Appeared in "Document Numérique", 3 (1-2), June 1999, pp. 75-91. Issue on "Les documents anciens".

An Electronic Edition of *Don Quixote* for Humanities Scholars

Shueh-Cheng Hu[*] — Richard Furuta[*] — Eduardo Urbina[**]

** Center for the Study of Digital Libraries*
and Department of Computer Science
Texas A&M University
{shuehu,furuta}@csdl.tamu.edu

Department of Modern & Classical Languages
Texas A&M University
e-urbina@tamu.edu

ABSTRACT. *A number of steps are required to create comprehensive, flexible electronic editions of ancient documents, especially those whose visual quality has degraded and those available in multiple versions. Besides acquisition and display of materials, support is required for interrelating variants, and interfaces are needed to allow deriving and justifying new editions. Furthermore, lower-level mechanisms are needed to allow the maintenance of discovered and specified relationships. Although solutions can be found in other applications for handling some portions of the specific steps, still there are challenges to meet the steps' new requirements and to integrate the individual steps seamlessly. This paper presents an overview of a work-in-progress to create an integrated system for processing printed documents with degraded visual quality and multi-variant contents. Additionally, the implementation of the three modules enabling creation of new editions, providing underlying database support, and customizing hypertext documents for readers are described in detail.*

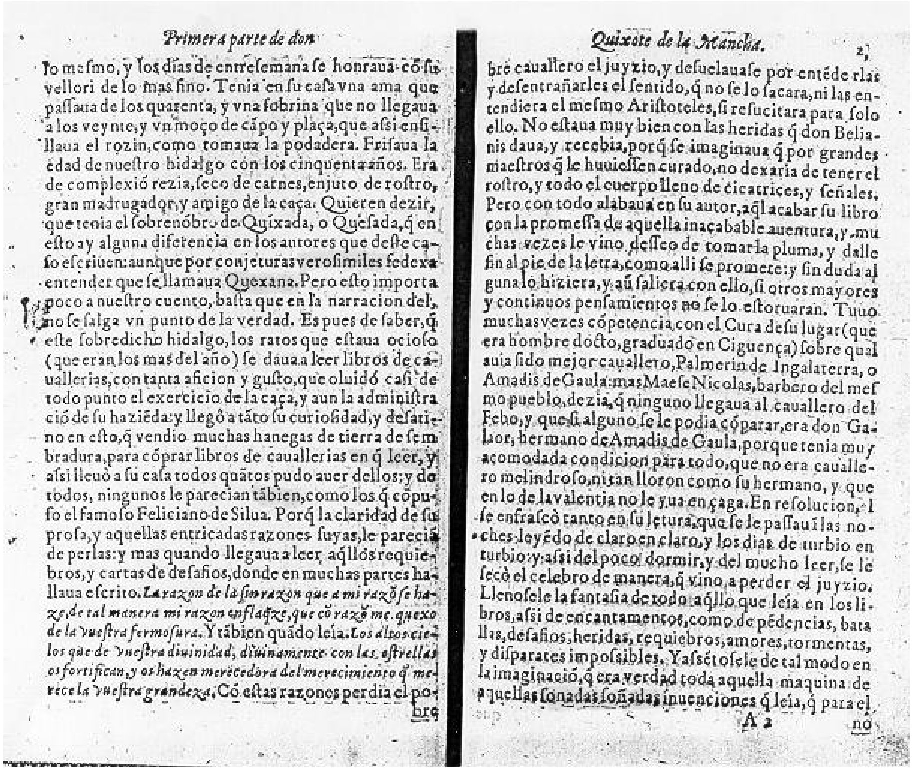
RÉSUMÉ. *à traduire*

KEYWORDS: *Multi-variant documents,* variorum *edition, user interface, data entities, hypertext*

MOTS-CLÉS : *à traduire*

1. INTRODUCTION

For several years, we have been building a digital archive based around the works of Miguel de Cervantes Saavedra (1547–1616), the author of *Don Quixote de la Mancha*. *Don Quixote*, often recognized as the first modern novel, provides illustration of many of the issues involved in creation of electronic editions. No manuscript of the work has survived, only a number of editions published during Cervantes' lifetime. The available images of these manuscripts are duplicated from archival microfilms and are typical quality for many rare and important documents received from distant libraries; while readable they reflect severe compromises in image quality. Unfortunately, due to administrative or budgetary restrictions, the availability of better copies is unlikely due to both the age of original material and the rareness of the original texts; the holders of rare printed texts usually are unwilling to subject them to additional scanning or copying.



Primera parte de don

lo meſmo, y los dias de entreſemana ſe honraua cõ ſu vellori de lo mas fino. Tenia en ſu caſa vna ama que paſſaua de los quarenta, y vna ſobrina que no llegaua a los veynte, y vn moço de cãpo y plaça, que aſsi enſillaua el rozin, como tomaua la podadera. Friſaua la edad de nueſtro hidalgo con los cinquenta años. Era de complexiõ rezia, ſeco de carnes, enjuto de roſtro, gran madrugador, y amigo de la caça. Quieren dezir, que tenia el ſobrenõbre de Quixada, o Queſada, q̃ en eſto ay alguna diferencia en los autores que deſte caſo eſcriuen: aunque por conjeturas veroſimiles ſe dexa entender que ſe llamaua Quexana. Pero eſto importa poco a nueſtro cuento, baſta que en la narracion del, no ſe ſalga vn punto de la verdad. Es pues de ſaber, q̃ eſte ſobredicho hidalgo, los ratos que eſtaua ocioſo (que eran los mas del año) ſe daua a leer libros de cauallerias, con tanta aficion y guſto, que oluidó caſi de todo punto el exercicio de la caça, y aun la adminiſtraciõ de ſu haziẽda: y llegó a tãto ſu curioſidad, y deſatino en eſto, q̃ vendio muchas hanegas de tierra de ſembradura, para cõprar libros de cauallerias en q̃ leer, y aſsi lleuó a ſu caſa todos quãtos pudo auer dellos: y de todos, ningunos le parecian tã bien, como los q̃ cõpuſo el famoſo Feliciano de Silua. Porq̃ la claridad de ſu proſa, y aquellas entricadas razones ſuyas, le parecian de perlas: y mas quando llegaua a leer aq̃llos requiebros, y cartas de deſafios, donde en muchas partes hallaua eſcrito. *La razon de la ſinrazon que a mi razõ ſe haze, de tal manera mi razon enflaqze, que cõ razõ me quexo de la vueſtra fermoſura.* Y tãbien quãdo leia. *Los altos cielos que de vueſtra diuinidad, diuinamente con las eſtrellas os fortifican, y os hazen merecedora del merecimiento q̃ merece la vueſtra grandeza.* Cõ eſtas razones perdia el po-

