes of original source material unavailable previously because of cost.

4. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. IIS-0081420. Support for this work was provided by the Humanities Informatics Initiative funded by the Office of the Vice President for Research through a Telecommunications and Informatics Task Force grant, with additional support from the Cervantes Chair, University of Castilla-La Mancha, College of Engineering, the College of Liberal Arts, and the Texas A&M University Libraries.

5. REFERENCES

[1] Crane, G., and Wulfman, C. Towards a Cultural Heritage Digital Library. *Proceedings of the third ACM/IEEE-CS joint conference on Digital Libraries*. May 2003. pp. 75-86.

[2] Fan, B. Women Writers Project. *Crossroads*, 6(2), winter 1999, pp. 19-23.

[3] Fekete, J., and Dufournaud, N., Compus: Visualization and Analysis of Structured Documents for Understanding Social Life in the 16th Century. *Proceedings of the fifth ACM Conference on Digital Libraries*. July 2000, pp. 47-45.

[4] Furuta, R., Siddarth, S. S., Kochumman, R., Urbina, E., and Vivancos-Pérez, R. The Cervantes Project: Steps to a Customizable and Interlinked On-line Electronic Variorum Edition Supporting Scholarship. *Research and Advanced Technology for Digital Libraries: 5th European Conference, ECDL 2001*, 2001, pp. 71-82.

[5] Urbina, E., and Smith, S. The Grangerized Copy of John Bowle's Critical Edition of Don Quixote at the Cushing Memorial Library of Texas A&M University. *Cervantes: Bulletin of the Cervantes Society of America* 23.2, 2003, pp. 85-118. Forthcoming 2004.

[6] Urbina, E., Monroy, C., and Furuta, R. Iconografía textual del *Quijote*: repaso y nueva aproximación de cara al IV centenario. *Limine al IV Centenario del* Quijote. Coloquio Internacional de la Associazione Cervantina di Venecia, Ateneo Veneto. Venice, Italy, April 2003.

[7] The TEI Consortium web site http://www.tei-c.org/ [Date accessed: January 2004]

[8] Guidelines for Best Encoding Practices, TEI Text Encoding in Libraries, http://www.indiana.edu/~letrs/tei/ [Date Accessed: April 3, 2003]

[9] TEI software utilities. http://www.umanitoba.ca/faculties/arts/linguistics/russell/ebenezer.htm [Date Accessed: February 10, 2004