LIBER QUARTERLY 21 (1), November 2011 — ISSN: 1435-5205. P48–68
http://liber.library.uu.nl/
Igitur publishing
This work is licensed under a Creative Commons Attribution 3.0 Unported License

Mass Digitization at the Complutense University Library: Access to and Preservation of its Cultural Heritage

José A. Magán

Director of the UCM Library

Manuela Palafox

Head of Digital Publishing and Digitization Projects at UCM Library,
mpalafox@buc.ucm.es

Eugenio Tardón

Vice-Director of the UCM Library

Amelia Sanz

Assistant Professor of French Literature at UCM

Abstract

In the mid-1990s, the Library of the Complutense of Madrid (UCM) applied the growing body of new technologies to a pioneering project: the digitization of a valuable collection of biomedical books. The result of the project was the Dioscorides Digital Library. Up to 2006, however, the pace of digitization was very slow, resulting in only 3,000 books and 50,000 digitized engravings.

In September 2006, the Complutense University and Google signed a partnership agreement for the purpose of carrying out large-scale digitization of our public domain collections. The Complutense University Library is participating in this mass-digitization project with the following goals:

- To improve the discovery of the cultural heritage of Complutense and increase the number of potential readers;
- To fulfil its public service mission;
- To preserve and protect the original books;
- To enhance students' and faculty research.

By early 2011, the Complutense Library had digitized roughly 120,000 out-of-copyright books from the 16th to the 19th centuries. All these materials are easily accessible via the Complutense Library Catalogue, Google Books and HathiTrust Digital Library. The Complutense University of Madrid joined the HathiTrust Digital Library in November 2010, which allowed books digitized by Google to be stored on HathiTrust servers. The HathiTrust Digital Library is both a digital preservation repository and a platform for access. It provides long-term preservation and access services for public domain and in-copyright content from a variety of sources, including Google, the Internet Archive, Microsoft, and in-house partner institution initiatives. In early 2011, the number of HathiTrust public domain volumes reached the two million mark and the collection exceeded eight million volumes.

In addition, and complementing the above, the digital content of the Complutense University Library will be aggregated to Europeana during 2011 and 2012.

Key Words: Digital preservation; HathiTrust Digital Library; Google Books; Europeana Libraries

1. Introduction

For more than two decades, academic and research libraries have been undergoing continuous transformation as a reaction to major changes inside and outside their environment. Our strategic lines of action are determined by three ideas:

- The life of libraries can no longer be conceived in pre-digital terms. We are experiencing the final transition from the analogue to the digital era (we only have to think of e-journals, e-books, repositories, etc.).
- There is a relationship between the digital content and the digital tools used in the creation and management of the former since both share the same characteristics.
- Increasing accessibility and use of information are the result of bringing both the content and the tools in line with each other in order to

> break down barriers to information flows related to learning, teaching and research activities. Any remaining barriers are essentially economic or political rather than technical.

In this globalized world universal access to culture and science is now considered a right. Libraries, as conspicuous custodians of cultural heritage, should be fully involved in the systems that are available for improving the dissemination of knowledge.

Therefore, like other libraries, the Library of the Complutense University of Madrid (henceforth referred to as UCM) has had to redefine its strategy for disseminating knowledge and research. Firstly, in order to serve the UCM academic community, the Library has made it easier to access the most recent outcomes from international R&D activities; secondly, to serve the global academic community and the general public, it has promoted the dissemination of historically accumulated research conducted by UCM scholars, the ultimate goal being that the cultural and scientific resources of UCM should be universally accessible and properly preserved.

In order to carry out this strategy, we have developed three lines of action within a framework that emphasizes international cooperation as the best way to ensure the survival of academic and research libraries in the future. The first line of action, which is the specific subject matter of this paper, is building collections through digitization and reformulation of the preservation policy of digital and printed documents. As Conway notes, it is necessary to distinguish between *digitization for preservation* (creating new digital products from physical artifacts) and *digital preservation* (protecting the value of the products created, irrespective of whether the original sources are tangible artifacts or born-digital data) (Conway, 2010). Each of these tasks has different standards, processes, technologies, costs and organizational challenges. The second line is related to the organization of knowledge and is aimed at enhancing access, while the third line furthers the development of mechanisms that enable the delivery and use of such knowledge.

In order to create new digital collections that can spread the Spanish cultural and scientific heritage hosted in the UCM Library, a partnership agreement was signed with Google in 2006 for digitizing our public domain collections. The amount of human and economic resources required for preserving a digital copy of more than 120,000 UCM books digitized by Google was so

large that the viability of the Library was jeopardized. In our view, only international cooperation can enable academic and research libraries to meet such a challenge. Therefore, UCM decided to join the HathiTrust Digital Library in November 2010.