bre

Quixote de la Mancha. 2

bre cauallero el juyzio, y deſuelauaſe por entẽderlas y deſentrañarles el ſentido, q̃ no ſe lo ſacara, ni las entendiera el meſmo Ariſtoteles, ſi reſucitara para ſolo ello. No eſtaua muy bien con las heridas q̃ don Belianis daua, y recebia, porq̃ ſe imaginaua q̃ por grandes maeſtros q̃ le huuieſſen curado, no dexaria de tener el roſtro, y todo el cuerpo lleno de cicatrices, y ſeñales. Pero con todo alabaua en ſu autor, aq̃l acabar ſu libro con la promeſſa de aquella inacabable auentura, y muchas vezes le vino deſſeo de tomar la pluma, y dalle fin al pie de la letra, como alli ſe promete: y ſin duda alguna lo hiziera, y aũ ſaliera con ello, ſi otros mayores y continuos penſamientos no ſe lo eſtoruaran. Tuuo muchas vezes cõpetencia con el Cura de ſu lugar (que era hombre docto, graduado en Ciguença) ſobre qual auia ſido mejor cauallero, Palmerin de Ingalaterra, o Amadis de Gaula: mas Maeſe Nicolas, barbero del meſmo pueblo, dezia, q̃ ninguno llegaua al cauallero del Febo, y que ſi alguno ſe le podia cõparar, era don Galaor, hermano de Amadis de Gaula, porque tenia muy acomodada condicion para todo, que no era cauallero melindroſo, ni tan lloron como ſu hermano, y que en lo de la valentia no le yua en çaga. En reſolucion, el ſe enfraſcò tanto en ſu letura, que ſe le paſſauã las noches leyẽdo de claro en claro, y los dias de turbio en turbio: y aſsi del poco dormir, y del mucho leer, ſe le ſecò el celebro de manera, q̃ vino a perder el juyzio. Llenoſele la fantaſia de todo aq̃llo que leia en los libros, aſsi de encantamentos, como de pẽdencias, batallas, deſafios, heridas, requiebros, amores, tormentas, y diſparates impoſsibles. Y aſsẽtoſele de tal modo en la imaginaciõ, q̃ era verdad toda aquella maquina de aquellas ſonadas ſoñadas inuenciones q̃ leia, q̃ para el

A 2 no

Figure 1. *Textual image of a page from the 1605* princeps *edition of Don Quixote*

Figure 1 shows the image of one page, converted from a 35mm microfilm, of the 1605 *princeps* edition of *Don Quixote*. Obviously, the quality of this facsimile image presents significant challenges for content acquisition as well as presentation. As Figure 1 shows, unacceptable visual quality results from background noise, shadows along spine and edges, distortions to the image introduced in copying, written annotations, and library stamps. Further challenges to automatic processing are caused by the unusual letter forms used in classical Spanish and typesetting conventions no longer used today. The effect of document image quality on the correction rate for optical character recognition (OCR) is significant, especially when the image quality degrades below specific thresholds [GAR 98, CAS 90]. A poor recognition rate increases the acquisition cost dramatically due to necessary heavy manual post-processing work such as proof-reading and correction.

As a separate factor, important texts in the humanities often have been preserved in multiple early versions. Even when the author's original manuscripts are no longer available, insights into the work can be gained by comparison of different impressions of the first edition and of editions prepared while the author was still alive. For example, there are eleven significant versions[1] of *Don Quixote de la Mancha*, including the *princeps* – the earliest printed edition, published in Madrid, 1605. Many *variances*, discrepant text segments in different versions derived from an identical original counterpart, exist among the different copies of the same edition.[2] Choosing one variance over another is the job of a Cervantes scholar, and different scholars may, of course, make different choices based on their interpretation of the raw material. Our acquisition process supports the specification of such choices, allowing the categorization of the choice and association of justification for it.

Our project team, comprised of researchers from the areas of Computer Science and Spanish Literature, is seeking to produce a unified, electronic *variorum* edition of *Don Quixote*. The virtual variorum edition will contain multiple copies of the 11 significant early editions of *Don Quixote*, annotation of the variances present among the editions to allow their comparison, derivative editions, generated as the result of scholarly analysis of the variances and bearing supporting reasoning, and scholarly commentary by experts. The reader of a virtual variorum edition will 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 virtual variorum edition will be interlinked, allowing easy traversal among the representations. To provide raw materials for the virtual variorum edition, we have obtained microfilmed copies of the text from multiple collections (the Spanish National Library, the Hispanic Society of America, the British Museum, Yale University, Harvard University, and others still in progress), which we are converting to digital images.

[1] These editions are, for volume 1: Madrid 1605 (*princeps*); Madrid 1605, 2nd ed.; Valencia 1605; Brussels 1607; Madrid 1608, 3rd ed.; Madrid 1637 (combined edition); Madrid 1647 (combined edition). For volume 2, they are: Madrid 1615 (*princeps*); Brussels 1616; Madrid 1637 (combined edition); Madrid 1647 (combined edition).

[2] According to an analysis done manually, there are 20 significant variances found in chapter one alone, excluding variances in the use of punctuation marks. There are total of 126 chapters in *Don Quixote*.

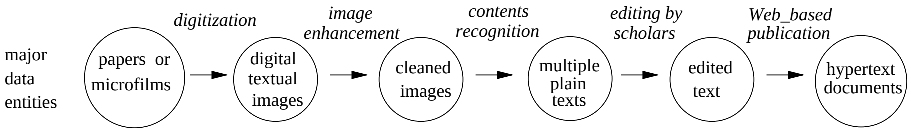


Figure 2. *General process of digitizing degraded printed texts with multi-variant contents*

In this paper, we report on initial experiences based on work with the early editions of *Don Quixote de la Mancha*. These experiences have lead to development of an architecture that allows economical conversion to digital form of ancient documents with degraded visual quality and multi-variant contents.

2. PREVIOUS WORK AND PROPOSED SOLUTIONS

Some effort has been devoted to the research of acquisition, preparation, and presentation of digital libraries material [BRE 96, SEA 97, LEC 98]. Issues about how to deal with texts with degraded image quality and multi-variant contents also have been discussed [GLA 98, BLA 95]. To offer better solutions for processing printed documents with degraded visual quality and/or multi-variant contents, methods presented in the prior work can be extended. In addition, new methods need to be developed to meet new requirements. An important requirement is to retain linkage among the related data entities that are generated during different processing steps. For instance, variances are identified during a collation phase and editing justifications are specified during a later editing phase. A relationship exists between a variance and the editing justifications that resulted from its analysis. It is important, therefore, to retain the correct relationship among data entities that might be modified or removed after generation. For instance, an update to a variance should lead to a corresponding review of the related editing justifications, or an inconsistency might be incurred. If there is no mechanism for retaining relationships among data entities, it will be difficult to maintain the consistency of related data entities and to present the inter-related data entities to readers correctly. To overcome the above drawback, an appropriate integration among individual steps is necessary, data entities as well as relationships among them will be retained and controlled in the integrated environment.

3. OVERVIEW OF THE PROPOSED SYSTEM

Although there might be differences among the procedures for handling various texts, a procedure for digitizing printed texts with poor visual quality and multi-variant contents can be generalized, as shown in Figure 2. To perform digitization work that follows the above procedure, a system, illustrated in Figure 3, is proposed. A num-

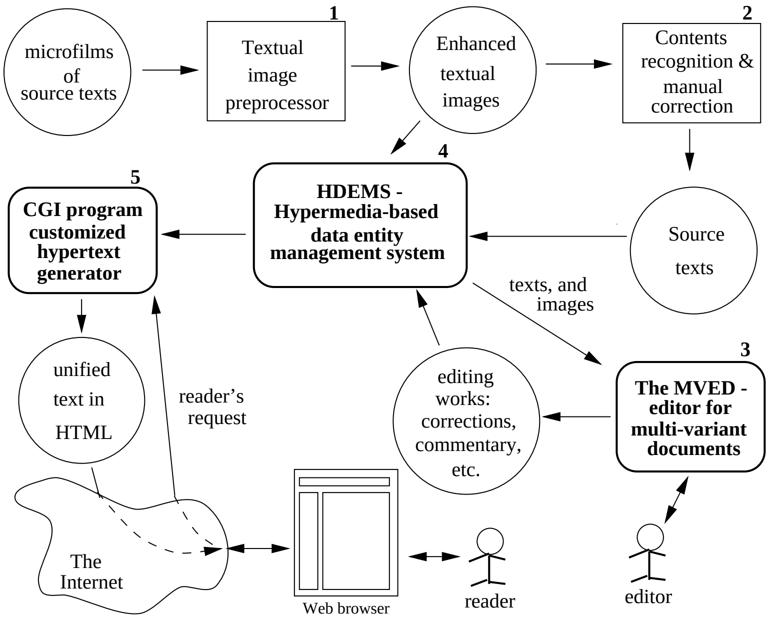

microfilms of source texts
1
Textual image preprocessor
Enhanced textual images
2
Contents recognition & manual correction
4
HDEMS - Hypermedia-based data entity management system
5
CGI program customized hypertext generator
Source texts
texts, and images
3
The MVED - editor for multi-variant documents
unified text in HTML
reader's request
editing works: corrections, commentary, etc.
The Internet
Web browser
reader
editor

