Case study 1: Google Books at the Complutense University of Madrid

CERL Annual Seminar 2012
October 30 2012, British Library

José Antonio Magán Wals
Antonio Moreno Cañizares
Manuela Palafox Parejo

Complutense University of Madrid Library
The Complutense University of Madrid and its Library

- The Complutense University of Madrid has 85,000 students and 6,289 scholars
- 34 Libraries
- 3 million books, the first academic library in Spain
- 11,300 seats for reading and 1,500 computers
- 411 librarians
- An important digital collection with more than 600,000 objects and millions of scanned pages
Our commitment to collaboration for digitization and dissemination of scientific production and heritage

- Opting for open access dissemination through both the called "green route" as in the "golden path".

- The library acts as digital publisher in collaboration with other university services.

- Collaboration for the dissemination and digitization:
  - **with external institutions and agencies:** Spanish Government, Federal Government of Madrid, Madrid Academic libraries consortium, Europeana, Hathi Trust, Internet Text Archive
  - **with private institutions:** Google, Santander Universities, Health Sciences Foundation, Editorial Extramuros,
  - **with commercial publishers and distributors:** Springer, Thomson Reuters, Proquest, E-Libro
Complutense Digital Collections: a) Academic works

- 25,000 digital dissertations (5,500 of them in open access)
- 30,000 articles from journals published by our university
- 11,000 e-prins in open access
Complutense Digital Collections:
b) Materials for research support

- 400,000 newspapers, photographs from Spanish Civil War, Drawings from the School of Arts...
Complutense University of Madrid Digital Colections:
c) Ancient books and cultural heritage

- 125,000 out of copyright books digitized
- 47,000 Prints
- The greatest digital collection in Spain of Ancient Books
Status of Complutense ancient books digitization in 2006 (Dioscorides Collection)

- 2,800 scanned books in 10 years (the largest university collection in open access in Spain).

- At this rate it would take 435 years scanning to digitize the number of works digitized with Google in 3 years.

- Portal deficiencies:
  - No long term digital preservation
  - No multilingual
  - No copyright managed
  - Not adapted to social web
Complutense-Google Agreement

- September 2006. Signing of the cooperation agreement to mass digitization of library collections free of copyright.

- Google Books has scanned more than 20,000,000 books, much of them from libraries:
  - U.S.: Michigan, California, Harvard, Stanford, New York Public Library...
  - Europe: Oxford, Bavarian State Library, Complutense of Madrid, Catalonia, Ghent, Lausanne, Lyon Municipal Library. ..
    Recently joined several national libraries: Florence, Rome, Czech Republic, Austria, Netherlands ...
  - Asia: Library of Keio University, Japan

- Each 6 months more than 90% of the books are visited
A controversial project

• For some people:
  - Violation of the rights of authors and publishers
  - Risk of monopolization of access to the content of books
  - Transfer of public cultural heritage to a commercial company
  - Scanning without enough quality: poor images and OCR.

• For others:
  - Unique opportunity to democratize knowledge by digitizing
  - Creates a free tool that allows you to query the contents of millions of books and download them for reading
  - Stimulates other public and private projects of mass digitization

• The facts:
  – Participating libraries have created with their digital copies important public collections of scanned books.
  – Now when you search Google you find not only information in websites but books that can be downloaded
What does Google do?

- Scans documents and is responsible for the costs: Books are scanned twice to avoid errors.

- Out of copyright scanned books are freely searchable and downloaded from Google Books.

- Create an exclusive interface for the University and its users, to access to and download digital works of the program.

- Gives to the Complutense a copy of the scanned books.
What does the Complutense Library do?

• We provide the books and experts who oversee the selection of the works to be scanned.

• Update metadata.

• Select and organize the movements of the books to ensure the integrity of collections.

• Preserve and disseminate its digital copy: these copies are used for academic projects.
Presentation

The goal of Dialogyca BDDH (Biblioteca Digital de Diálogo Hispánico) is to progressively provide the corpus of all Hispanic literary dialogues written in the different languages of the Peninsula, including Latin-based, Spanish-Hebrew and Spanish-Moorish. During this first phase, we have concentrated especially on the dialogues of the 15th and 16th centuries. Each data sheet or record corresponds to a single dialogue for which the following information is provided: data on the authorship and literary characteristics, primary bibliography, secondary bibliography and, in addition, the reproduction of the testimony most relevant for the transmission of the text. Thus, when the dialogue on one record is part of a volume along with another type of text or other dialogues, two data sheets have been created. Both of these have internal links to make the consultation easier: one link is for the dialogue and other is for the book in which it is conserved.

