Texts, Images, Knowledge: Visualizing Cervantes and Picasso

Carlos Monroy, Richard Furuta
Center for the Study of Digital Libraries
Texas A&M University

Eduardo Urbina, Enrique Mallen
Department of Hispanic Studies
Texas A&M University

Abstract

In this paper we present two cases of the application of information technology to the humanities. The collections we are working with incorporate linked image-text collections, and tools to explore and analyze their contents. The first case is the Cervantes Project [1], which includes a digital collection of copies of the first edition of *Don Quixote* published in 1605, as well as texts of other Cervantes' works. The second case is the On-line Picasso Project [2], a web-based repository of images and history about Pablo Picasso's artistic creation. We also discuss the use of Interactive Timeline Viewer (ItLv), a visualization tool used to depict and explore the contents of both collections.

The Cervantes Project: An Overview

The Cervantes Project hosted at Texas A&M University, is a joint project between the Department of Hispanic Studies and the Center for the Study of Digital Libraries. The main goal of the Cervantes Project is to provide a comprehensive repository of information, texts, and images about Miguel de Cervantes life and works. To achieve this, the project is divided into three main components: a) The Cervantes Digital Library, an electronic repository of novels, plays, and other writings in different formats and versions, b) the Cervantes International Bibliography On-line, a cumulative annotated bibliography of studies, editions, and translations of Cervantes' works, and c) the Cervantes Digital Archive of Images, an archive of images documenting Cervantes' life and time.

Another important component of the Cervantes Project is the creation of an Electronic Variorum Edition of Cervantes' *Don Quixote*[3], to accomplish this, we have been acquiring multiple microfilmed copies of several of the early editions of *Don Quixote*. The novel, which is divided in two parts, was first published in Madrid in 1605—part one—and 1615 —part two—, this first edition is also known as *princeps*. Of the 1605 princeps edition, it is believed that only approximately 18 copies have survived, and of these only 12 are accessible to scholars. Our collection includes nine digital copies of the princeps edition for both parts.

Additionally, we have also obtained more than thirty copies of the nine key early editions published between 1605—second edition—and 1637, and considered of special textual relevance. The Electronic Variorum Edition enables scholars to collate, classify, justify, and annotate textual variants. It also has a browsing mechanism, so users can navigate through the texts and images. A synchronized tool allows users to visualize texts and their corresponding images simultaneously, along with links to the variants, critical commentary, and scholarly annotation.

A new component to be incorporated to the project is the creation of a Textual Iconography of *Don Quixote*[4]. This component will include illustrations, engravings, and drawings of about two hundred illustrated editions of *Don Quixote* published since 1620. These illustrations will be linked to the editions already in the collection, which in turn will enable users to read a hyperlinked multi-year illustrated edition of *Don Quixote*.

The Electronic Variorum Edition of *Don Quixote*

As mentioned earlier, one of the components of the Cervantes Project is the creation of an Electronic Variorum Edition; to accomplish this, we have acquired from different institutions, about forty copies of the early editions of Don Quixote, ranging from 1605 to 1637, including nine copies of the first edition published in 1605 in Madrid. After the acquisition of the copies in facsimile format, they have been transcribed into text files. The transcription was done manually because of the low accuracy results obtained during initial optical character recognition (OCR) attempts. Using a software tool called Multi Variant Editor for Documents (MVED), these texts are then electronically collated. To collate a set of texts using the MVED, the user chooses one of the texts as the base text, which in turn is compared against all of the other texts –called comparison texts--, automatically identifying all the variants or differences among them.

Scholars can then classify these variants as printing errors, typographical errors, spelling variants, substantive certain, or uncertain. Along with the variant's classification, scholars can justify their editing decision, emendate the text, or annotate a section of the text, which can also be attached to any of the texts used in the collation. All these information is then stored in a database. Finally, these texts and their corresponding images along with the critical scholarly commentary are available on the Internet through a reader's interface, which can be accessed by users with different backgrounds and goals, from a school student to a Cervantes scholar.

Let us imagine for a moment building a Variorum Edition without the help of information technology and computer science advances. Foremost, access to several copies would be impossible, unless more than one copy were available at the same place. The collation process would be error prone due to its repetitive and time-consuming nature. After this, it would require a lot of time and effort to first make these results available to Cervantes' scholars for their classification, justification, and annotation; and then gather them to be placed in the final edition. Assuming of course that access to the original copies was granted and after a long period of time that this process has been completed, users will not have access to copies of the original sources. It is easy, after imagining the previous

scenario, that advances in information technology can make this process feasible and much simpler.