In addition, and reinforcing this strategy, UCM joined the new 'Europeana Libraries'[1] project, funded by the European Commission, in January of 2011. This project complies with the provisions laid down by the European Union in its European Digital Agenda (European Commission, 2010) and with recommendations by the Comité des Sages on the Digitization of European Cultural Heritage (European Commission, Comité des Sages, 2011), which both draw attention to the urgent need for all materials digitized with public funding to be aggregated to Europeana by 2016.

In a move to improve the organization of, access to and use of our digitized collections by the UCM community, two Complutense research teams have combined their efforts to work on a method that will allow collaborative annotations of literary texts and the collaborative creation of annotation schemas. This project, entitled 'Collaborative Annotation of Digitalized Literary Texts', is funded with one of the 12 grants awarded by Google in its 2010 Digital Humanities Research Awards Program.[2]

These innovative projects have been actively supported by our academic authorities (President, Vice-Presidents, Deans, etc.) and have been welcomed by the UCM community.

2. Digitization of the Cultural Heritage of UCM

2.1. Previous Experiences

In the mid-1990s, the UCM Library created the Dioscorides Digital Library with the aim of digitizing and providing access to a selection of works from its historical collection of biomedical books. By 2006, approximately 3,000 books and 50,000 engravings from the fifteenth to the nineteenth centuries, including manuscripts and printed books, had been digitized.

After the results of this project had been assessed, three conclusions were drawn: a) the pace of digitization had been slow, which made it very difficult

to extend the project to include all of UCM's cultural heritage within a reasonable period of time, b) budgetary constraints limited any expansion of the project, and c) dissemination on the Internet was poor, owing to the local nature of the project. It was therefore necessary to find an alternative digitization model and migrate from the 'boutique model' to the large-scale model that by then was already being implemented by Google through the Book Search project.

2.2. The UCM and Google Partnership Agreement

In September 2006, UCM and Google signed a cooperation agreement with the aim of digitizing public domain collections in its Library and offering free online access. The University contributed its holdings and the staff needed to select and handle the documents to be digitized, while Google took care of the costs of digitization. Two copies of the digitized material were produced, one for UCM and the other for Google. The specific goals of the UCM Library were:

- To promote access to and dissemination and preservation of the cultural heritage of UCM in the public domain.
- To add the UCM holdings to those of major libraries already participating in the Google project (Michigan University, Stanford University, Harvard University, Oxford University and New York Public Library) in order to facilitate access to scholarly knowledge by the global academic community and the general public.
- To provide a cutting-edge information system for accessing the cultural heritage of the Library.
- To design a plan for the conservation and restoration of damaged books.
- To provide UCM scholars, especially those ones working in the fields of social sciences and humanities, with a corpus of digital materials that enable digital projects to be developed.

2.2.1. Digitization project planning

In accordance with the agreement with Google, the UCM library established a working group to draw up the digitization plan, monitor its progress and prepare reports on its development. The group undertook the following activities:

a) Analysis of the collections and storage spaces. Firstly, the collections and the spaces where they were stored were examined, with the following objectives:
   - To ascertain the number of public domain volumes to be digitized and their distribution over the 32 locations of the UCM Library.
   - To obtain information on the facilities and accessibility in the above-mentioned locations and any potential difficulties that may arise.
   - To carry out a survey of the collections, including the organization system, conservation conditions and number of non-catalogued books.
   - To specify, before and after selecting the material, the tasks required for completion of the project, as well as the staff needed for this purpose.

b) Cataloguing Plan. In December 2006, a cataloguing plan was drawn up for non-catalogued materials dating from the 16th to the 19th centuries. In total, approximately 220,000 volumes were catalogued.

c) Selection Guidelines. The selection of volumes was based on three main variables: date of publication (from the 16th century to 1870), physical condition of the item (dimensions, text block and binding conditions) and fitness for scanning.

d) Binding of 19th-century books. A sample of data revealed that binding issues could be a serious obstacle to digitization. To overcome this, the restoration department of the UCM Library drew up a list of technical recommendations for the proper binding of 19th century books. As a result, it was possible to digitize many previously discarded volumes and improve their conservation.

2.2.2. Implementation of the digitization project

In mid-2007 the workflows and logistic operations required for the digitization process were established for all UCM Library locations. In order to control the project workflows, the Library developed several applications for logistics management in cooperation with the university's IT services.

