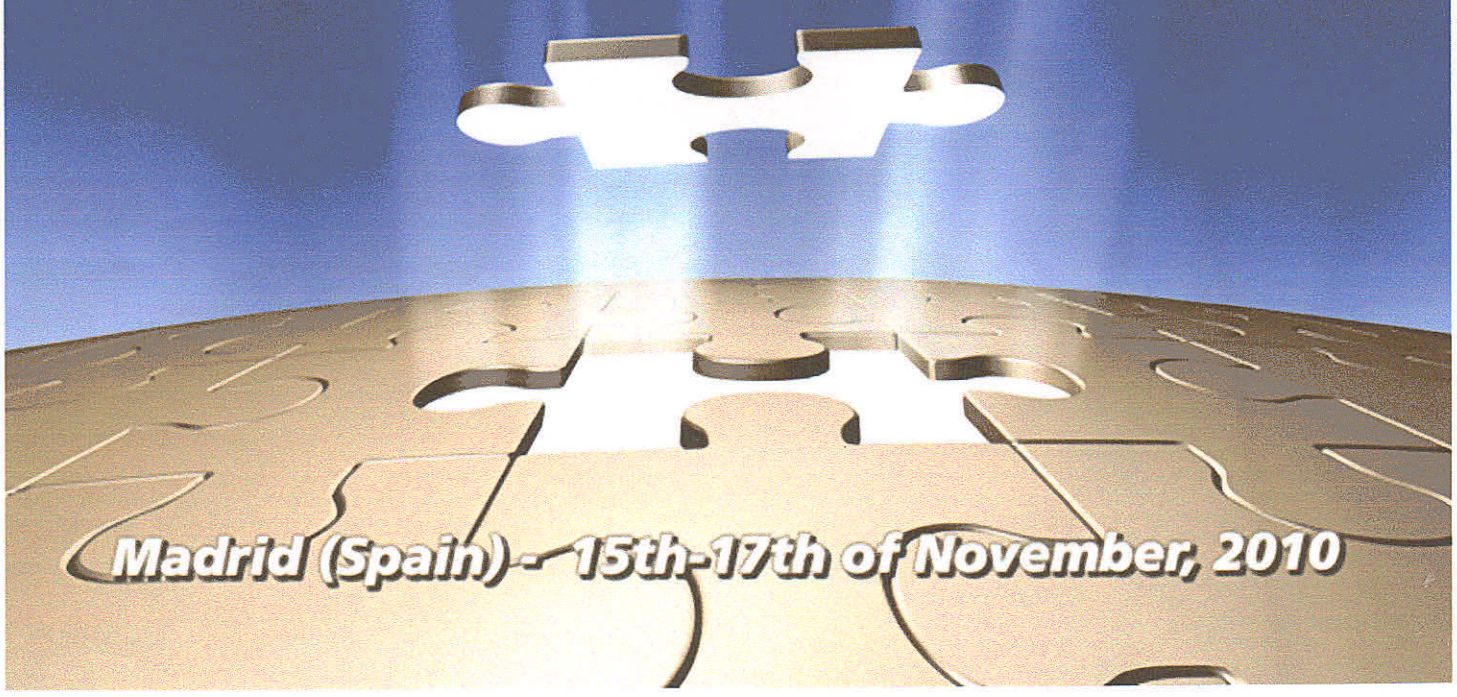
iCERi 2010

International Conference of Education, Research and Innovation

CONFERENCE PROCEEDINGS



Madrid (Spain) - 15th-17th of November, 2010

TEI ENCODING IN CONJUNCTION WITH MYSQL, JAVASCRIPT, AND HTML TO CREATE A VISUAL VARIORUM EDITION OF DON QUIXOTE

[1]Víctor E. Agosto, [1]Eduardo Urbina, [2]Fernando González Moreno, [3]Richard Furuta

[1]Department of Hispanic Studies, Texas A&M University (UNITED STATES)
[2]Department of Art History, Universidad de Castilla-La Mancha (SPAIN)
[3]Center for the Study of Digital Libraries; Department of Computer Science and Engineering, Texas A&M University (UNITED STATES)

