Sapienza Libraries and Google Books Project

Adriana Magarotto, Maura Quaquarelli, and Mattia Vallania

Sistema Bibliotecario Sapienza, Sapienza - Università di Roma, Italia
{adriana.magarotto,maura.quaquarelli,
mattia.vallania}@uniroma1.it

Abstract. The report shortly examines the experience of Sapienza libraries as partners of Google Books project, signed in July 2011. The goal is to digitize 35,000 books from 1500 to 1872 during the first year of activity. The issue concerns management and the optimization of bibliographic data set, development of web-based instruments for ruling the workflow and sharing records and information between the ILS system (Sebina Open Library) and external data bases.

Keywords: libraries, digitization, organization, catalogues, bibliographic data.

1 Introduction and Contest

A research library's mission is to set up document collections to satisfy user needs, so it's necessary to ensure access to these collections and to make it possible to spread historical memory and knowledge kept in these libraries. Recent digitization projects, many economical and space problems for the printed collections' growing have given new opportunities and changed the concept of research library. To pursue this transformation, a process going on in libraries worldwide, Sapienza decided to take part in the Google Books project, a precious chance to jumpstart complete digitization. Previously other similar projects have been realized by single departments or in partnerships, such as ProDigi (2008-2009). These projects have made it possible to create an archive of digitized texts, and helped spread knowledge and best practices which we can now use for this brand new project, a complex effort to create and increase Sapienza Digital Library.

After the agreement between Sapienza University of Rome and the Ministry of Cultural Heritage-MiBAC, signed by the University Rector in July 2011, Sapienza was willing to give a large part of its own book collections for the digitization. Now the project is running and many books have been sent since November 2011. We think that we can digitize about 35,000 books during the first year. We start with ancient printed books, from 1500 to 1700 up to but not over1872, a conventional date established by international copyright laws. Only 10 libraries have taken part in the first step of the project, because of time constraints and the experimental nature of the project. The Google contract expects maximum privacy on their technical solutions, so the following description briefly shows problems, solutions and results of the first part, only for SBS-Sapienza.

T. Catarci, N. Ferro, and A. Poggi (Eds.): IRCDL 2013, CCIS 385, pp. 165–170, 2014.

2 Technical and Organizational Characteristics

In the first operative phase we had to select, prepare and send all those books conform to Google standards. We have tackled some problems such as the format for cataloguing data, books shipment and organization of the work between the SBS Centre and ten libraries.

The main requirements asked by Google to digitize the documents are:

All the books must have metadata. We get the metadata from the collective catalogue of polo RMS (we call polo a local part of our National Library Service, SBN). This catalogue has SOL software, an integrated library system (ILS), realized by Data Management; the software is a web application that manages all the main librarian functions (cataloguing, purchasing, lending, users database management), for back and front office. Most of scientific books and rare editions of Sapienza have already been catalogued. The software makes the data export from SBN format possible in order to exchange bibliographic data formats used both in Europe and USA, Unimarc and Marc21. We send a Marc21/xml file to identify our records in Google Books.

All the books must have a barcode with a univocal identification code, unique for all Sapienza libraries. The barcode reading is necessary in every phase: book shipment, digitization, metadata and book linking that will be available in Google Research Interface.

The Sapienza collection must be considered as one collection, even if we have several collections in different places that belong to economically and organizationally independent centers. The collective catalogue is very useful in this situation, but it gives us solutions just for a part of our problems. The organization of the activities and the communication with Google make it necessary to produce too many files and printed texts.

3 Solutions

First, it's necessary to develop an instrument, an operative context because of two main reasons:

- to limit the costs of developing a new component in SOL, used just in this temporary project
- to have system components which are quickly increased and adapted to our operative needs

For the realization of these instruments there is a team of SBS staff, librarians and software developers of Cineca, Sapienza's partner with a great experience in bibliographic data management and technical manager of RMS polo.

Here we have a brief description of the problems we tackled.