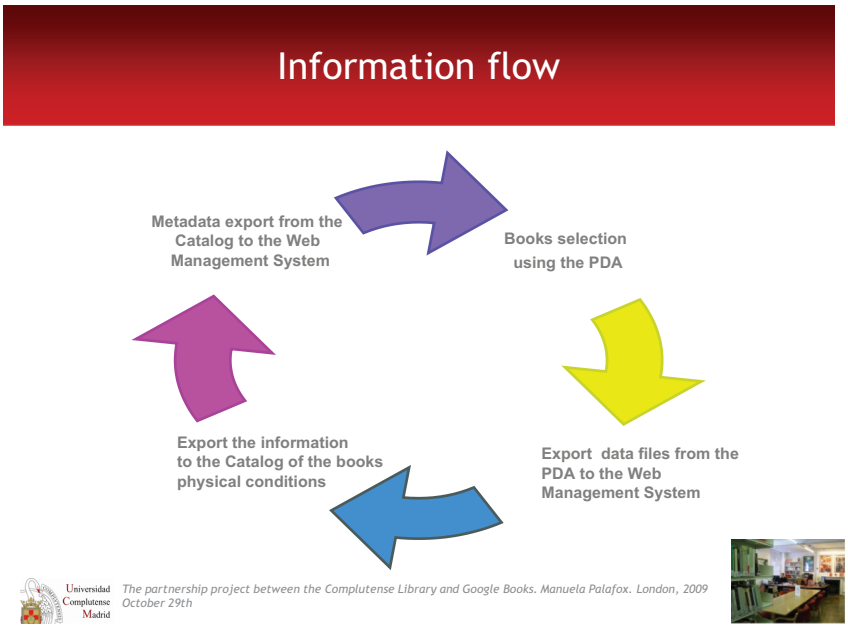
- A web application providing real-time information on all the processes involved in scanning, with concrete data and statistics. This application stores book metadata in the catalogue included in the digitization programme and also offers a complete overview of the

daily movements of books in each location and books selected to be sent to Google. At the same time, it shows automatically updated information on the availability or status of the item in the UCM Cisne catalogue.

- An application used directly through a PDA by selection teams in the stacks. When the application reads the book's barcode, a form is displayed on the touch screen for the librarian to evaluate features relating to the physical conditions of the book. This information is then dispatched to the item record in the web application, where the dimensions of the books are registered, together with aspects of their state of conservation (presence of fungi, degree of physical deterioration, loose or brittle pages, whether the binding is valuable, weak or absent, etc.). A summary of this information is automatically forwarded to the Cisne catalogue records in a note field (Figure 1).

This process has provided the Library with very comprehensive information on the conservation of its historic collection: some 13% of the pre-1871

Fig. 1: Information flow.



volumes (about 19,000 out of 140,000) have issues that exclude them from the digitization project. The main issues are related to text block (2% brittle pages, 10% fungi, 20% loose pages, 21% uncut pages and the remaining 47% other physical damages) (Figure 2) and binding (5% non-open, 10% lost, 25% weak, 30% requiring rebinding, and 30% very much damaged) (Figure 3). Knowing these figures and being able to identify each item with conservation issues allows us to devise a restoration plan including the economic cost. However, budgetary constraints have made it impossible to implement such a plan at the present time.

Fig. 2: Text block problems.

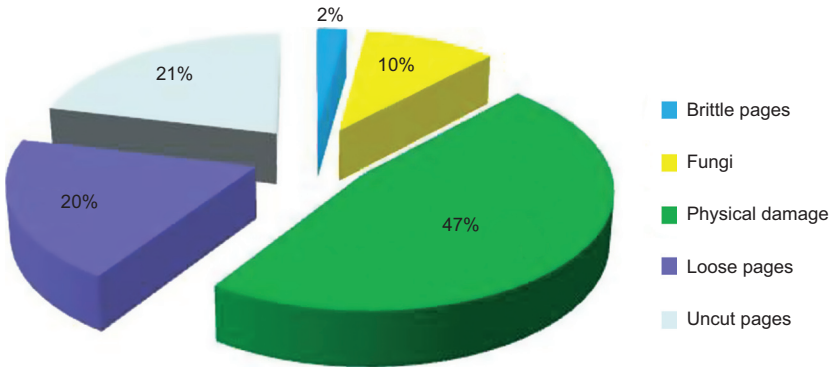
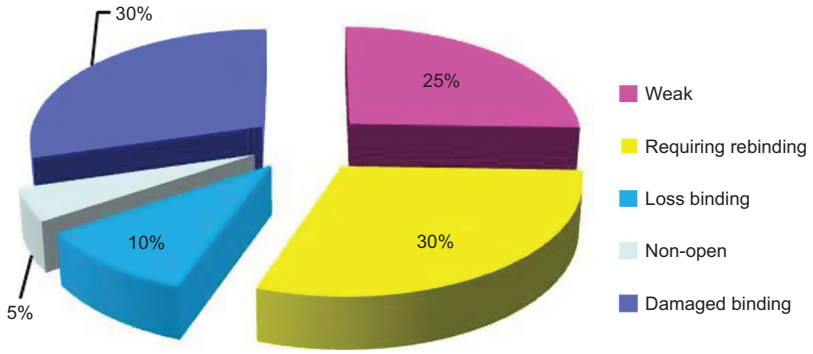


Fig. 3: Binding problems.



