Mass-digitization at the Complutense University Library. Access and preservation of its bibliographic heritage

José Antonio Magán, Manuela Palafox, Eugenio Tardón, Amelia Sanz. Universidad Complutense de Madrid

Barcelona, June 29, 2011
Overview

- Bibliographic Heritage Digitization in the UCM
- Preservation of the Complutense digital contents
- Collaborative Annotation of Digitized Literary Texts
Goals of the digitization Project with Google

- Promoting access, dissemination and preservation of the bibliographic heritage
- Adding UCM holdings to those of some major libraries in order to facilitate the access of the global academic community to the scholarly knowledge
- Designing a plan for the conservation and restoration of damaged books
- Providing to UCM scholars a corpus of digital materials which makes possible to develop digital projects
UCM and Google partnership agreement

- Cataloging Plan: 220,000 volumes were cataloged, from the 16th to 19th centuries
- Criteria Selection Guidelines:
  - Date of publication: from 16th Century to 1870
  - Book physical conditions: dimensions, text block and binding
- Binding of the 19th century books
- Over 120,000 books digitized
- Dissemination of the bibliographic heritage of the Complutense University Library

Information flow

Books selection using the PDA

Metadata export from the Catalog to the Web Management System

Export the information to the Catalog of the books physical conditions

Export data files from the PDA to the Web Management System

Physical conditions of the digitized books by Google

Non digitized for text-block

- papel frágil: 21%
- hongos: 10%
- deterioro físico: 47%
- hojas sueltas: 20%
- intonso: 2%
Physical conditions of the digitized books by Google

Non-digitized for the binding

Dissemination and access

Preservation of the UCM digital contents

- For the implementation of a digital preservation program, 2 needs were identified:
  - A large-scale digital content repository
  - A scalable technological and organizational potential

- Conclusion:
  - The digital preservation could only be achieved as a result from the cooperative involvement of academic institutions following the international library community standards.
  - UCM joined HathiTrust on November 2010
Books digitized by Google and in-house in HathiTrust

HathiTrust Digital Library

Source: Jeremy York (data as of May 1, 2011)
### UCM Ingest

<table>
<thead>
<tr>
<th>Date</th>
<th>Volumes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>December 2010</td>
<td>78,256</td>
<td>78,256</td>
</tr>
<tr>
<td>January 2011</td>
<td>1,156</td>
<td>79,412</td>
</tr>
<tr>
<td>February 2011</td>
<td>6,125</td>
<td>85,537</td>
</tr>
<tr>
<td>March</td>
<td>2,774</td>
<td>88,311</td>
</tr>
<tr>
<td>April</td>
<td>15,486</td>
<td>103,797</td>
</tr>
<tr>
<td>May</td>
<td>2,947</td>
<td>106,744</td>
</tr>
</tbody>
</table>
Visits of the UCM book pages in HathiTrust

Europeana Libraries

Europeana Libraries

• 1,200 film and video clips
• 850,000 images
• 4.3 million texts (books, journal articles, theses, letters)
• Complutense University Library will aggregate:
   Complutense Rare Books from 16th Century to 1870: Metadata of more of 120,000 books digitized by Google.
   Engravings of Dioscorides Collection: 50,000 engravings approximately.
   Complutense University Theses: 6,000 theses.
   Journal articles published by Complutense University Press: 31,000 articles
   Fine Art Old Drawings from the Fine Arts Faculty (between 1752 and 1914): 287 drawings.
   Photos of the Spanish Civil War from the Historical Archive of the Communist Party of Spain: 3,200 photographs approximately.
   Personal Archive of Ruben Dario: 5,000 documents.
   Complutense Manuscripts & Incunabula
   Cartographic material, maps and city views


Europeana Data Model
Collaborative Annotation of Digitalized Literary Texts Project
Our commitment to the digital humanities

Posted by Jon Orwant, Engineering Manager for Google Books, Magazines and Patents
(Cross-posted from the Official Google Blog)

We've given awards to 12 projects led by 23 researchers at 15 universities:

- Steven Abney and Terry Szymanski, University of Michigan. Automatic identification and Extraction of Structured Linguistic Passages in Texts.
- Dan Cohen and Fred Gibbs, George Mason University. Reframing the Victorians.
- Miles Efron, Graduate School of Library and Information Science, University of Illinois. Meeting the Challenge of Language Change in Text Retrieval with Machine Translation Techniques.
- David Mimno and David Blei, Princeton University. The Open Encyclopedia of Classical Sites.
- Amelia del Rosario Sanz-Cabrerizo and José Luis Sierra-Rodríguez, Universidad Complutense de Madrid. Collaborative Annotation of Digitalized Literary Texts.
I am not able to light on any map or work giving the exact locality of the Dracula, as there are no maps of this country as yet to compare with the Ordnance Survey Maps; but I found that Bistritz, the post town by Count Dracula, is a fairly well-known place. I shall enter here part of my notes, as they may refresh my memory when I talk over my life with Mina.

In the population of Transylvania there are four distinct nationalities: Walachians in the South, and mixed with them the Wallachs, who are the descendants of the Dacians; Magyars in the West, and Szekelys in the East and North. I am going among the latter, who claim to be descended from Attila and the Huns. This may be so, for when the Magyars conquered the country in the eleventh century they found the Huns settled in it.

I read that every known superstition in the world is gathered into the Carpathians, as if it were the centre of some sort of imaginative whirlpool; if so my stay may be very interesting. (Mem., I must ask the Count all about them.)

I did not sleep well, though my bed was comfortable enough, for I had all sorts of queer dreams. There was a dog howling all night under my window, which may have had something to do with it; or it may have been the paprika, for I had to drink up all the water in my carafe, and was still thirsty. Towards morning I slept and was wakened by the continuous knocking at my door, so I guess I must have been sleeping soundly then.

I had for breakfast more paprika, and a sort of porridge of maize flour which they said was "mamaliga", and egg-plant stuffed with forcemeat, a very excellent dish, which they call "imletata". (Mem., get recipe for this also.)

I had to hurry breakfast, for the train started a little before eight, or rather it ought to have done so, for after rushing to the station at 7:30 I had to sit in the carriage for more than an hour before we began to move.

It seems to me that the farther east you go the more unpunctual are the trains. What ought they to be in China?
Conclusions

• The large-scale digitization with Google will allow the Complutente University Library to face successfully to the new challenges are posing to research libraries.

• Complutense digital collection of over 120,000 books is in HathiTrust, the most innovative digital preservation project for the library community.

• Complutense Library is collaborating with Europeana Libraries to increase the dissemination of our digitized heritage and participate in an innovative research area related to the semantic web and linked data.
Thank you very much!
Questions?

mpalafox@buc.ucm.es