Building a Digital Collection of Manuscripts from the Library of the Royal Palace of Spain

Soledad Vélez, Manuel Sánchez-Quero, Juan Carlos García and Alejandro Bia
Miguel de Cervantes DL, University of Alicante
P.O. Box 99, E-03080
Alicante, Spain
{sole, manuel.sanchez, jcgarcia, alexbia}@cervantesvirtual.com

ABSTRACT

With an aim of bringing cultural contents to cyberspace and spreading some unknown aspects of the history of the Americas, the Miguel de Cervantes Digital Library has embarked in a joint effort with the Library of the Royal Palace of Spain, to develop the digital web publication of the Manuscripts of the Americas in the Royal Collections funds. In this joint venture, the Library of Royal Palace supplied its invaluable contents for digitization, and the Miguel de Cervantes DL its technology and experience as a digital publisher. The goal was to join the ancient and the new, the most precious and carefully preserved documents with the new electronic publishing technologies. The result was to make freely available to a worldwide public those otherwise unreachable treasures of the Royal Collections. We expected this effort to have an impact both on the general reader and on the specialized researcher. In both cases they would be spared the time and cost of a trip to Madrid, and the risk of an access denial. In all cases, the funds would gain exposure with a reduction in physical contact which is desirable for long term preservation.

With a long term vision we set high level quality requirements to obtain accurate digital facsimiles of the originals not only for the present state of the art of web technology but for future higher standard of graphic display. With an innovative spirit we applied the best available technology at the service of culture.

1. THE LIBRARY OF THE ROYAL PALACE

The Library of the Royal Palace of Spain has its origins in the private libraries of the kings of the House of Borbones. In the National Patrimony collections, which include the Library of the Royal Palace, we can find unique editions, rare books, incunabula, manuscripts, old printings, ancient maps, musical scores, drawings, engravings, rare book bindings, coins, tapestry, furniture and much more. Most of the Manuscripts of the Americas collection comprises manuscripts of the 17th and 18th centuries, the kings' letters, royal credentials, royal bans, etc., all of great historical and cultural value: a wide and valuable testimonial set of the American colonial period.

2. TECHNICAL OVERVIEW

We started building this digital collection in two parallel fronts. On one hand, cataloguing the funds, and on the other hand, digitizing them. For cataloging, we defined an XML (eXtensible Markup Language) mark-up scheme based on the Master guidelines [5], and developed our own software for the automatic transformation of catalogue records to nicely formatted HTML (HyperText Markup Language) for web publication.

The digitization of the funds was partially done by the Miguel de Cervantes DL, and partially assigned to a private company that followed our quality standards.

After digitization was done, partially-automated digital image processing was applied to the top quality big-size preservation images to obtain both thumb-nail images to create graphical indexes and medium-size medium-quality images for adequate web transmission and display on nowadays screens. In this process, high quality TIFF images were converted to lower quality JPEG ones. Let's remember that the JPEG format offers a high level of compression for higher speed transmission but with some loss of quality.

The last stage in digital facsimile production was the automatic generation of HTML facsimile ensembles using FacsBuilder, a tool we have developed to speed-up de mounting of facsimile sets for web display.

To complement this huge digital publishing effort, a lot of care and meticulous work was devoted to graphic design of the web site where the digital collection would be displayed.

2.1 Cataloguing

For cataloguing, we followed the MASTER guidelines (Manuscript access through Standards for Electronic Records). MASTER is a project partially funded by the European Union with the purpose of defining a general norm based on XML for the description of manuscripts. This working group developed a DTD (Document Type Definition) for manuscript descriptions that is compatible with the recommendations of

the TEI Consortium (Text Encoding Initiative) and is expected to be included in future editions of the TEI Guidelines. In practice, the Master tag set is larger and more complex than the TEI Header, the TEI subset devoted to bibliographic information.

The use of the MASTER guidelines for cataloguing manuscripts offered us a wide range of possibilities through its rich markup system for authorities, entities and toponyms. It allows the librarian to develop very rich descriptions of manuscripts, while it allows the user to make both simple and also highly accurate searches to fulfill all kinds of requirements.