Abstract

In this paper we present the steps taken to TEI encode both parts of Don Quixote and the approach and solutions developed to fit the requirements of our archive regarding visual elements and dynamic links utilizing the structural divisions and taxonomic categories to be able to navigate from images to texts and from texts to images. We describe the four level division utilized to create the TEI One Document Does is All (ODD), including the Part and Chapter headers linked to all the images in the database for a particular chapter. In addition, we illustrate the Text/Image that corresponds with the terms found in the Browse image archive by content finding aid, which makes possible to filter results by individual chapters. This multilevel approach facilitates the search of information as well as the maintenance and update of the TEI tags.

1 INTRODUCTION

Since 2003, the Cervantes Project (http://cervantes.tamu.edu) has been developing a fully documented hypertextual archive to make accessible the textual iconography of the Quixote (Madrid 1605, 1615). With support from a Preservation and Access grant from the National Endowment for the Humanities (2006-2009) we have digitized, indexed, and annotated over 30,000 illustrations and have developed search tools and finding aids to enable the identification of discreet images or sets of items based on a newly created taxonomy of episodes and adventures of the text, which includes over 500 categories, as well as a 400 control vocabulary of key words. The contents of the image database and metadata are dynamically updated and the results are available immediately online through the collection index and a multilayered search engine. In 2009 we applied tags related to both the taxonomy categories and the key words to create two online editions of the Quixote, one in Spanish and one in English, linking the illustrations to the texts and thus producing a virtual visual variorum edition suitable for both research and teaching purposes.

2 TEI/XML/XLST ENCODING PROCESS

We utilized the Oxygen 9 software packet as the bases to encode both parts of Don Quixote (1605 and 1615) for the Text Encoding Initiative (TEI) process. We also have used the Element Classes of the TEI XML P4 version guidelines [5] to supplement the Oxygen's Editor, when the Oxygen instructions lacked the necessary elements in the encoding process of the project. Both, the Oxygen program and the Element Classes were used in conjunction with the P5: Guidelines for Electronic Text Encoding and Interchange website [4]. Utilizing the editor and author screens in Oxygen, four TEI encoding levels were created for the project.

Level 1 divides the header from the Text sections. The head section displays all the concerned documentation pertaining to the writing and publishing of Don Quixote and also the creation information of the project. This section is dynamical in that it is always filled with information that references the rest of the work. The text section encompasses the rest of the levels encoding instructions. An example is shown below:

- <TEI xmlns="http://www.tei-c.org/ns/1.0">
- <teiHeader>
- <fileDesc>
- <titleStmt>

ISBN: 978-84-614-2439-9

```
- <title>
- <biblFull>
- <hi rend="italic">
<hi rend="bold">FIRST PART OF THE INGENIOUS Hidalgo don Quixote de La Mancha </hi>
                    </biblFull>
                    </title>
                     </titleStmt>
- <publicationStmt>

                    <p>MIGUEL DE CERVANTES SAAVEDRA</p>
                    <p>Don Quixote</p>
                    <p />
                    <p />
                     <p>Translated and with notes by</p>
                     <p>TOM LATHROP</p>
                     <p>Founder Member, The Cervantes Society of America</p>
                    <p>Asociación de Cervantistas</p>
                    <p />
                    <p>Consulting Editors</p>
                    <p>ANNETTE GRANT CASH</p>
                    <p>Georgia State University</p>
                    <p />
                    <p>VICTORIA RICHARDSON</p>
                    <p>Cervantes Prize Winner, University of Delaware</p>
                    <p />
                    <p>JAMES K. M. SADDLER</p>
                    </publicationStmt>
```