Figure 3. *Architecture of the system for digitizing degraded printed texts with multi-variant contents*

ber of functional modules in the system are necessary in the procedure, which are described as follows: (1) After digitization, the document image enhancement module is provided to reconstruct raw images to obtain sufficient quality for Web presentation. Additionally, reconstructed images improve the accuracy rate of the later contents recognition process. Reconstructed document images also are important for later editing work because scholars need them to verify correct conversion from document images to the corresponding plain texts and to obtain additional clues that are not present in the plain text form. (2) To cope with the degraded visual quality of available facsimile images, a custom recognition module is required. The recognition module is expected to be more tolerant of noise in document images by making use of known features of the printed text, such as the lexicon and other higher level linguistic information in the processed text. (3) An editing tool (the MVED) is under development to support the simultaneous editing of texts with multiple versions, which is not supported by general editing tools. The editing tool needs to facilitate the work of locating, analyzing and classifying variances among different versions; comparing, contrasting, and linking versions that could be in different formats (either facsimile image or plain text). (4) A Hypermedia-based Data Entity Management System (HDEMS) is required for controlling access of data entities, recording derivation relationships among data entities, and keeping related data entities in consistent state. In this paper, data entities are called related if there exist relationships among them.

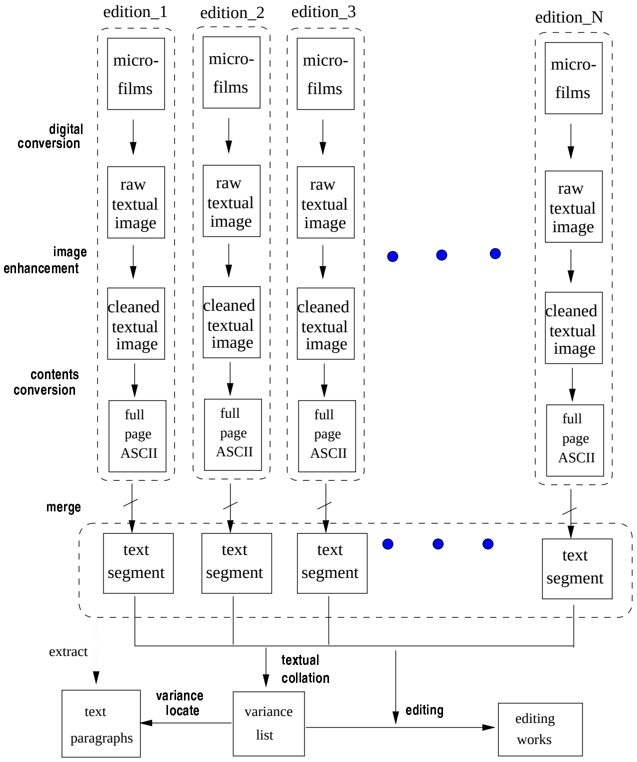


Figure 4. *Flow chart of data entities during the process of digitizing printed texts*

Inter-relationships among data entities are created due to two major reasons: format mutation and contents derivation. As Figure 4 shows, a large volume of various kinds of data entities are generated by different functional modules during the process of digitizing degraded printed texts with multi-variant contents. Thus, a framework like the HDEMS is necessary to prevent inconsistencies among related data entities during the digitalization process. (5) A module for customizing hypertext documents is necessary to serve readers with various backgrounds and different interests. Through adaptable documents, readers can view different levels of material that includes editing rationale, variances among multiple versions, and original facsimile images, besides the edited texts.

Our early activities have focused on modules (3), (4), and (5). These modules will be described in the following three sections.

4. THE MULTI-VARIANT CONTENTS EDITOR (MVED)

Scholars need to prepare various editions based on their analysis of the raw material, interpreted through their knowledge and perceptions. To support the editing of documents with multiple versions, there are requirements that can not be met by current text editing tools. These requirements include the capability of locating, analyzing and classifying variances among different versions; and comparing and contrasting versions that could be in different formats (either facsimile image or plain text).

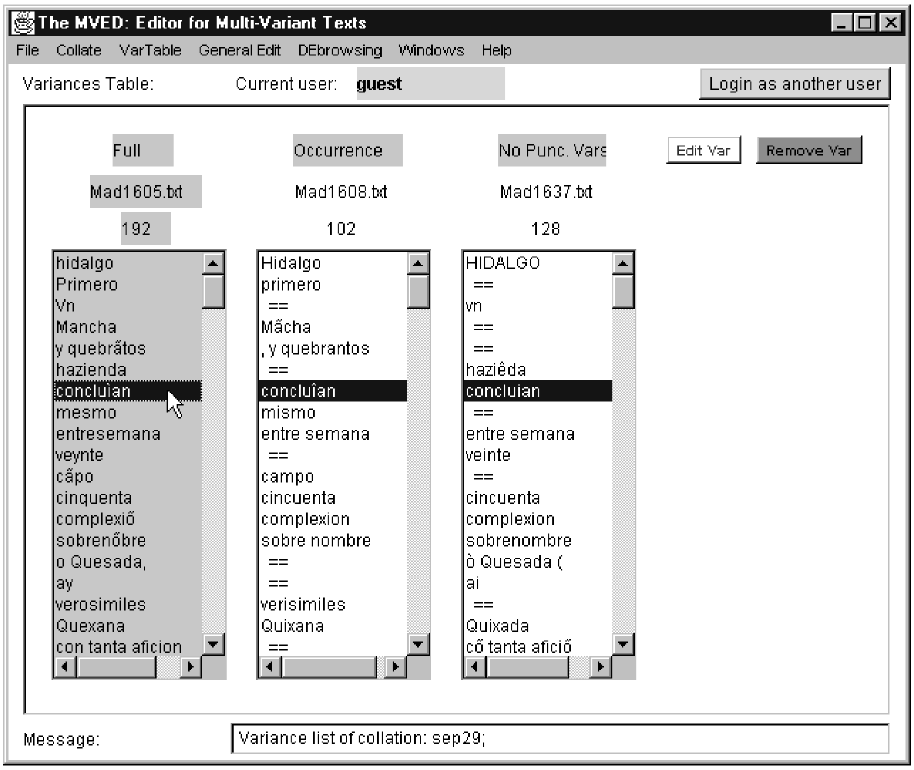
4.1. *Structural Overview*

To fulfill these requirements, the major components in the MVED include: (1) a collator, for comparing multiple versions of the same text to locate all textual variances among versions. The outcome of a collation session includes the offset and length of all variances found during the session. The offset and length data will be used by the text synchronization mechanism to facilitate the analysis of textual variances; (2) a mechanism for text and image synchronization, which facilitates comparing and contrasting multiple texts and their raw images. Since repeated comparison and contrast among the multiple sources are required to conduct the analysis of variances existing in those texts and editing decisions, an efficient mechanism for performing intensive comparison and contrast work among multiple texts/images sources is important; (3) a variance classification mechanism, which allows scholars to classify variances interactively, based on their knowledge and analysis of variances found by the collator.