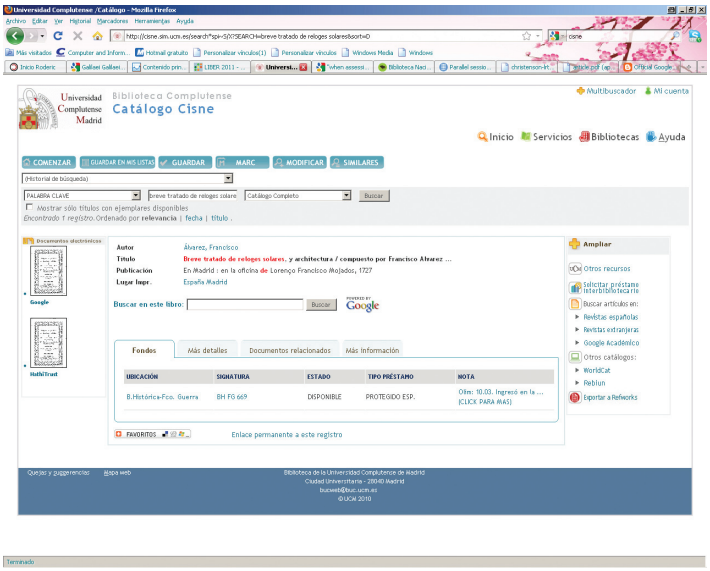
2.2.3. Dissemination and access via the Cisne catalogue

Digitized books are available through three websites: the Cisne catalogue of the UCM Library, Google Books and HathiTrust. The Cisne catalogue has links to the digital volume and also a box for full-text searching. The web application for logistical management produces a daily script for accessing GRIN, the Google partners management interface, obtains the item identifiers (barcodes) with the full text from Google Books and adds an 856 MARC label to the Cisne record, thus providing a link to the digital copy in Google Books. Cisne also runs another script that uses a Google API[3] to check whether there is a digital copy of a book in Google Books; if this is the case, the Cisne catalogue will show a search box. Another link will shortly be added to the UCM digital copy stored in HathiTrust, as shown in Figure 4.

3. Preservation of the UCM Digital Content

One of the main motivations behind the UCM agreement with Google was to ensure that our public domain collections would remain accessible to future

Fig. 4: Cisne Catalogue, access to full text books (Google and HathiTrust).



generations. If users access digital copies instead of the original materials, then the handling of original books is kept to a minimum and the preservation function comes into play.

According to the UCM Management Collections Policy,[4] 'the UCM Library policy of preservation of digital resources is based on activities and interventions required to ensure long-term accessibility and legibility of reliable digital objects required by the University for the purposes of learning, teaching, researching and other activities related to institutional objectives.'

With respect to digitization specifications, the preservation formats for images from in-house digitization projects in the UCM Library are TIFF and JPEG. The Library uses PREMIS[5] structured according to the METS[6] (Metadata Encoding & Transmission) scheme, in accordance with recommendations from the Spanish Department of Culture.[7]

Throughout the year 2010, our IT services and Library staff were engaged in the joint exploration of a common infrastructure strategy, aimed at significantly enhancing our capability for preserving and accessing the UCM cultural heritage digitized by Google. Two needs were identified for implementing a digital preservation programme: 1) a large-scale digital content repository, and 2) a scaleable technological and organizational potential. We concluded that these requirements were so important that digital preservation could only be achieved through the cooperative involvement of academic institutions following international library community standards. As a result, in November 2010 the UCM joined HathiTrust: 'Because scholarly needs can often be more readily addressed by organizations and initiatives that are focused on academia, the creation of HathiTrust and its relationship with Google is especially promising. In essence, Hathi will bring an enormous collection of digitized books, including many of those scanned through Google partnerships, under the control of an enterprise principally driven by a scholarly mission. Hathi's digital preservation mission is vital, and its partnership model allows participating libraries to achieve scale and quality far greater than that which any of them could do on their own' (Schonfeld, 2010).

3.1. The HathiTrust Digital Library

The HathiTrust digital repository was launched in 2008 by a partnership of major US research libraries whose aim was to preserve and provide access to digitized

books and journals through collaboration. Currently, more than fifty libraries who are partners in HathiTrust have digitized their collections in mass digitization projects, including digitization by Google and Internet Archive, as well as in-house initiatives. 'These libraries have a common vision to build a new kind of library for the 21st century — a cooperative library founded on principles of commitment, deep resource sharing, and trust, that was so comprehensive in its representation of the published record, so available to users anywhere in the world, and so broad in its impact on the fundamental business of what libraries do, that it could rise to be called a universal library' (York, 2010).