Hyperlinked and Synchronized Texts and Images

Since our collection includes images of *Don Quixote*'s original copies along with their textual transcriptions, both the MVED and the reader's interface presents the user with a synchronized browsing option, displaying the text and its corresponding image in parallel. Clicking on the image, the text is scrolled to its corresponding offset. Conversely, clicking on the text, a rectangle is drawn in the image's coordinates that correspond to the text. Along with this synchronized browsing mechanism, more than one image can be depicted in separate windows for the user to compare them. In some instances, the images are not clear because the original page in the book was damaged or had stains; therefore, we are using a tool called Tilepic[5], which allows the user to zoom in sections of the image to a much higher size, helping to identify unclear or blurring letters and accents. This synchronized and interlinked browsing feature is possible because of the flexibility provided by the hypertextual nature of the collection.

In the case of the reader's interface for the Electronic Variorum Edition, the user is presented with the base text, and highlighted in yellow are all the variants. Clicking on a variant, a new window presents the string of characters both in the base text as the ones in the comparison texts, so the user can see all instances of the variant across all the texts in the collation. Along with the variant, the user can also look at the justification made by the editor, as well as commentaries and annotations by different scholars. It is also possible to compose a customized critical edition, to do so, the user can specify to include only the commentaries done by one editor, or limit the edition to some of the categories of variants.

Discovering *Don Quixote* Using Interactive Timeline Viewer (ItLv)

Until now we have been discussing the main features of our project, but what knowledge about Don Quixote can be discovered from this collection? Initially the creation of a Variorum Edition including copies of the first edition of *Don Quixote* can show its evolution during a period of time, along with valuable commentary from several scholars. Analyzing the variants, users are able to eliminate printing and compositor's errors, producing a closer version of the original manuscript written by Cervantes.

At this point, one of the challenges is how to present the results of the collation in such a way that the user can analyze them. We are currently using ItLv[6] to help users with this process. In ItLv, the results of a collation are depicted in a two-dimensional display, in which the X-axis can depict any attribute related to a variant; for example, the edition in which it appears. Similarly, the Y-axis depicts the offset in characters from the beginning to the end of a chapter (this is the unit we chose for our analysis). The variants are depicted as rectangles whose length is proportional to the variant's length in characters. Mousing over any of the rectangles—representing a variant—, a new window pops up,

depicting all the information related to that variant. Users can also filter the contents of the collation to narrow the scope of the analysis or to discover patterns across the texts.

With the use of ItLv as visualization tool we expect to analyze textual patterns across the texts. For instance, it can be used to identify the use of certain abbreviations in the texts; in this case, depicting and analyzing one chapter of Don Quixote, we found for that chapter that the abbreviation—replacing the string "que"—was used across some pages but not in others. Considering that the string "que" appears regularly across the whole chapter, we can therefore raise the hypothesis that the pattern of the abbreviation as a compositional variant could reflect the preference of an individual compositor in charge of those sections of the book.

Users can also visualize how the text has evolved over time. For example it is possible to identify in what edition a variant was introduced, and how that variant survived through other later editions. Similarly, it gives evidence to identify what edition or editions were used to compose another edition. However, ItLv's added value is not only the possibility to depict the evolution of the variants and editions, but also the richness of the scholarly commentary attached to them along with the possibility of navigating through the texts and their corresponding images.

The Textual Iconography of *Don Quixote*

Although the first edition of *Don Quixote* published in 1605 was a non-illustrated edition, hundreds of illustrated editions have been published since then. Illustrations in these editions constitute a rich visual narrative of the novel. Several attempts have been made to create a comprehensive textual iconography of the *Quixote*. López Fabra's edition[7] published in 1879 included 101 illustrations from 60 editions. In 1905, Henrich[8] publishes an edition including 611 illustrations; however, this edition was limited to title pages. In 1947, Juan Givanel's edition included 77 engravings, finally, a more ambitious edition including 800 engravings and drawings was published in 1836.

With the advances in information technology and computer science, we are working on the creation of a repository, interfaces, and visualization tools to archive, access, and visualize a collection of digitized illustrations of more than 200 editions of *Don Quixote*, ranging from the early 17th century to 1930 --a span of nearly 400 years--. Our initiative focuses especially on English, Spanish, and French illustrated editions of the 18th and 19th centuries. We estimate that our repository of illustrated editions will incorporate about 7,000 digital images from over 450 volumes.

The illustrations will then be catalogued, and their corresponding segment in the text stored. However, texts will need to be tagged to provide structure and semantics. With a tagged text it is possible to find meaningful segments within the text, e.g. adventures, episodes, characters, and locations; which in turn makes it possible to associate and link them to the illustrations.