4.2. *Interactions with the MVED*

The MVED is implemented in Java and is deployed on Windows NT. The MVED can find all textual variances in a collection of up to 32 texts with multi-variant contents by comparing each to a pre-selected base text. After collation, editors can select a specific set of variant contents in all collated texts with a mouse click. When such a selection is made, both plain text and facsimile representation are synchronized on the display. With support of the text/image synchronization mechanism, editors are able to focus on comparing and contrasting contents of documents with minimum cognitive overhead.

Figure 5 shows the MVED's main window, displaying a list of variances found in a collation session in which three different texts are analyzed. The three texts correspond to the chapter one in three versions of *Don Quixote*, denoted as *Madrid1605*, *Madrid1608*, and *Madrid1637*, respectively. Figure 6 shows an alternative, compact list of the same variance list. In either the full or compact variance list, clicking any item among a set of variances will enable text synchronization; i.e., the items of the same set of variance in other texts will be located at the same time. As seen in Figure 7, a document viewer shows a segment of text, with the variance highlighted. If desired, a facsimile image, also synchronized, can be displayed. Figure 8 shows three



Variances Table:

Current user: guest

Mad1605.txt	Mad1608.txt	Mad1637.txt
192	102	128
hidalgo	Hidalgo	HIDALGO
Primero	primero	==
Vn	==	vn
Mancha	Mãcha	==
y quebrãtos	, y quebrantos	==
hazienda	==	haziêda
concluìan	concluîan	concluian
mesmo	mismo	==
entresemana	entre semana	entre semana
veynte	==	veinte
cãpo	campo	==
cinquenta	cincuenta	cincuenta
complexiõ	complexion	complexion
sobrenõbre	sobre nombre	sobrenombre
o Quesada,	==	ò Quesada (
ay	==	ai
verosimiles	verisimiles	==
Quexana	Quixana	Quixada
con tanta aficion	==	cõ tanta aficiõ

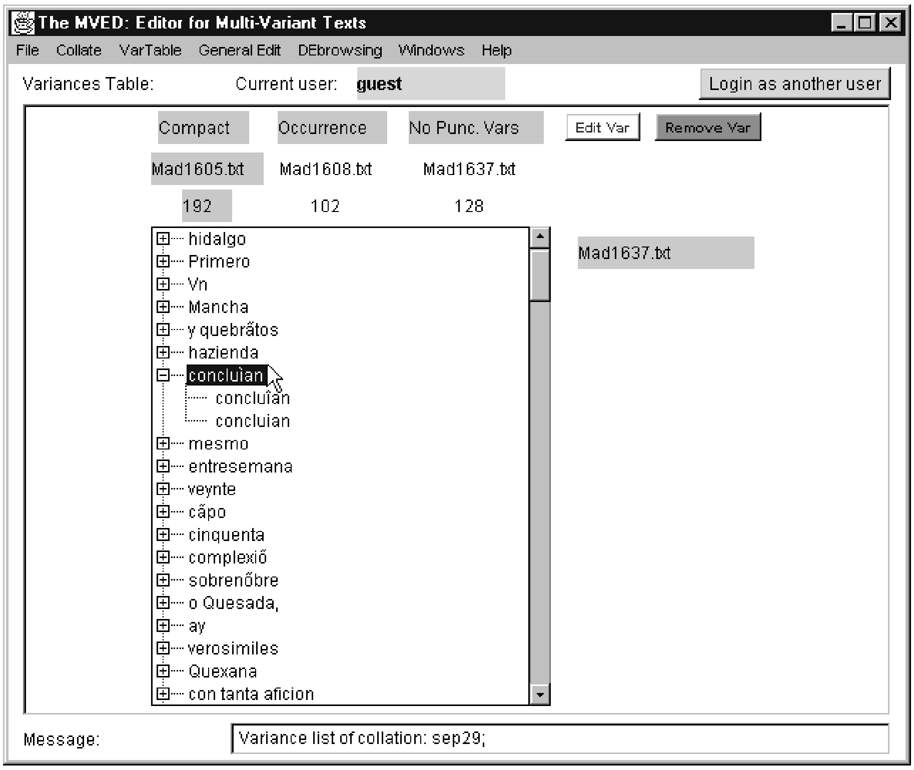
Message: Variance list of collation: sep29;

Figure 5. *Full variance list in the MVED*

synchronized viewers on a particular variance – *concluìan*, *concluîan*, and *concluian* in this example. An editor also may select any region of text in any of the viewers and synchronize the other viewers with it.

Since facsimile images are the sources of all data entities generated during the digitizing process, access to those images is important during editing. As Figure 7 shows, there is a scrollable image display area within each document viewer window. The synchronization mechanism between a text and its facsimile image counterpart allows users to locate corresponding regions. In addition to this text-to-image synchronization, clicking on facsimile images and the choice list located in the top row of each document viewer window allow users to perform image-to-text synchronization.

Whenever an editor decides to make a correction and/or commentary on a particular set of variances, pressing the *Edit Var* button will lead him/her to an editing dialog, as Figure 9 shows. Through the editing dialog, the editor is able to select the correct contents or make a correction directly, classify the edited variance into one of the five



Variances Table:

Current user: guest

Mad1605.txt	Mad1608.txt	Mad1637.txt
192	102	128

Mad1637.txt

- hidalgo
- Primero
- Vn
- Mancha
- y quebrãtos
- hazienda
- concluian
  - concluîan
  - concluian
- mesmo
- entresemana
- veynte
- cãpo
- cinquenta
- complexiõ
- sobrenõbre
- o Quesada,
- ay
- verosimiles
- Quexana
- con tanta aficion

Message: Variance list of collation: sep29;

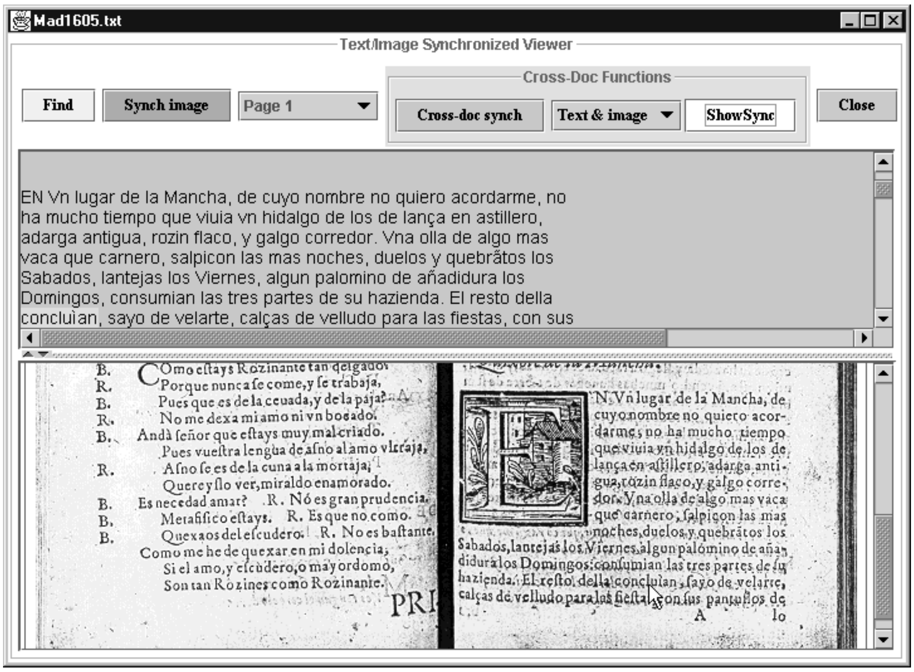
Figure 6. *Compact variance list in the MVED*

pre-determined categories[3], give commentary about the editing transaction, and cite supporting material. Pressing the *Save* button in the editing dialog will activate the HDEMS to store the contents of the editing transaction into the underlying database.