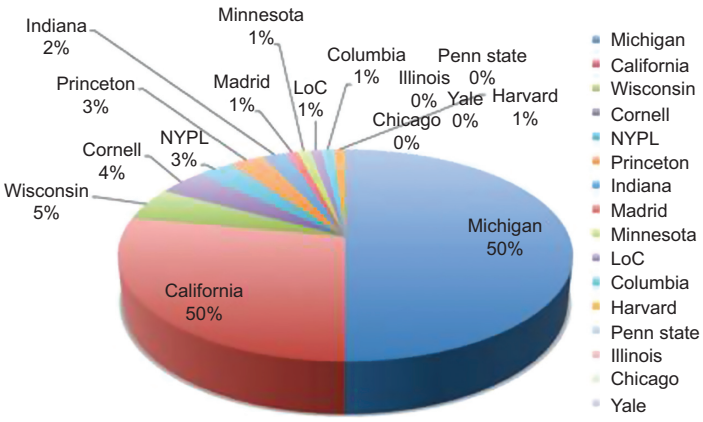
HathiTrust is committed to bit-level preservation and format migration of materials as technology, standards, and best practices in the digital library community change. HathiTrust strives to ensure that the digital content it preserves is accurate, complete, suitable for long-term preservation, and useful for a variety of access purposes. It does this through consideration of content file formats, preservation and descriptive metadata, validation routines, and attention to quality.[8]

HathiTrust serves a dual role. First, as a trusted repository it guarantees the long-term preservation of the materials it holds, providing the expert curation and consistent access long associated with research libraries. Second, as a service for partners and a public good, HathiTrust offers persistent access to the digital collections. This includes viewing, downloading, and searching access to public domain volumes, and searching to in-copyright volumes. Specialized features are also available which facilitate access by persons with print disabilities, and allow users to gather subsets of the digital library into 'collections' that can be searched and browsed.[9]

As of June 8, 2011 HathiTrust partners had contributed 8,800,666 volumes, of which 4,800,890 were book titles and 213,244 serial titles. The number of volumes in the public domain was 2,400,247 (~27% of the total).[10] HathiTrust is investigating issues relating to the storage and delivery of electronic publications (in the ePub format in particular) and digital audio and image files (Beers *et al*., 2010).

As shown in Figure 5 (York, 2011), over 75% of the content is contributed by the collections of Michigan and California Universities, which have digitized both out-of and in-copyright collections. The Complutense University of Madrid owns 1% of the total content.

Fig. 5: HathiTrust content (data as of May 1, 2011).



The repository was designed in accordance with the Open Archival Information Systems (OAIS) reference model.[11] Preservation information is recorded using PREMIS, and is compliant within the context of community-wide standards and criteria for Trustworthy Digital Repositories Audit & Certification (TRAC).[12] The repository utilizes a robust technology and has the geographic redundancy of two mirror sites at the University of Michigan and Indiana University.

3.1.1. Ingest of bibliographic metadata and content

Two components are required for ingest in HathiTrust: bibliographic metadata and content. Accurate bibliographic records for submitted content are provided by the institutions before content ingestion. The records act as a manifest of the digital content and are used as part of HathiTrust's digital object management strategy. The UCM Library delivers the bibliographic metadata via download from a source. Specifications for bibliographic metadata are:

- Data in MARCXML format in utf-8 encoding
- One bibliographic record per item (multi-volume works should have the same record repeated for each item)
- Complutense system number in 001 field
- Barcode in 955 | b
- Item description (enumeration/chronology) in 955 | v.

The namespace identifier is determined by the institution. HathiTrust distinguishes content in the storage system by creating a new namespace for each institution, and each body of content within an institution has a unique identifier scheme. The namespace is four characters long at most and is followed by a unique (in that namespace) identifier. This is often, but not always, the physical item barcode. The two together (namespace plus identifier) comprise an object's repository identifier. For example: ucm.5325109853.

HathiTrust uses the Handle system () to assign permanent URL's to repository objects. For example: http://hdl.handle.net/2027/ucm.5325109853.

The UCM Library uses a hybrid format for scanned volumes containing bitonal TIFF files for all-text pages and JPEG2000 files for images. UCM ingests the digital content digitized by Google and other materials digitized in-house, such as incunabula and manuscript collections. Ingest started in December, 2010 (Table 1).