Target audience

Dialogyca BDDH is aimed at researchers and people interested in a wide variety of specialities, since dialogue is a genre that has served for the transmission of information from all of the branches of knowledge: philosophy, medicine, engineering, spirituality, astronomy, etc. The purpose is to provide the community with comprehensive, coherent, stable and rigorously prepared information created with the tools of philology.

Access

There are two types of access: A) Free: for those wishing to access general information about the work, and B) Restricted: for researchers who would also like to consult more specialised documents (digitised text, bibliographic descriptions and secondary bibliographies). The project is open to all interested researchers. To make collaboration easier, the team has developed a set of users’ guidelines or standard operating manual (Plan de Trabajo Normalizado) that can be found at: http://eprints.ucm.es/8638/
<table>
<thead>
<tr>
<th>Código</th>
<th>Autor</th>
<th>Título</th>
<th>Datos</th>
<th>Copia</th>
<th>Descripción</th>
<th>Reproducción digital</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>López de Villalobos, Francisco</td>
<td>Diálogo que pasó entre un grande deste reino de Castilla esteando con el hijo de la cuartana y el doctor Villalobos estando allí con él en presencia de varios nobles y señores.</td>
<td>1520-1608</td>
<td>Londres, British Library Add.6129 fols. 149v-154r</td>
<td>[BDDH141M1 - 229KB]</td>
<td>[BDDH141M1 - 1.61MB]</td>
</tr>
<tr>
<td>2</td>
<td>López de Villalobos, Francisco</td>
<td>Este es el resumen de un diálogo que pasó entre un grande deste reino de Castilla esteando con el hijo de la cuartana y el doctor Villalobos estando allí con él en presencia de varios nobles y señores.</td>
<td>1347</td>
<td>Madrid, Real Biblioteca (Palacio) II-2057, fols. 41v-47v</td>
<td>[BDDH151M1 - 222KB]</td>
<td>[BDDH151M1 - 3.48MB]</td>
</tr>
<tr>
<td>3</td>
<td>López de Villalobos, Francisco</td>
<td>Este es un diálogo que pasó entre el duque de Salamanca y el doctor Villalobos estando allí con él en presencia de varios nobles y señores.</td>
<td>1347</td>
<td>Madrid, Real Biblioteca (Palacio) II-2057, fols. 41v-47v</td>
<td>[BDDH151M1 - 222KB]</td>
<td>[BDDH151M1 - 3.48MB]</td>
</tr>
<tr>
<td>1</td>
<td>López de Villalobos, Francisco</td>
<td>Libro intitulado: Los problemas de Villalobos, que trata de cuerpos naturales y morales, y dos diálogos de medias de tres grandes, y una canción y la comedia de Amphytion.</td>
<td>1543, 9 de febrero</td>
<td>Madrid, Real Biblioteca (Palacio) II-2057, fols. 41v-47v</td>
<td>[BDDH151M1 - 222KB]</td>
<td>[BDDH151M1 - 3.48MB]</td>
</tr>
</tbody>
</table>

Ejemplar digitalizado: Madrid UCM Bibl. Histórica "Marqués de Valdencilla" BH FG 1560

Link: [afuma.sim.ucm.es/diosconides/consulta_libro.aspx?ref=B33239189&droma=0]

Código: 2

Autor: López de Villalobos, Francisco
Título: Libro intitulado: Los problemas de Villalobos, que trata de cuerpos naturales y morales, y dos diálogos de medias de tres grandes, y una canción y la comedia de Amphytion.
Lugar: Zamora, a espescias de Pedro Bermúdez y Bartolomé Núñez
Impresor: Juan Cobiá
Año: 1544, 15 de enero

Ejemplar digitalizado: Madrid UCM Bibl. Histórica "Marqués de Valdencilla" BH MED 1626

Link: [afuma.sim.ucm.es/diosconides/consulta_libro.aspx?ref=X53375076&droma=0]

Código: 3

Autor: López de Villalobos, Francisco
Título: Libro intitulado: Los problemas de Villalobos, que trata de cuerpos naturales y morales, y dos diálogos de medias de tres grandes, y una canción y la comedia de Amphytion.
Lugar: Sevilla
Impresor: Cristóbal Alvarez
Año: 1530, 30 de octubre
Project Planning and Design: 2007