4.3. *Implementation Issues*

Due to the limits of available OCR techniques, it is not feasible to perform real-time synchronization between texts and their facsimile images counterparts. The feasible solution is to collect the coordinates of every line in facsimile images with a supporting tool. These pre-collected coordinates are used by the text-image synchronization mechanism to locate the region of an image approximately, which derives the corresponding paragraph of plain text.

[3] The five categories: printing errors, typographical errors, spelling variants, substantial-certain, and substantial-uncertain, were developed by a domain expert after examination of the texts and existing scholarly practices.



EN Vn lugar de la Mancha, de cuyo nombre no quiero acordarme, no
ha mucho tiempo que viuia vn hidalgo de los de lança en astillero,
adarga antigua, rozin flaco, y galgo corredor. Vna olla de algo mas
vaca que carnero, salpicon las mas noches, duelos y quebrãtos los
Sabados, lantejas los Viernes, algun palomino de añadidura los
Domingos, consumian las tres partes de su hazienda. El resto della
concluìan, sayo de velarte, calças de velludo para las fiestas, con sus

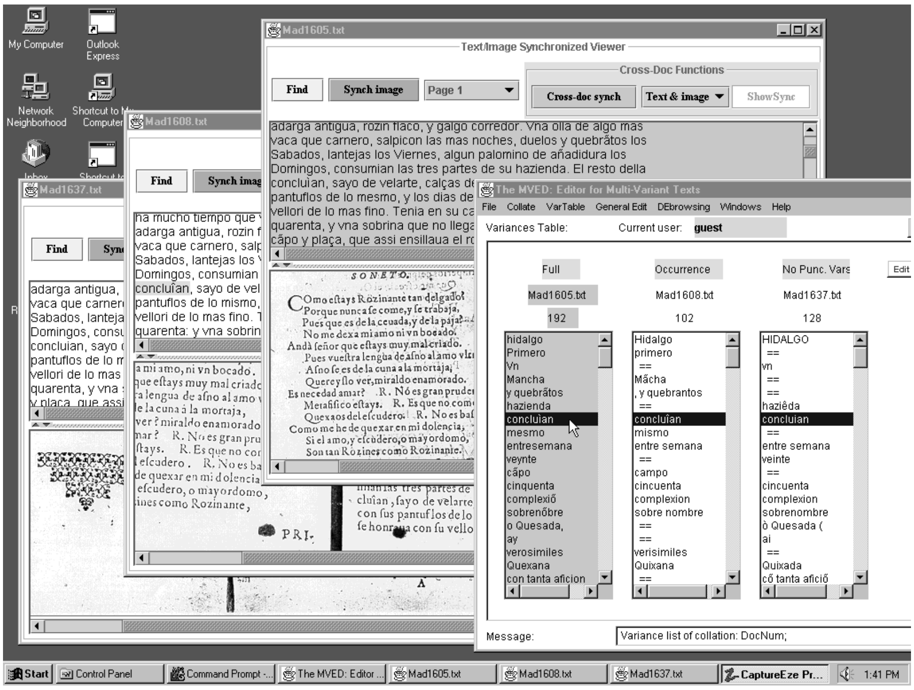
Figure 7. *Highlighted variance in the version of Madrid 1605*

5. THE HYPERMEDIA-BASED DATA ENTITY MANAGEMENT SYSTEM

Besides enforcing consistency control, two other objectives of the HDEMS are (1) providing common functionality, such as navigating among related data entities, for different upper-layered applications through uniform interface, which increase the modularity; and (2) facilitating efficient project management, such as being able to keep track of relationships among data entities.

5.1. *Architecture of the HDEMS*

Figure 10 shows the structure of the HDEMS and the relationship between the HDEMS and other applications. As Figure 10 shows, data entities can be stored in two places: either in the underlying database or in the file system. However, for the sake of efficient access control and verification, both the metadata and relationships of data entities, which are accessed frequently, are stored in the underlying database. All upper-layered applications access data entities through the application interface (API) of the HDEMS.



adarga antigua, rozin flaco, y galgo corredor. Vna olla de algo mas
vaca que carnero, salpicon las mas noches, duelos y quebrãtos los
Sabados, lantejas los Viernes, algun palomino de añadidura los
Domingos, consumian las tres partes de su hazienda. El resto della

Variances Table: Current user: guest

Full	Occurrence	No Punc. Vars
Mad1605.txt	Mad1608.txt	Mad1637.txt
192	102	128
hidalgo	Hidalgo	HIDALGO
Primero	primero	==
Vn	==	vn
Mancha	Mãcha	==
y quebrãtos	, y quebrantos	==
hazienda	==	haziêda
concluìan	concluîan	concluian
mesmo	mismo	==
entresemana	entre semana	entre semana
veynte	==	veinte
cãpo	campo	==
cinquenta	cincuenta	cincuenta
complexiõ	complexion	complexion
sobrenõbre	sobre nombre	sobrenombre
o Quesada,	==	ò Quesada (
ay	==	ai
verosimiles	verisimiles	==
Quexana	Quixana	Quixada
con tanta aficion	==	cõ tanta aficiõ

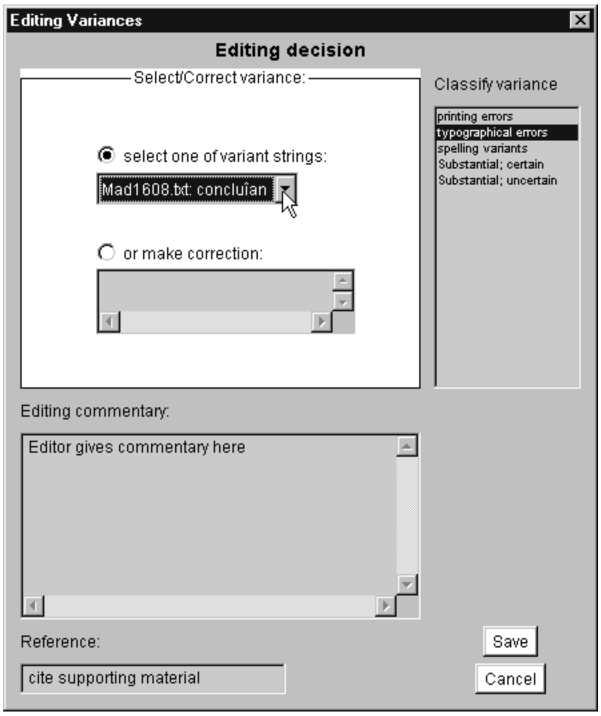
Message: Variance list of collation: DocNum;

Figure 8. *Three Synchronized Documents with Texts and Images*

5.2. *Underlying Database Design*

We chose to place data entities, relationships among them, and their metadata into a relational database for the following reasons. First of all, there are large number of instances of each type of entity and there exists a regular pattern of relationships among specific types of entities. Second, with features (metadata) of entities stored in a structured format, it is easier to locate particular entities precisely, by using metadata to specify the entities of interest. Third, the relationships among entities can be modeled by the entity-relationship (E-R) model [CHE 76, OZK 90, OZK 86], which is the underlying model of relational databases.