Level 2 provides a unique identifier to each chapter for both part of the Quixote. First, a <divGen > tag is implemented as the principal tag for each chapter. This division supplies the necessary flexibility at the time encoding is implemented within each chapter. In the parenthesis of this tag, the attributes of id, n, and type are used to give them a unique identification. The id attribute declares the year that the book was published (1605 for part I and 1615 for part II). The n attribute represents the following categories: (1) The number of the chapter, (2) Preliminaries, (3) table, (4) Poetry, and (5) Prologue. The type produces the class of the divisions in the n values. These differences are made because of the use of other division class types such as poetry, encountered in Don Quixote. Below are samples of three different <divGen> tags found in the text:

```
<divGen id="DQ1605" n="1" type="chapter"></divGen>
<divGen id="DQ1605" n="Prologue" type="poetry"></divGen>
<divGen id="DQ1605" n="14" type="song"></divGen>
```

Level 3 identifies the various unique elements found on both parts. For instance, the <lg></lg> tag is utilized to separate the different kind of poetry (i.e. Sonnet, free style, etc). To separate these poetries the type attribute is applied inside the <lg> tag with two main classifications reflected throughout both parts of Don Quixote: (1) Sonnet, and (2) Free Style. Examples of these classification encodings are found below:

```
- <divGen id="DQ1605" n="Prologue" type="poetry">
- <lg type="free">
                    To the book about don Quixote de La Mancha
                    <l>URGANDA THE UNKNOWN</l>

- <lg type="Sonnet">
                    AMADÍS DE GAULA
l>To don Quixote de La Mancha</l>
- </lg>
- </divGen>
```

Level 4 implements further element's classes within the level 3. For sample, in level 3, tags are used to identify regular paragraphs. Inside this paragraph a written word in another language may appear, which is tagged with the <foreign></foreign>. For instance, the <lang></lang> tag is used to

identify what is foreign in the word (i.e. "Spanish", "English", "Latin"). Below there are a couple examples of level 4:

```
<divGen id="DQ1605" n="1" type="chapter">
    <p>[...]
        <foreign lang="lat">tantum pellis ossa fuit</foreign>
    </p>
</div>
```

The use of these levels produced two different TEI encoding versions of the text. The first is an ODD (One Document Does it All), which through the Oxygen software, processed the XML/TEI and, parsing it through XSLT, converted this ODD into an XHTML format. The second product is comprised of dividing the ODD and its XHTML format into individual chapters. This was done to facilitate navigation within the chapters of both parts of Don Quixote. It also, provides users with the manageability of reading chapters in the form of a book, in that the readers have the option to jump chapters without the cumbersome need of scrolling up or down a 700 plus page ODD.

3 PROBLEMS AND SOLUTIONS

Three major problems were found encoding both parts of Don Quixote in the English and Spanish languages. The first problem encountered in this project was the selection of the TEI tags to satisfy the parameters of the various genres found in Don Quixote in both English and Spanish editions. Because Don Quixote includes a wide diversity of genres (i.e. poetry, short stories, pastoral novels, etc), their presence presented unique challenges at the time of selecting encoding tags since they are included with little or no textual marking. For example, inside the short novels found in the text, there are secondary stories which provide a more realistic environment to the book's overall fiction. When selecting the tags for these secondary stories, a careful approach in identifying each section was implemented to ensure users could differentiate between a sentimental episode (Luscinda and Cardenio) and a pastoral narration (Marcela and Grisóstomo). In this case, the <textDesc></textDesc> was selected, where <textDesc id=" "> reflects the chapter and <textDesc n=" "> the types of narratives found within that chapter. Although a series of TEI tags have been identified to start, monitor and continue with the project, this problem will not be completely resolved until all the texts has been fully encoded in both languages. Since the Quixote contains the major, if not all, of modern genres, the texts will need to be reviewed and re-tagged on an ongoing basis to satisfy various types of research and analysis.

A second problem found in encoding the illustrated Quixote was the avoidance of duplication within the same document in English or Spanish. This problem has been discussed by Alejandro Bia and Manuel Sánchez Quero and they conclude that there must be only one original document, disregarding its language, as a starting point [1]. This means that the one edition in English and one in Spanish were chosen to represent the many editions found of the texts from their first publishing dates of 1605 and 1615. As Arantza Casillas and Raquel Martínez explain, adopting an original document permits an automation of bilingual documentation and relieves the very time consuming and expensive processes otherwise created [2]. Also, for future editions in any other language, the English and Spanish TEI encoded editions will serve as an overall guide to encode those editions.