MASTER permits a multilevel bibliographical description: a general descriptor of the manuscript (main record), and several analytic descriptors for each one of the parts that conform the document.

2.2 Advantages of XML-Master

The use of an XML format for bibliographic metadata grants compatibility and easy interchangeability between platforms and applications by means of simple transformations as XML and XSLT are gaining ground as data exchange standards [4]. This ease of transformation to other formats plays an important role in the preservation of metadata by preventing obsolescence of the format. The open text nature of XML markup makes it easy to read and to process giving independence from other obscure proprietary formats.

Another plus in using XML is the possibility of generating different output formats with different renderings by means of XSLT (XML Stylesheet Transformations) [6]. An example of these transformations is the automatic processing of XML records for research purposes, statistical analysis, insertion in databases, complex searches, etc.

2.3 Advanced searches

The use of XML for digital publishing allows for complex searches based on semantic tags. For instance, we can search the word "Aragn" only in responsibility statements and only when it is contained between proper name tags.

Searches based on the structure of the document, which are not possible with ordinary relational database architectures are becoming more and more common. For example, we can search for the word "Aragon" in responsibility statements but now only when it is contained in an organization name.

2.4 Software

At the Miguel de Cervantes DL, we currently do research in new technologies and technological innovation concerning digital publishing [2], natural language processing [3], computational tools for linguistic research [7] and web publishing technologies. For this project we have developed highly complex software, like FacsBuilder [1], which saves time, dramatically reducing production costs. Starting from a sequence of numbered images and two templates the facsimile edition is automatically assembled.

2.5 Transformation process

The original bibliographic records from the Royal Palace were sent in MARC format. We had to decode them and develop the necessary structures to keep the semantic information associated to the cataloguing data. For this purpose, we had to identify relationships and dependences. The result was the automatic generation of MASTER records from MARC ones. However, expert revision by a librarian was necessary to assure the quality of the resulting data.

3. REFERENCES

[1] A. Bia. A Versatile Facsimile and Transcription Service for Manuscripts and Rare Old Books at the Miguel de Cervantes Digital Library. In *Proceedings of First ACM/IEEE-CS Joint Conference on Digital Libraries*, page 477, Roanoke, Virginia, USA, June 2001.

[2] A. Bia and R. C. Carrasco. Automatic DTD simplification by examples. In *ACH/ALLC 2001. The Association for Computers and the Humanities, The Association for Literary and Linguistic Computing, The 2001 Joint International Conference*, pages 7–9, New York University, New York City, June 2001.

[3] A. Bia and R. Muñoz. Aplicación de Técnicas de Extracción de Información a Bibliotecas Digitales (Applying Information Extraction Techniques to DLs). In M. V. Ferro, editor, *Proceedings of the XVI Conference of the SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural)*, volume 26, pages 207–214, University of Vigo, Spain, September 2000. SEPLN. (published in: Procesamiento del Lenguaje Natural, journal of the SEPLN).

[4] N. Bradley. *The XML Companion*. Addison-Wesley Longman Limited, 2000. BIMICESA.

[5] L. Burnard and P. Robinson. Vers un standard européen de description des manuscrits: le project master. In J. André and M.-A. Chabin, editors, *Les documents anciens*, volume 3 of *Document numérique*, pages 151–169. Hermes Science Publications, Paris, juin 1999.

[6] M. Kay. *XSLT Programmer's Reference*. Wrox Press, 1102 Warwick Road, Acocks Green, Birmingham, B27 6BH, UK, 1st edition, 2000.

[7] A. Zaslavsky, A. Bia, and K. Monostori. Using Copy-Detection and Text Comparison Algorithms for Cross-Referencing Multiple Editions of Literary Works. In P. Constantopoulos and I. Solvberg, editors, *Research and Advanced Technology for Digital Libraries:5th European Conference, proceedings/ECDL 2001*, volume 2163 of *Lecture Notes in Computer Science*, pages 103–114, Darmstadt, Germany, 4-9 September 2001. Springer-Verlag.