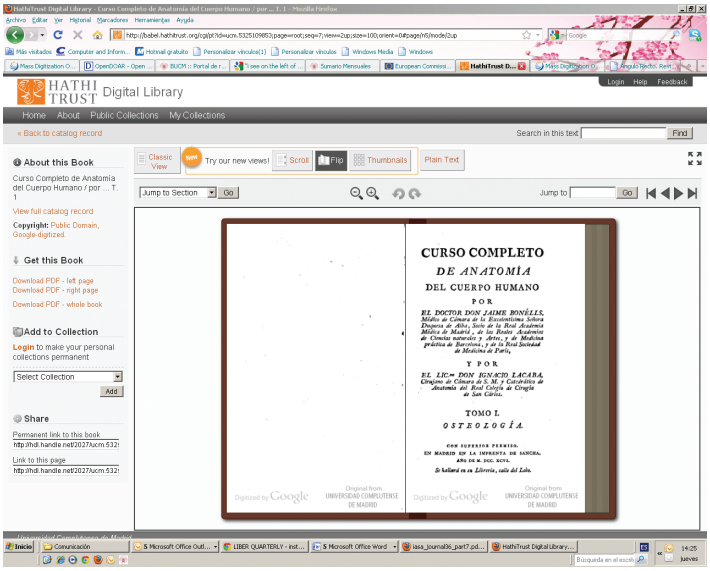
Table 1: UCM Ingest into HathiTrust

Date	Volumes	Total
December 2010	78,256	78,256
January 2011	1,156	79,412
February 2011	6,125	85,537
March 2011	2,774	88,311
April 2011	15,486	103,797
May 2011	2,947	106,744

3.1.2. Access services

Bibliographic and full text searching is available for all items in HathiTrust. Digital volumes are accessible by various means depending on copyright status. The bibliographic search uses an aggregation of records contributed by partner libraries and is thus based on rich descriptive metadata that is the output of decades of library cataloguing (Christenson, 2011, p. 98). Volumes in the public domain are freely accessible to all users, but in order to download the whole book, authentication by persons affiliated with partner institutions is needed. All public domain volumes can be viewed on the web in a page-turner application (Figure 6).

Fig. 6: HathiTrust Digital Library.



In addition, HathiTrust offers a collection builder tool. 'The Collection Builder has great potential for integration within local services, such as online courses and themed collection portals built by local institutions. Once a collection is created, the full text of those volumes can be searched as a set' (Christenson, 2011, p. 99).

3.1.3. Supporting research

The mission of HathiTrust is to contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge.[13] 'The presence of a critical mass of research institutions in the HathiTrust partnership enables an aggregation of digital resources not seen before, hosted by libraries for the long term in a continuation of their traditional role as stewards of the scholarly record and supporters of research and other scholarly pursuits' (Christenson, 2011, p. 95).

HathiTrust uses multiple strategies to support data mining. It plans to create a Research Centre to perform computational research and is developing tools

to make the texts of public domain works in its corpus available for research purposes.[14] These fall into two categories: non-Google-digitized volumes and Google-digitized volumes:

- Non-Google-digitized volumes: There are no restrictions on the availability or use of texts for non-Google-digitized public domain volumes. As of February 1, 2011, there are approximately 120,000 public domain volumes primarily, though not exclusively, English language materials published prior to 1923.
- Google-digitized volumes: There are approximately two million public domain volumes as of February 2011, representing a wide variety of languages, subjects, and dates. In general, limits on the use of these materials are as follows:

    - They can only be used for scholarly research purposes.
    - They may not be used commercially.
    - They may not be re-hosted or used to support publicly available search services.
    - They may not be shared with third parties.

As can be seen in Figure 7, which shows the number of visits to the UCM book pages from December, 2010 to June, 2011, considerable use is being made of UCM books.

3.2. Europeana Libraries

UCM is also collaborating on the Europeana Libraries project which includes major library networks (CERL, CENL, LIBER) and 19 leading European research libraries from: Austria, Belgium, Estonia, Germany, Hungary, Ireland, Romania, Serbia, Spain, Sweden, Switzerland and the UK. 'The aim of the "Europeana Libraries" project is to both significantly increase the amount of high-quality digital content from the libraries of Europe to aggregate to Europeana and to produce a sustainable model for it to continue as a major aggregator for Europeana into the future' (Chambers and Schallier, 2010) (Figure 8).

The bedrock of the Europeana Libraries project consists of 4 associations:

- Conference of European National Librarians (CENL)
- Consortium of European Research Libraries (CERL)

Fig. 7: Visits to the UCM book pages in HathiTrust Digital Library.