The last major problem came when trying to parse the encoded JavaScript images' links through the TEI/XML and XLST processes. In the Cervantes Project's Iconography, the images are shown on the web with the use of Hypertext Transfer Protocol (http) addresses, which are called from the MySQL database through the use of JavaScript's encoded procedures. In these procedures, the sign "=" is utilized to match images with a text search engine that present the user with all the images for a particular keyword. This equal sign is interpreted in the parsing in two ways: 1) as a delimiter, which, in accordance with Oxygen, must be a semicolon instead of the equal sign, or 2) as an Element of the type in which is searching of its attribute. In both cases, these incompatibilities prevent the linking of the TEI text with the JavaScript images. Even, when trying to run an XML document through normal parsing (utilizing another program beside Oxygen) the result is a semi colon character was expected. Error processing resource 'file:///F:/DQ TEI/DQ.xml'.Line 53, Position 102.
To solve this problem, the decision of not linking the text with the images was taken and instead the approach of using an XHTML encoding with the free to use TEI CSS (Cascading Style Sheet) was adapted separately to link text and image within a TEI environment.

4 XHTML AND TEI CSS NETWORK

When utilizing the free to use TEI CSS [3], an XMTML (Extensible Hypertext Markup Language) has been created to hyperlink the keywords found in the Browse image archive by content of the Iconography Section at the Cervantes Project with their images into both parts of Don Quixote. These keywords are divided in four divisions: 1) People, 2) Places, 3) Objects, and 4) Animals. Also, each keyword is hyperlinked, through the encoding of JavaScript, to its correspondence image of the MySQL database called Iconography.

The XHTML has been divided by chapters for both parts to generate a network of chapters that allows navigational interaction within the texts. In this case, there are 52 individuals XHTML pages for Part I and 74 for Part II. There are two reasons for having these divisions. First, each keyword may or may not have images present in each chapter. Second and more important, each keyword contains metadata in which a place for Chapter has been assigned to that particular image, thus, allowing the exclusive selection of images by chapter. This allows users the advantage of viewing only the image or images found in an individual chapter hyperlinked to the keywords, avoiding any other images belonging to any other chapter in the work.

For example, in Figure 1 the keyword lanza (lance) is shown to contain 17 images in the first chapter of Part I, while the same keyword shows 19 images in chapter 2 of the same part as seen in Figure 2.