Actions

- Collection Analysis of the works and the libraries involved.
- Progress reports: Data on facilities, access to repositories of books, etc.
- Selection criteria guide: fixed criteria (date of publication) and criteria for the scanning condition.
- 19th Century books bookbinding plan and Recommendations Guide.
- Scanning Program: workflows, schedule and logistics operations (even cleaning books and the book repositories)
Project Planning and Design: 2008-2011
Actions

- Cataloging Plan: 220,000 books cataloged.
- Analysis of the conservation status and selection criteria for scanning in 145,000 books.
- Scanning in the Google Scanning Center of more than 200,000 Books (120,000 from Complutense Library, the rest from Catalonia libraries)
- In June 2011 we finished the operations
Technological Developments: Web application for project management:

- Offers information online and in real time of all the daily movements of books, Google shipments, returns, preservation status of the books, etc. with statistical data on project operations.

- Stores the metadata of the books included in the digitization project.
Technological Developments: PDA application

- For selection tasks in storage libraries
- It reads the bar code book.

- We present the characteristics and condition of the book according to the selection criteria guide
  - Dimensions: height, width and thickness.
  - Binding Type: Valuable, weak, lost, rebinding, with opening problems, impaired.
  - Sheets: fungi, physical deterioration, flyers, fragile paper, uncut.

- Such information is exported to Web Management System and then to the Library Catalog.
How do you access the Complutense digitized books? 1. Searching anything in Google (or Google Books or Google Play)
Los misterios de Londres, Volumenes 8-10 (Google eBook)

Ebook (gratuito)

Apto para:
Web
Tablet/Pad
eReader
Teléfono
Inteligente

Funciones:
Secuencia de lectura
Páginas originales
Ayuda con dispositivos y formatos

Leer el libro GRATIS
Descarga ePub, PDF - leer instrucciones para eReader

Más información sobre los libros de Google Play

Comentarios de usuarios - Escribir una reseña

No hemos encontrado ninguna reseña en los lugares habituales.

Libros relacionados

https://play.google.com/store/books/details?id=WMkBt3z0C&rdid=book=WMkBt3z0C&dote=1
Los misterios de Londres: v. 8-10
Paul Fieral

Has seleccionado...

Google Play

madagascen1@gmail.com

Gratis
Podrás leer este libro en la Web y en dispositivos móviles compatibles.

Ya puedes leer tu libro.

Compartir

Google Play Books
Google Inc.

Las millones de libros en la Web. Lleva contigo tus libros favoritos y sigue leyendo estos donde estés.

Leer ahora en

Android

Chrome

Google Play Books
Los misterios de Londres, Volumenes 8-10 (Google eBook)

Comentarios de usuarios - Escribir una reseña

No hemos encontrado ninguna reseña en los lugares habituales.

Libros relacionados
Remember: Every 6 months more than 90% of the 20.000.000 Google Books are visited
How do you access the Complutense digitized books?

2. Exclusive Google search interface for searching Complutense books.
How do you access the Complutense digitized books?

How do you access the Complutense digitized books?

4. HathiTrust Digital Library.
What’s Hathi Trust?

- A library consortium to ensure that the cultural record is preserved and accessible long into the future.
- 10,557,655 total volumes digitized: 5,556,767 book titles and 274,642 serial titles
- 31% of total in the public domain
- Complutense joined HT in 2010. The only non-American partner
- Partners: Library of Congress, New York Public Library, California Digital Library and some academic libraries: Columbia, Cornell, Harvard, MIT, Princeton, Stanford, California, Chicago, Michigan, Yale...
How do you access the Complutense digitized books?
5. In your own catalogue if you have a discovery tool as Summon (only one click)
How do you access the Complutense digitized books?

6. More: Internet Archive, Europeana...
Europeana Libraries Project

- Complutense collaborates with 18 research libraries from 14 countries: Bavarian State Library, Oxford University, Welcome Library, University College London, Ghent, Trinity College, etc.

- The objective is to incorporate 5 million digital objects: manuscripts, films and texts belonging to bibliographic and scientific heritage of the participating libraries to Europeana.
### Scanning process total data

<table>
<thead>
<tr>
<th>Checked Books</th>
<th>Scanned Books</th>
<th>Not scanned Books</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>143,000</td>
<td>120,000</td>
<td>23,000</td>
<td>17%</td>
</tr>
</tbody>
</table>

![Pie chart showing 83% of books scanned and 17% not scanned]
### Access to Complutense Books in Google

**Most visited books (one week)**

- Every week more than 60% of the scanned books are visited
- The most visited book of the Google European partners is from Complutense University.