The first step of designing the underlying database is identifying all necessary entities, along with their attributes. There are two types of data entities generated during the process of creating electronic variorum editions: single-edition entity and cross-edition entity. Single-edition entities are those entities which contents come from one edition only. In our system, identified single-edition data entities include (1) raw image: the class for the images obtained from digitizing microfilms that record original texts, (2) enhanced image: the class for the enhanced images that will result in more efficient automatic contents recognition, (3) full page in ASCII: the class for the



● select one of variant strings: Mad1608.txt: concluîan

○ or make correction:

Classify variance: printing errors; **typographical errors**; spelling variants; Substantial; certain; Substantial; uncertain

Editing commentary: Editor gives commentary here

Reference: cite supporting material

Figure 9. *Variance Editing Dialog Interface*

plain texts obtained based on the facsimile images, (4) text paragraph within a page: a paragraph of text within one page, used for recording a variant string belonging to a particular variance, and (5) text segment: comprises continuous full pages in ASCII, that will be the basis for text collation to find textual variances among different editions of the same text. In contrast to single-edition entity, cross-edition entities' contents come from multiple editions. The identified cross-edition entities are (1) collation session: specifies one set of text segments, which textual collation was performed to find the variances among the given set of text segments, (2) variance: includes all information about a particular variance found during the process of a collation session, and (3) editing transaction: records contribution from the editors.

After identifying the entities and their attributes, the entity-relationship (E-R) model is used to model the relationships among the data entities. The derivational relationships among the data entities are modeled in two ways, depending on the type of the relationship. The one-to-many relationship is is represented by means of primary key (PK) - foreign key (FK) pairing. The requirements of the system determine which types of relationships are defined among entities. For example, the one-to-many relationship between the raw image class and the enhanced image class reflects that

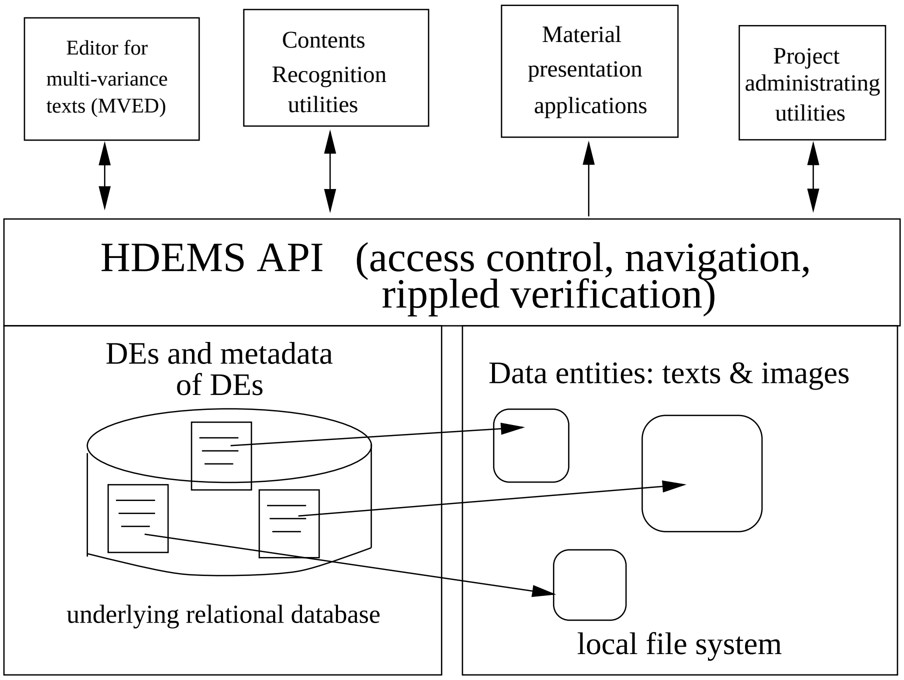
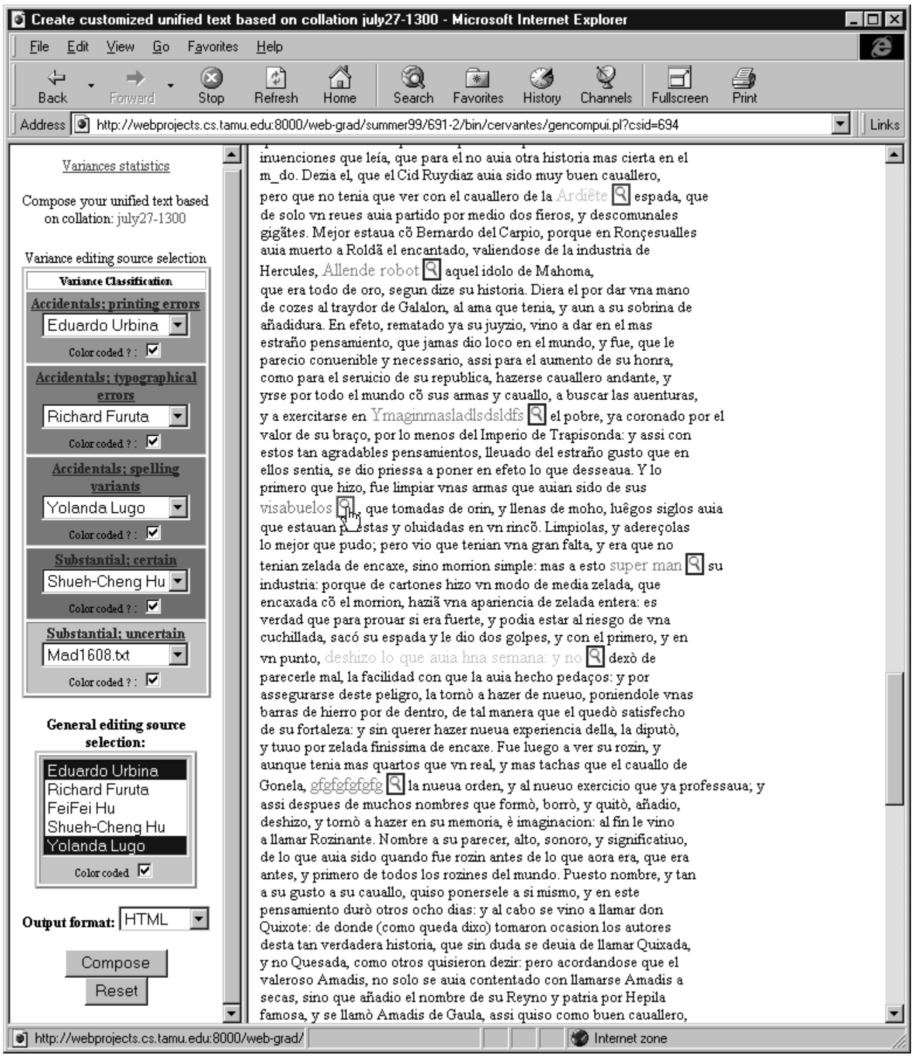

Editor for multi-variance texts (MVED)
Contents Recognition utilities
Material presentation applications
Project administrating utilities
HDEMS API (access control, navigation, rippled verification)
DEs and metadata of DEs
underlying relational database
Data entities: texts & images
local file system

Figure 10. *The architecture of the HDEMS*

more than one images could be produced by enhancing the same one raw image. The many-to-many relationship, also known as association, need to be represented as tables in E-R models. These associative tables are used by the HDEMS API to perform navigation; upper-layered applications do not know the existence of these tables.

5.3. *Implementation*

The relational database management system used in this work is Oracle 8. Tables in the database store various data entities, their metadata, and relationships. To access data entities in the database, the MVED uses JDBC — a standard relational database interface for Java programs. Another module needing access to the data entities is a unified document composer which will be discussed later, to access the data entities by using “oraperl” — a package for connecting Perl programs and the Oracle database.