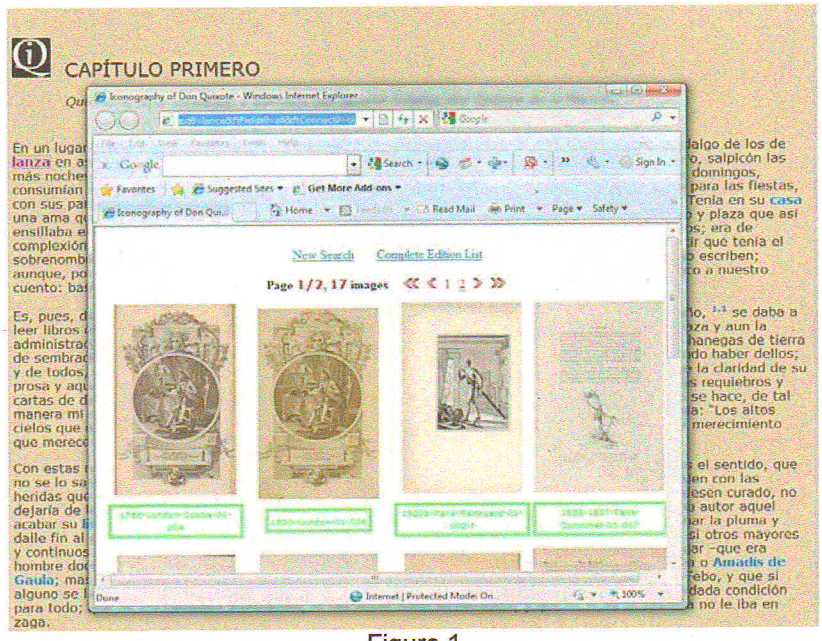


Figure 1

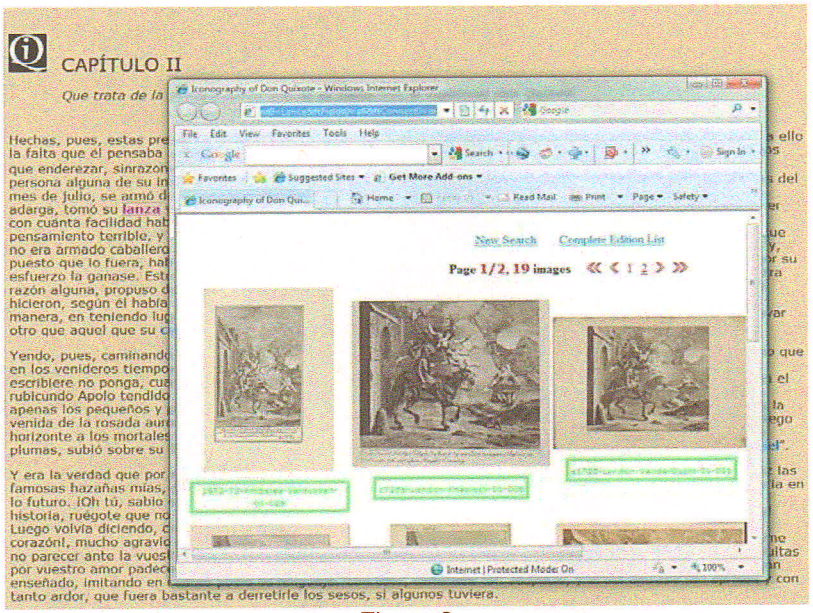


Figure 2

As seen in theses figures, when hyperlinked keywords are pressed, a separate window opens, showing corresponding images. This allows users the opportunity to analyze the images without exiting the text. Also, it offers the benefit of having multiple open windows at one time.

Through the use of the XHTML and the TEI CSS, a dynamically text/image is presented for the user to interact while at the same time research is conducted analyzing the content, author or year of the images in a TEI environment. Furthermore, combining XHTML and the TEI CSS, an illustrated and annotated edition of both parts of Don Quixote has been created to better represent and comprehend the narrative texts. Lastly, a secondary XHTML project is being created in which each keyword will link and visualize selected images by chapter and part. This approach offers the user the ability to explore and analyze the illustrations associated with each keyword without having to go through the chapters one by one.

5 CONCLUSION

We have discussed the ongoing procedures developed to utilize the Text Encoding Initiative in parallel editions of Don Quixote (Spanish and English) in association with a rich image archive of illustrations and, in particular, the processes and solutions developed to fit the requirements of our archive regarding visual elements and dynamic links by utilizing the structural divisions, taxonomic categories and keywords to be able to visualize and navigate from images to texts and from texts to images. We have described the four divisions utilized to create the TEI One Document Does is All (ODD) along with the creation of an XHTML/TEI CSS network that make possible to associate the terms found in the Browse image archive by content finding aid with their correspondent word in the texts, and thus produce a variorum illustrated edition of both parts of Don Quixote. Also, we have discussed the three major difficulties and the solutions encountered while working in this project.

Finally, in the course of our research a secondary project has emerged which involves the future generation of a TEI schema to facilitate the analysis of the terms found in the Browse image archive by individual chapter and part.

ACKNOWLEDGEMENTS

This material is based upon work supported by the National Endowment for the Humanities under Grant no. PA/51993-06 and has been conducted under the direction of Dr. Eduardo Urbina, Editor of the Textual Iconography Archive project and Director of the Cervantes Project at Texas A&M University, and Dr. Richard Furuta, Director of the Center for the Study of Digital Libraries in the Department of Computer Science at Texas A&M University. Further information about the Cervantes Project may be found at http://cervantes.tamu.edu/

REFERENCES

[1] Bia, Alejandro and Sánchez Quero, Manuel. "Desarrollo de Herramientas Informáticas para el Marcado Multibilingüe: Conversión del TEI al Español". La Biblioteca Virtual Miguel de Cervantes (BVMC). http://cervantesvirtual.com. July 21, 2010.