<table>
<thead>
<tr>
<th>Accesos</th>
<th>Título</th>
<th>Autor</th>
<th>Año</th>
<th>Centro</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.490</td>
<td>Diccionario etimológico de la lengua castellana (ensayo)</td>
<td>Pedro Felipe Monlau</td>
<td>1856</td>
<td>FLL-DER</td>
</tr>
<tr>
<td>12.008</td>
<td>Diccionario geográfico-estadístico de España y sus posesiones de ultramar</td>
<td>Pascual Madoz</td>
<td>1830</td>
<td>VET</td>
</tr>
<tr>
<td>8.637</td>
<td>La iliada</td>
<td>Homero</td>
<td>1788</td>
<td>FOA</td>
</tr>
<tr>
<td>8.275</td>
<td>Vida y viajes de Cristóbal Colón</td>
<td>Washington Irving</td>
<td>1852</td>
<td>GHI</td>
</tr>
<tr>
<td>7.520</td>
<td>Enciclopedia moderna</td>
<td>Francisco de Paula Mellado</td>
<td>1851</td>
<td>DER</td>
</tr>
<tr>
<td>6.468</td>
<td>Diccionario de la lengua castellana</td>
<td>Real Academia Española</td>
<td>1852</td>
<td>FLL</td>
</tr>
<tr>
<td>4.450</td>
<td>Mitología universal</td>
<td>Juan Bautista Carrasco</td>
<td>1864</td>
<td>DER</td>
</tr>
<tr>
<td>4.205</td>
<td>Linajes nobles de España</td>
<td>Juan José Vilar Psayla</td>
<td>1867</td>
<td>FLL</td>
</tr>
<tr>
<td>4.179</td>
<td>Diccionario de agricultura práctica y economía rural</td>
<td>Agustín Esteban Collantes, Agustín Alfaro</td>
<td>1855</td>
<td>MED</td>
</tr>
<tr>
<td>4.035</td>
<td>Anatomie descriptive</td>
<td>Jean Cruveilhier</td>
<td>1837</td>
<td>MED</td>
</tr>
<tr>
<td>3.984</td>
<td>Anatomia do corpo humano</td>
<td>Bernardo Santucci</td>
<td>1739</td>
<td>FOA</td>
</tr>
<tr>
<td>3.671</td>
<td>Diccionario universal latino-español</td>
<td>Manuel de Valbuena</td>
<td>1808</td>
<td>FOA</td>
</tr>
</tbody>
</table>
How do we preserve our digitized books? Hathi Trust

- Hathi Trust: Long term preservation (and dissemination).

- For us, digital preservation could only be achieved as a result of cooperative involvement with other academic institutions following the standards of the international library community.

- What’s Hathi Trust?
  - A repository for storing high quality
  - A scalable technological and organizational potential
  - A portal to access scanned books and journals
HathiTrust Characteristics:

- Bibliographic and full-text search.
- Shibboleth authentication system.
- Bibliographic metadata are managed in a Library Management System (Aleph).
- Access to disabled users
- Offers access to bibliographic data via API for your catalog
- Bibliographic data (and access to scanned books) is included in discovery tools as *Summon*.
- You can download the books in the public domain (pdf, epub).
- Additionally, you can make collections, make them public and share with others.
Characteristics of HathiTrust

• Content Preservation: Image and text representation (if possible)
• Open code technologies: PERL, Linux, MySQL
• International Standards:
  • Trustworthy Repositories Audit & Certification (TRAC)
  • Open Archival Information System (OAIS) Reference Model
  • Preservation Metadata Implementation Strategies (PREMIS)
• Image formats: TIFF, JPEG 2000
• Permanent URL
HathiTrust Digital Library Collection

- Michigan: 50%
- California: 28%
- Wisconsin: 4%
- Yale: 1%
- Columbia: 1%
- Harvard: 1%
- Chicago: 1%
- Penn State: 0%
- Indiana: 2%
- Princeton: 3%
- Cornell: 3%

Source: Jeremy York (data as of May 1, 2011)
Conclusions: Objectives achieved in the project

• Scanning an important number of our ancient books (83%)
• Increasing the use of the collection by the general public.
• Supporting researchers offering digitized materials for text analysis.
• Increasing the visibility and long term preservation of our collections.
• Bringing on the library catalog all the books before the twentieth century (much of them in full text).
• Knowing the exact preservation condition of each book
• Establishing a plan of conservation and restoration of damaged books.
Thank you for your attention!

José Antonio Magán Wals  
Antonio Moreno Cañizares  
Manuela Palafox Parejo

Complutense University of Madrid Library