The On-line Picasso Project: An Overview

The second project is the On-line Picasso Project, a digital reasoned catalogue of Picasso's artworks hosted at Texas A&M University, as a joint project between the Department of Hispanic Studies and the Center for the Study of Digital Libraries. The main goal of the On-line Picasso Project is to provide a comprehensive repository of information, texts, and images about Pablo Picasso and his pictorial creation. To achieve this, the project is divided into six main components: a) Biography, a detailed day to day textual documentation of Picasso's artistic life, b) Database of Works, a visual catalogue of the complete works of Pablo Picasso organized by additive arbitrary numeration, c) Museum List, a database of real locations where the works may be viewed, d) Bibliography, a comprehensive list of references to scholarly publications on Pablo Picasso, e) Archives, a list of news items referring to Pablo Picasso's exhibitions, auctions, etc., and f) Search Engine.

Alternating Simultaneous Reasoned Catalogue

The On-line Picasso Project currently holds over 7,000 images of paintings, drawings, sketches, and sculptures; with items constantly being added to the collection. Using a browser, users can explore the collection from at least two simultaneous different perspectives; in the first one they are presented—divided by year—with a table containing thumbnails of each artwork. Users can then navigate through the collection to see Picasso's creation in different years. If there is a particular interest to know more about an item, clicking on its thumbnail a new window pops up, depicting all the metadata associated with that item, such as title, place of execution, medium, dimension, date, and location where it can be found. In some instances it also includes price, past exhibitions, along with some scholarly annotations. In the second perspective, users are presented—also divided by year—, with a chronology of historical events, including places, and people relevant to Picasso's artistic creation in some month or season of the year, along with links to his artworks. Thus, as users read through the chronology they can visualize the related artworks.

This browsing environment allows users to navigate and explore the entire collection. However, due to the static nature of the hypertext markup language (html), it lacks flexibility for more advanced analysis; for instance: a) searching for patterns in two different periods of time, b) determining whether the technique used, the place of execution, the period of the year, and the theme painted have any correlation, c) finding trends in the artist creation in two distant periods in time, and d) discovering correlations between historic events, Picasso's personal life, and his artistic creation. Hence the need to incorporate and experiment with other tools.

Although Picasso wrote only between 1935 and 1958, his writings are full of words referencing elements in his paintings, which makes very difficult if not impossible to understand his writings without browsing through his artworks. ItLv provides a parallel text/artworks display that helps users to simultaneously read the text and navigate through the paintings.

Another characteristic of Picasso, is the creation of series, we can mention for example, *The Minotaur Series*, *The Cabana Series*, and *The Magic Series*. In the context of series, it is often the case that scholars disagree either on the membership of a particular item to the series, or the order of an item within the series, particularly when it was not stated by Picasso. ItLv provides a feature that enables users to create, modify, and compare series of artworks.

Exploring Picasso's Artistic World

From a technical point of view, it is worth to point out some issues—regarding dates—that give special characteristics to the Picasso collection. In some instances it is not well know when Picasso began painting an artwork, nor when he completed it. Thus, a mechanism to represent uncertainty is required. We adopted a color-fading scheme, in which a fading color indicates that the date is uncertain; whereas a solid color represents a certain date. Also, the dating format is not uniform; for example, in some cases, he dated his paintings as "Summer of 1907", "Late 1927", or "Early Spring 1908", in other cases the exact date appears in the painting. This lack of uniformity requires an initial process to convert the dates into a normalized format.

When using ItLv with the Picasso's collection[9], time is depicted in the X-axis. Users can select any attribute of the collection to be depicted in the Y-Axis. Then the content is plotted in a two dimensional display, where each artwork is depicted by a rectangle whose length is proportional to the duration of the creation of that piece. Any of the attributes of the metadata can be depicted as a label next to the rectangle, which can be toggled on and off. However, it is very common that the display gets too cluttered due to the high number of entries being depicted. Therefore, a second level of detail is provided by another window, which depicts only the items of the category the user wants. To further explore the information of any given item, users are presented with another window containing all its associated metadata. The user can also customize another pop up window to depict only the attributes relevant to a specific analysis.

We mentioned earlier that discovering facts was one of the goals of using this tool with the Picasso collection. Let us illustrate this with one example. While visualizing the entries for 1927, we selected the place of execution as label. It was immediately evident that Picasso was very productive during the summer of that year while he was in Cannes (France)—where seventy five percent of his artworks thus far catalogued were made—, in contrast to the rest of the year while he was in Paris –producing the rest twenty five percent--. Also given that the color of the rectangles represents the medium used, the display showed that Picasso mostly used India ink with pencil and pencil alone. From the art scholar point of view, it is possible to raise the hypothesis that, along many other factors, the place where he was living at that time might have influenced the technique he used that summer. As our work progresses, we expect to use ItLv to perform more advanced analysis about Picasso. In particular analysis related to trends and patterns across several years and decades.