Create customized unified text based on collation july27-1300 - Microsoft Internet Explorer
File Edit View Go Favorites Help
Back
Forward
Stop
Refresh
Home
Search
Favorites
History
Channels
Fullscreen
Print
Address
Links
Internet zone

http://webprojects.cs.tamu.edu:8000/web-grad/summer99/691-2/bin/cervantes/gencompui.pl?csid=694

Variances statistics

Compose your unified text based on collation: july27-1300

Variance editing source selection

Variance Classification

Accidentals; printing errors: Eduardo Urbina
Color coded ? : ☑

Accidentals; typographical errors: Richard Furuta
Color coded ? : ☑

Accidentals; spelling variants: Yolanda Lugo
Color coded ? : ☑

Substantial; certain: Shueh-Cheng Hu
Color coded ? : ☑

Substantial; uncertain: Mad1608.txt
Color coded ? : ☑

General editing source selection:
[✓ Eduardo Urbina]
Richard Furuta
FeiFei Hu
Shueh-Cheng Hu
[✓ Yolanda Lugo]
Color coded ☑

Output format: HTML

Compose
Reset

inuenciones que leía, que para el no auia otra historia mas cierta en el
m_do. Dezia el, que el Cid Ruydiaz auia sido muy buen cauallero,
pero que no tenia que ver con el cauallero de la Ardiête espada, que
de solo vn reues auia partido por medio dos fieros, y descomunales
gigãtes. Mejor estaua cõ Bernardo del Carpio, porque en Ronçesualles
auia muerto a Roldã el encantado, valiendose de la industria de
Hercules, Allende robot aquel idolo de Mahoma,
que era todo de oro, segun dize su historia. Diera el por dar vna mano
de cozes al traydor de Galalon, al ama que tenia, y aun a su sobrina de
añadidura. En efeto, rematado ya su juyzio, vino a dar en el mas
estraño pensamiento, que jamas dio loco en el mundo, y fue, que le
parecio conuenible y necessario, assi para el aumento de su honra,
como para el seruicio de su republica, hazerse cauallero andante, y
yrse por todo el mundo cõ sus armas y cauallo, a buscar las auenturas,
y a exercitarse en Ymaginmasladlsdsldfs el pobre, ya coronado por el
valor de su braço, por lo menos del Imperio de Trapisonda: y assi con
estos tan agradables pensamientos, lleuado del estraño gusto que en
ellos sentia, se dio priessa a poner en efeto lo que desseaua. Y lo
primero que hizo, fue limpiar vnas armas que auian sido de sus
visabuelos, que tomadas de orin, y llenas de moho, luêgos siglos auia
que estauan puestas y oluidadas en vn rincõ. Limpiolas, y adereçolas
lo mejor que pudo; pero vio que tenian vna gran falta, y era que no
tenian zelada de encaxe, sino morrion simple: mas a esto super man su
industria: porque de cartones hizo vn modo de media zelada, que
encaxada cõ el morrion, haziã vna apariencia de zelada entera: es
verdad que para prouar si era fuerte, y podia estar al riesgo de vna
cuchillada, sacó su espada y le dio dos golpes, y con el primero, y en
vn punto, deshizo lo que auia hna semana: y no dexò de
parecerle mal, la facilidad con que la auia hecho pedaços: y por
assegurarse deste peligro, la tornò a hazer de nueuo, poniendole vnas
barras de hierro por de dentro, de tal manera que el quedò satisfecho
de su fortaleza: y sin querer hazer nueua experiencia della, la diputò,
y tuuo por zelada finissima de encaxe. Fue luego a ver su rozin, y
aunque tenia mas quartos que vn real, y mas tachas que el cauallo de
Gonela, gfgfgfgfgfg la nueua orden, y al nueuo exercicio que ya professaua; y
assi despues de muchos nombres que formò, borrò, y quitò, añadio,
deshizo, y tornò a hazer en su memoria, è imaginacion: al fin le vino
a llamar Rozinante. Nombre a su parecer, alto, sonoro, y significatiuo,
de lo que auia sido quando fue rozin antes de lo que aora era, que era
antes, y primero de todos los rozines del mundo. Puesto nombre, y tan
a su gusto a su cauallo, quiso ponersele a si mismo, y en este
pensamiento durò otros ocho dias: y al cabo se vino a llamar don
Quixote: de donde (como queda dixo) tomaron ocasion los autores
desta tan verdadera historia, que sin duda se deuia de llamar Quixada,
y no Quesada, como otros quisieron dezir: pero acordandose que el
valeroso Amadis, no solo se auia contentado con llamarse Amadis a
secas, sino que añadio el nombre de su Reyno y patria por Hepila
famosa, y se llamò Amadis de Gaula, assi quiso como buen cauallero,

http://webprojects.cs.tamu.edu:8000/web-grad/

Figure 11. *Options and unified hypertext document in a browser*

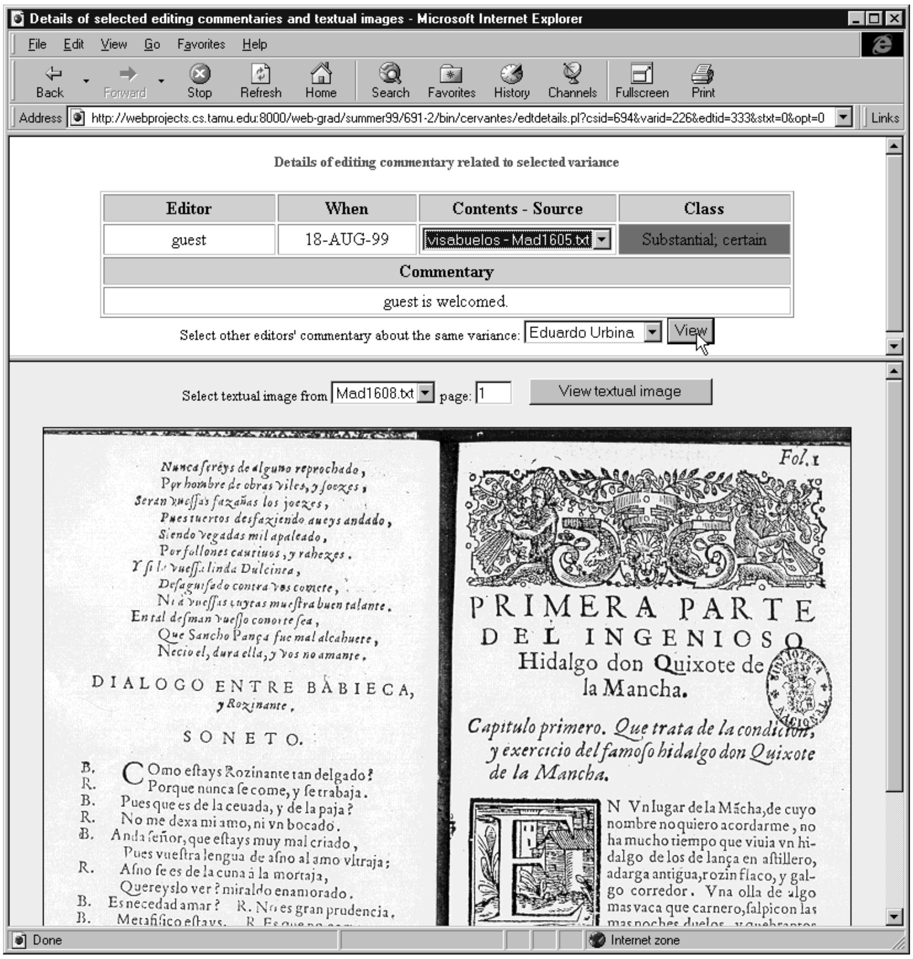
6. The Reader's Interface

Even with raw facsimile images, texts, and editing transactions stored in the system, the whole system will not be complete without an interface through which readers are able to access the raw material and editing records. Two goals of the reader's interface in our system are (1) to provide higher availability (allow more readers to access the contents of the system) and (2) to allow readers to customize unified documents, i.e., to generate any combinations of raw material and editing records, based upon their preferences. To achieve the goals, a Web-based reader's interface is developed, which uses standard Web browsers to accept a reader's input and to display unified documents, generated by a CGI program based on reader's input. As the left-side frame in Figure 11 shows, through the composition options, readers can select the sources of material. The unified document, shown in the right-side frame in Figure 11, will be created based on the selected options.