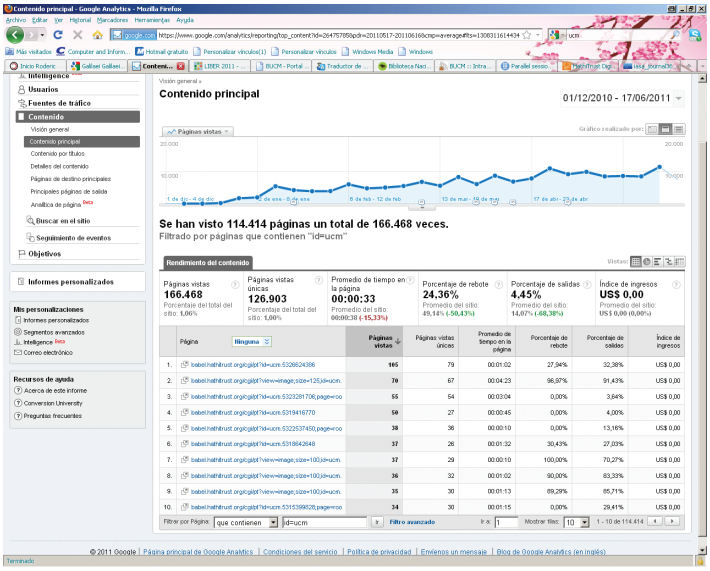


Fig. 8: Europeana Libraries.

- Europeana Foundation
- Ligue des Bibliothèques Européennes de Recherche (LIBER)

The aims of Europeana Libraries are:

- To create a valuable resource for scholars, with full-text search capabilities for written works.
- To build a robust network of national, university and research libraries.
- To establish an efficient aggregation model that can be used by research libraries across Europe.[15]

The Europeana Libraries project was launched in January 2011 and will last for 24 months. Five million digitized objects will be aggregated to Europeana. The project is coordinated by The European Library and hosted by the National Library of the Netherlands. The scientific coordinator is the University College London (UCL) and aggregation tools are being developing by the Instituto Superior Técnico. The project is been managed by MDR Partners. There are 19 content providers from major European universities and national libraries; some of these will aggregate metadata for books digitized by Google at UCM, the University of Ghent, the University of Oxford, and the Bavarian State Library.

The UCM Library will provide the following collections:

- Complutense rare books from the 16th Century to 1870: metadata of more of 120,000 books digitized by Google.
- Engravings from the Dioscorides Collection: approximately 50,000 engravings.
- Complutense university theses: 6,000 theses.
- Journal articles published by the 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: approximately 3,200 photographs.
- The Personal Archive of Ruben Dario: 5,000 documents.
- Complutense manuscripts.
- Complutense incunabula: 362 incunabula.
- Cartographic material, maps and city views: 300 items.

Europeana is currently based on the 'Europeana Semantic Elements' (ESE) metadata format, but has begun to implement the new 'Europeana Data Model'. This is 'a new proposal for structuring the data that Europeana will be ingesting, managing and publishing. It will help enable users to browse the content in Europeana in new ways and will facilitate Europeana's participation in the semantic web' (Chambers and Schallier, 2010, p. 116).

4. Collaborative Annotation of Digitized Literary Texts

Two of Complutense's research teams, the LEETHI Group (European Literatures: From Text to Hypermedia), coordinated by Professor Amelia Sanz, and the ILSA Group (Language Software Engineering and Applications) coordinated by Professor José Luis Sierra, have combined their efforts to work on a method that allows collaborative annotations of literary texts and the collaborative creation of annotation schemas. The teams point out that texts that are only digitized cannot be presented to student-readers as they are. The teams argue that reliable, reader-friendly texts are needed at four levels:

- Full documentation on the volume.
- Perfect reproduction in image mode and perfect reproduction in text mode.
- Presentation of an enriched text thanks to inserted notes and the wording of questions.
- Educational publishing to provide inexperienced readers with easier access to and reading of digitized texts.

The teams envision an annotation framework allowing communities of scholars to annotate the corpus of literary texts with evolving annotation schemas: instead of relying on pre-established categories of annotations, scholars can create and re-organize types of annotations according to their particular expressive needs. Moreover, these annotations can be classified using key words that are in turn identified as instances of concepts in an ontology. As a metadata schema, this ontology can also be created and maintained by scholars in order to tailor it to their specific needs. The annotation framework is currently being implemented in a system integrated with Google Books. The system will be tested with other digitalized collections at UCM, such as the Ruben Dario collection.[16]

The system will include the following functions:

- The recovery of digitized texts from the web and annotation of these texts using the shared annotation schema.
- The addition of new terms and relationships to the annotation schema as required during the annotation process; these terms will be immediately available for all researchers in the community.
- Search and navigation of the annotations, and therefore the collections of digitized texts, using the schema structures (terms and relations).
- The possibility for privileged users to edit and restructure the schema and perform administrative tasks (e.g. community creation, privileges assignment, etc.).
- Finally, the annotation system will contribute to introducing young students to electronic textualities at academic level, because the tool will be integrated into the learning process through any of the Learning Management Systems of our Virtual Campus at the Complutense University of Madrid.[17]

5. Conclusions

The large-scale digitization agreement with Google for scanning the cultural heritage of UCM in the public domain means that the Complutense Library can successfully face the new challenges that the globalization of knowledge, science and the economy will pose to research libraries.