Thematic Series in Pablo Picasso

Pablo Picasso is also known for the creation of series of paintings. Quite often events in his life influenced the theme and contents of such series; Stassinopoulos [10] illustrates Picasso's artistic production after he was abandoned by his lover at the end of his life; "But he suffered alone, and on November 28, a month after his seventy-second birthday, he stopped talking, took his despair in hand, and started working. He worked feverishly, and in just over two months produced 180 drawings." Although in some cases Picasso stated the order of the items within a series, there are instances where the order is not indicated, thus leading to a scholarly discussion about the order within the series.

Manipulating series of artworks without the advances in information technology is a very difficult task. Scholars, while taking notes of their observations and findings, have to rely on printed catalogs, where they have to go back and forth in different catalogs. With ItLv, the user can compose series by selecting items from the collection. The position of an artwork in the series can be changed as many times as necessary. As the user is composing the series, a slide show presents the items added to the series. The slide show's speed can be changed from very slow to very fast, as well as the direction between backward and forward. As the user is presented with each item in the series, its associated metadata is also displayed; therefore, the user can visualize how the series evolves over time.

The user can create as many series as needed, and up to four series can be compared simultaneously. To do this, one of the series is selected as the "base series", which in turn is compared against the other series. Those items in the comparison series that do not match its corresponding item in the base series are highlighted in a different color, helping users to find evidence about the differences among the series.

We discovered that flexible, multiple organizations within specific series in 1927 –the *Cabana Series*, for instance– allow for a more enlightened interpretation of the individual works making up the series. From the art scholar's perspective, the different visualizations allow us to raise a wide range of hypotheses. Thus, we could stipulate that, along many other factors, place and time might have influenced the technique the artist used, and that the integration of a work in one of the series might have determined the extension of its features to other members of the same series and the simultaneous demarcation of differences with respect to works in other series.

Conclusion

Our work with this two collections shows the importance of advances in information technology and digital libraries applied to the humanities –literature and arts- in this case. And also how their application can lead to the creation of tools, which in turn provide new ways to explore and study a literary or artistic collection. We also experience not only the benefits new technologies provide to scholars in the humanities, but also the challenges that practices in the humanities presents to the development of new technologies.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant no. IIS-0081420. Support for this work was also 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 College of Engineering, the College of Liberal Arts, the Glasscock Center, and the Texas A&M University Libraries.

References

[1] The Cervantes Project, http://www.csdl.tamu.edu/cervantes/

[2] Enrique Mallen, The On-line Picasso Project http://www.tamu.edu/mocl/picasso/

[3] Eduardo Urbina, Richard Furuta, Arpita Goenka, Rajiv Kochumman, Eréndira Melgoza, and Carlos Monroy, "Texto, contextos e hipertexto: la crítica textual en la era digital y la Edición electrónica variorum del Quijote," *Quaderni di Letterature Iberiche e Iberoamericane* (Milan) 27 1999-2000 [2002]: 21-49.

[4] Eduardo Urbina, Carlos Monroy, and Richard Furuta, "Iconografía textual del *Quijote*: repaso y nueva aproximación de cara al IV centenario." *In limine al IV Centenario del Quijote. Coloquio Internacional de la Associazione Cervantina di Venecia, Ateneo Veneto.* Venice, April 2003. Forthcoming.

[5] Tilepic. Berkeley Digital Library Project, http://elib.cs.berkeley.edu/tilepic/

[6] Carlos Monroy, Rajiv Kochumman, Richard Furuta, Eduardo Urbina, Eréndira Melgoza, and Arpita Goenka. "Visualization of Variants in Textual Collations to Analyze the Evolution of Literary Works in The Cervantes Project," Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries, 2002, pp. 638-653.

[7] *"Iconografía de Don Quixote"*. Barcelona: P. Riera, 1879.

[8] Henrich, Manuel. *Iconografía de las ediciones del* Quijote *de Miguel de Cervantes*: *Reproducción en facsímile de las portadas de 611 ediciones con notas bibliográficas*... Barcelona: Henrich y Cía., 1905.

[9] Carlos Monroy, Richard Furuta, and Enrique Mallen. "Visualizing and Exploring Picasso's World" To appear in Proceedings of the Joint Conference on Digital Libraries, Houston, Texas 2003.

[10] Arianna Stassinopoulos *Picasso: Creator and Destroyer*, Simon and Schuster, New York, 1988.