The CGI program is responsible for composing a hypertext document dynamically. Upon receiving a request from the reader, the hypertext document composer performs a sequence of operations to generate a customized unified document. These operations include (1) parsing the given arguments from Web browser to determine the reader's requirement. (2) Matching the reader's requirements with available material by sending queries to the underlying data-entity database. (3) Fetching the best-matched material from the data entity database. If the system can not find the requested data entities, the best alternative data entities will be sent for display. For example, if a reader want to see the variances categorized as printing errors by Editor-A but there is no editing transaction classified as "printing error" by Editor-A, the system will search for a variance identified as resulting from printing errors by any other editor for display. The system will use the corresponding variance in the base text as default replacement if no appropriate editing classifications exist in the database. (4) Associating hyperlinks with inter-relationships among data entities, for example, associating hyperlinks with editing records that contain detailed information about raw material, editing commentaries, and participating editors. Thus, in addition to browsing unified documents, readers also can investigate the edited material in detail through these hyperlinks. As Figure 11 shows, in the unified document, each variance is color-coded, enlarged, and associated with a hyperlink represented by a magnifying glass icon. Clicking an icon will lead reader to another browser window where all data entities related to the associated variance are displayed or can be reached through other hyperlinks; see Figure 12. If more than one editing transactions has been made on the same variance by different editors, reader can select any one of the editing transactions by using the pull-down menu in the upper-side frame. Besides editing transactions, readers also can view facsimile images, which contain the variances of interest in the lower-side frame. (5) Composing a unified documents by concatenating fetched data entities in the order of occurrence. The unified hypertext document will be sent back to reader's browser through the Web server.

7. CONCLUSIONS

Due to rareness, access to ancient documents usually is restricted, which hinders the research work of domain experts and awareness of rare documents by general readers. Besides presenting the overview of an integrated system for digitizing ancient documents with degraded visual quality and multi-variant contents, this paper details three modules that are under development. First, this paper describes how an



http://webprojects.cs.tamu.edu:8000/web-grad/summer99/691-2/bin/cervantes/edtdetails.pl?csid=694&varid=226&edtid=333&stxt=0&opt=0

Details of editing commentary related to selected variance

Editor	When	Contents - Source	Class
guest	18-AUG-99	visabuelos - Mad1605.txt	Substantial; certain

Commentary

guest is welcomed.

Select other editors' commentary about the same variance: Eduardo Urbina

Select textual image from Mad1608.txt page: 1

Nunca ſereys de alguno reprochado,
Por hombre de obras viles, y ſoezes,
Seran vueſſas fazañas los joezes,
Pues tuertos desfaziendo aueys andado,
Siendo vegadas mil apaleado,
Por ſollones cautiuos, y rahezes.
Y ſi la vueſſa linda Dulcinea,
Deſaguiſado contra vos comete,
Ni à vueſſas cuytas mueſtra buen talante.
En tal deſman vueſſo conorte ſea,
Que Sancho Pança fue mal alcahuete,
Necio el, dura ella, y vos no amante.

DIALOGO ENTRE BÀBIECA, y Rozinante.

SONETO.

B. Como eſtays Rozinante tan delgado?
R. Porque nunca ſe come, y ſe trabaja.
B. Pues que es de la ceuada, y de la paja?
R. No me dexa mi amo, ni vn bocado.
B. Anda ſeñor, que eſtays muy mal criado,
Pues vueſtra lengua de aſno al amo vltraja;
R. Aſno ſe es de la cuna à la mortaja,
Quereyslo ver? miraldo enamorado.
B. Es necedad amar? R. No es gran prudencia.
B. Metafiſico eſtays. R. Es que no como.

Fol. 1

PRIMERA PARTE DEL INGENIOSO Hidalgo don Quixote de la Mancha.

BIBLIOTECA NACIONAL

Capitulo primero. Que trata de la condicion, y exercicio del famoſo hidalgo don Quixote de la Mancha.

EN Vn lugar de la Mãcha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que viuia vn hidalgo de los de lança en aſtillero, adarga antigua, rozin flaco, y galgo corredor. Vna olla de algo mas vaca que carnero, ſalpicon las mas noches, duelos, y quebrantos

Figure 12. *Browser window displaying details of a variance*

editing tool for documents with multi-variant contents (MVED) can locate variances, facilitate comparison and contrast of multiple versions of the same document, and allow scholars to correct and comment on particular parts of the document. With it's embedded synchronization mechanism, the MVED can locate a set of variant contents in every collated texts as well as the corresponding facsimile image based on a particular segment of texts. Following the MVED, this paper describes how derivational relationships among related data entities can be stored in a relational database, thus less inconsistencies will result from the frequent update on the linkage structure of related entities and the contents of entities. The third module presented in detail is a Web-based readers' interface. The pervasiveness of Web browsers increases the availability of both raw and edited material. Through the Web-based interface, readers can customize the composition of unified texts, as well as traverse among all related data entities through hyperlinks.

Integrating the above modules together, a generic solution is formed for creating electronic editions of ancient textual documents. The integrated system benefits scholars by associating raw materials with their editing rationale closely and precisely. The association also allows dynamic links among raw materials and support for multiple scholarly interpretations. This facilitates serving readers with different backgrounds, interests, and goals. In addition, readers are not limited by the decisions made by editors, readers can create customized editions at will. More broadly, we believe that the concepts, working procedures, and software techniques will be helpful in handling issues of mid-to-large scale processing of ancient documents for a wide variety of additional collections.

8. References

[BLA 95] BLAKE N. F., ROBINSON P., SOLOPOVA E., *The Canterbury Tales Project*, , 1995, http://www.shef.ac.uk/uni/projects/ctp/index.html.

[BRE 96] BREWER A., DING W., HAHN K., KOMLODI A., "The Role of Intermediary Services in Emerging Digital Libraries", *DL96: Proceedings of Digital Libraries*, Bethesda MD, USA, 1996, ACM, p. 29-35.

[CAS 90] CASEY R. G., YONG S. Y., "*Image Analysis Applications*", chapter 1, p. 1-36, Marcel Decker, Inc., 1990.

[CHE 76] CHEN P., "The Entity Relationship Model: A Basis for the Enterprise View of Data", *ACM Trans. on Database Systems*, vol. 1, num. 1, 1976, p. 9-36.

[GAR 98] GARRIS M. D., JANET S., KLEIN W. W., "Impact of Image Quality on Machine Print Optical Character Recognition", report , January 1998, National Institute of Standards and Technology.

[GLA 98] GLADNEY H. M., MINTZER F., SCHIATTARELLA F., BESCOS J., TREU M., "Digital Access to Antiquities", *Communications of ACM*, vol. 41, num. 4, 1998, p. 49-57.

[LEC 98] LECOLINET E., ROLE F., LIKFORMAN-SULEM L., LEBRAVE J.-L., ROBERT L., "An Integrated Reading and Editing Environment for Scholarly Research on Literary Works and their Handwritten Sources", *Proceedings of the Third ACM international conference on Digital Libraries, DL '98*, Pittsburgh, PA, June 1998, ACM, p. 144-151.

[OZK 86] OZKARAHAN E., *Database Machines and Database Management*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, USA, 1986.

[OZK 90] OZKARAHAN E., *Database Management: Concepts, Design, and Practice*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, USA, 1990.

[SEA 97] SEAMAN D., "The User Community as Responsibility and Resource: Building a Sustainable Digital Library", *D-Lib Magazine*, , 1997, http://www.dlib.org/dlib/july97/07seaman.html.