We have created a new digital collection of over 120,000 volumes. This has enabled us to improve the preservation of original and printed sources and, at the same time, to integrate these new digital volumes into the HathiTrust Digital Library, the most important and innovative digital preservation project for the library community.

In this field, similar innovative initiatives are emerging in Europe, as is the case of the Europeana Libraries project, in which UCM is participating. We have taken advantage of previous efforts in digitization with a dual goal: to increase the dissemination of our digitized heritage and to participate in an innovative research area, related to the semantic web and linked data.

Finally, the newly created digital collections have strengthened the relationships between the Library, IT Services and the UCM Faculty, as reflected in the field of digital humanities, where UCM researchers are working on a system for\ Collaborative Annotation of Digitized Literary Texts.

References

Beers, Shane, Jeremy York and Andrew Mardesich (2010): 'Adding new content types to a large-scale shared digital repository', *Austrian Computer Society* (OCG). http://hdl.handle.net/2027.42/83275.

Chambers, Sally and Wouter Schallier (2010): 'Bringing Research Libraries into Europeana: Establishing a Library-Domain Aggregator', *Liber Quarterly*, 20/1, available from http://liber.library.uu.nl/.

Christenson, Heather (2011): HathiTrust. 'A Research Library at Web Scale', *Library Resources & Technical Services*, 55(2), April 2011.

Conway, Paul (2010): 'Preservation in the age of Google: digitization, digital preservation, and dilemmas', *Library Quaterly*, 80(1).

European Commission. 'COM, (2010): 245. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: a Digital Agenda for Europe.' http://ec.europa.eu/information_society/digital-agenda/documents/digital-agenda-communication-en.pdf.

European Commission, Comité des Sages (2011): ' The New Renaissance: Report of the 'Comité des Sages' Reflection group on bringing Europe's Cultural Heritage online, Brussels, http://ec.europa.eu/information_society/activities/digital_libraries/doc/refgroup/final_report_cds.pdf.

Rieger, Oya Y. (2008): 'Preservation in the Age of Large-Scale Digitization. A White Paper', Washington, Council on Library and Information Resources.

Schonfeld, Roger C. (2010): Conclusion. In: *Library of Congress Cataloging-in-Publication Data. The Idea of Order: Transforming Research Collections for 21st Century Scholarship*, June, http://www.clir.org/pubs/reports/pub147/pub147.pdf.

Wilkin, John (2011): 'Bibliographic Indeterminacy and the Scale of Problems and Opportunities of "Rights" in Digital Collection Building', CLIR and DLF 'Ruminations' series, February, http://www.clir.org/pubs/ruminations/01wilkin/wilkin.html.

York, Jeremy (2010): 'Building A Future By Preserving Our Past: The Preservation Infrastructure of HathiTrust Digital Library', *World Library and Information Congress: 76th IFLA General Conference and Assembly*. 10–15 August 2010, Gothenburg, Sweden. http://www.ifla.org/files/hq/papers/ifla76/157-york-en.pdf.

York, Jeremy (2011): HathiTrust Open Webinar (slides), http://www.hathitrust.org/documents/HathiTrust-OpenWebinar-201105.ppt.

Notes

[1] http://www.europeana-libraries.eu/

[2] http://googleresearch.blogspot.com/2010/07/our-commitment-to-digital-humanities.html

[3] http://code.google.com/intl/es-ES/apis/books/docs/viewer/developers_guide.html

[4] Política de gestión de las colecciones de la biblioteca de la Universidad Complutense de Madrid (junio de 2009). http://www.ucm.es/BUCM/intranet/30336.php

[5] PREMIS (Preservation Metadata Implementation Strategies). http://www.loc.gov/standards/mets/

[6] METS (Metadata Encoding & Transmission) http://www.loc.gov/standards/mets/

[7] http://travesia.mcu.es/documentos/pautas_digitalizacion.pdf

[8] HathiTrust Preservation. http://www.hathitrust.org/preservation

[9] Complutense University of Madrid joins to HathiTrust, http://www.ucm.es/BUCM/servicios/doc16172.pdf

[10] HathiTrust: http://www.hathitrust.org/

[11] Reference Model for an Open Archival Information System (OAIS); CCSDS 650.0-B-1; Consultative Committee for Space Data Systems: Washington, DC, 2002.

[12] TRAC (Trustworthy Repositories Audit & Certification): Criteria and Checklist. Center for Research Libraries and OCLC, 2007. http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

[13] HathiTrust, http://www.hathitrust.org/mission_goals

[14] HathiTrust, http://www.hathitrust.org/datasets

[15] Europeana Libraries website, http://www.version1.europeana.eu/web/europeana-libraries

[16] The latest release of this annotation system is available at http://fdiejemplolector.appspot.com

[17] https://www.ucm.es/campusvirtual/CVUCM/index1.php