[2] Casillas, Arantza and Martínez, Raquel. "Bitext Generation through Rich Markup". *Computer and the Humanities* 38: 223-251, 2004.

[3] TEI CSS. Text Encoding Initiative. http://www.tei-c.org/release/xml/tei/stylesheet/tei.css July 3, 2010.

[4] P5: Guidelines for Electronic Text Encoding and Interchange, Text Encoding Initiative. http://www.tei-c.org/Guidelines/P5/ July 3, 2010.

[5] The XML Version of the TEI Guidelines. Text Encoding Initiative http://www.tei-c.org/Guidelines/P4/html/REFCLA.html

Published by
International Association of Technology, Education and Development (IATED)
www.iated.org

ICERI2010 Proceedings CD
Edited by
L. Gómez Chova, D. Martí Belenguer, I. Candel Torres
International Association of Technology, Education and Development
IATED, Valencia, Spain

ISBN: 978-84-614-2439-9
Depósito Legal: V-3998-2010

Book cover designed by
J.L. Bernat

SCIENTIFIC COMMITTEE AND ADVISORY BOARD

Agustín López	SPAIN	Julia Williams	UNITED STATES
Aldon Hartley	NEW ZEALAND	Karen Woodman	AUSTRALIA
Ali Simsek	TURKEY	Kari Kumpulainen	FINLAND
Amparo Girós	SPAIN	Karl Hain	GERMANY
Ana Paula Cláudio	PORTUGAL	Katie Goeman	BELGIUM
Antonio García	SPAIN	Linnea Stenliden	SWEDEN
António Gomes	PORTUGAL	Luis Gómez Chova	SPAIN
Astrid Ramirez Valencia	COLOMBIA	Lurdes Babo	PORTUGAL
Audrey Cooke	AUSTRALIA	Mª Jesús Suesta	SPAIN
Christoph Rappl	GERMANY	Manfred Meyer	GERMANY
Claude Doom	BELGIUM	Mari Lahti	FINLAND
David Edelman	UNITED STATES	Maria Porcel	SPAIN
David Martí	SPAIN	Maritta Välimäki	FINLAND
Diana Phillips	UNITED STATES	Matthew Temple	UNITED STATES
Don Burwell	UNITED STATES	Mónica Fernández	SPAIN
Dorina Ionescu	SOUTH AFRICA	Norma Barrachina	SPAIN
Elena Ors	SPAIN	Osama Shata	QATAR
Elena Shoikova	BULGARIA	Patrizia Lùperi	ITALY
Georg Öttl	AUSTRIA	Paul Morris	AUSTRALIA
Gerd-Michael Hellstern	GERMANY	Penelope Bidgood	UNITED KINGDOM
Gihane Endrawes	AUSTRALIA	Peter Haber	AUSTRIA
Hannelie Nel	SOUTH AFRICA	Richa Malhotra	INDIA
Huseyin Yolcu	TURKEY	Sergio Pérez	SPAIN
Ignacio Ballester	SPAIN	Serkan Perkmen	TURKEY
Ignacio Candel	SPAIN	Sue Cobb	UNITED KINGDOM
Insook Choi	UNITED STATES	Susana Raya	SPAIN
Irene Lee	HONG KONG	Tala Vaziri	UNITED ARAB EMIRATES
Ismael Serrano	SPAIN	Tamara Bianco	GERMANY
Javier Domenech	SPAIN	Timothy Hornberger	UNITED STATES
Javier Martí	SPAIN	Valentina Gullà	ITALY
Jeramie Silveira	UNITED STATES	Vladislav Denishev	BULGARIA
John B. Stav	NORWAY	Wei-wen Chang	TAIWAN
Jose F. Cabeza	SPAIN	Wing-sat Chan	CHINA
Jose Luis Bernat	SPAIN	Xavier Lefranc	FRANCE
Judith Aponte	UNITED STATES	Zainal Abidin Koemadji	INDONESIA