3.1 Data Format

Bibliographic data export to Marc21, generated by SOL software, is transformed in Marc21/xml according to a standard; but some modifications are necessary on metadata, as requested by Google, especially on multi-volume works' titles that must start and end in field 245. So this field has been appropriately modified, using $n and $p subfields for hierarchical links (see the example)

For the administrative section we add the own identification code in field 955.

```
<record>
<leader>00869nam 22001937i 4500</leader>
<controlfield tag="001">PAR0736263</controlfield>
<controlfield tag="008">100224s1869 it ||||  |
|||||||ITAod</controlfield>
<datafield tag="041" ind1="0" ind2=" ">
<subfield code="a">ITA</subfield>
</datafield>
<datafield tag="100" ind1="1" ind2=" ">
<subfield code="a">Curioni, Giovanni</subfield>
</datafield>
<datafield tag="245" ind1="1" ind2="0">
<subfield code="a">L'arte di fabbricare, ossia Corso
completo di istituzioni teorico-pratiche per gli
ingegneri, per gli architetti, pei periti in costruzione
e pei periti misuratori</subfield>
<subfield code="p">Operazioni topografiche</subfield>
<subfield code="c">per Giovanni Curioni</subfield>
</datafield>
<datafield tag="260" ind1=" " ind2=" ">
<subfield code="a">Torino</subfield>
<subfield code="b">A. F. Negro</subfield>
<subfield code="c">1869</subfield>
</datafield>
<datafield tag="300" ind1=" " ind2=" ">
<subfield code="a">351 p.</subfield>
<subfield code="c">25 cm.</subfield>
</datafield>
<datafield tag="774" ind1=" " ind2="0">
<subfield code="t">L'arte di fabbricare, ossia Corso
completo di istituzioni teorico-pratiche per gli
ingegneri, per gli architetti, pei periti in costruzione
e pei periti misuratori</subfield>
<subfield code="w">RMS191070</subfield>
</datafield>
<datafield tag="852" ind1=" " ind2=" ">
```

```
<subfield code="a">RMSAR</subfield>
<subfield code="c">ARlibro MINN. C 838 </subfield>
<subfield code="t">AR 33621 </subfield>
</datafield>
<datafield tag="955" ind1=" " ind2=" ">
<subfield code="a">BIBLIOTECA CENTRALE DELLA FACOLTA' DI
ARCHITETTURA</subfield>
<subfield code="z">RMSAR$$$000033621$$$D</subfield>
</datafield>
</record>
```

[Example of a modified marc21/xml file]

3.2 Barcode

The barcode is realized according to standard "code 39", it's made of 20 symbols (letters or numbers) and a check digit. Every barcode starts with RMS, the library system name; then we have 2 letters that identify the specific library and 15 symbols that represent the inventory number linked with the book. We use $ when we have an empty space, as Google wants, to make the search by barcode easier.

Table 1. Example of a barcode Barcode: RMSSTA$$000000497$$$F

Barcode: RMSSTA$$000000497$$$F	
RMS: polo code	ST: Earth Science Library code
A$$000000497$$$: inventory A 000000497	F: check digit

We decided to create a univocal ID with significant elements (not random univocal sequences) just to have an ID with an own link to the material object.

3.3 Such as One Collection

This process give us a global vision of all the books selected by libraries, so we send without mistakes just one of the items of a work and we don't digitize the same edition two or three times, for example, in two or three different libraries.

From the beginning, we decided not to not duplicate records, because, when we have more copies of a specific edition kept in different libraries, first we have to analyze the conditions of every single book to choose a suitable copy for digitization.

4 Work-Flow

4.1 Selection of the Books for Date and Realization of Lists for Next Steps

Often the bibliographic descriptions in our catalogue don't follow the standard punctuation of cataloguing rules, so the files exported in marc are not necessarily suitable

to select books for year of publication. The solution was to search for year of publication both in marc subfields and in ISBD description field (using a temporary database) with a specific procedure that recognizes frequent mistakes in cataloguing. This way we have more results than just searching in the marc field (10% more).

4.2 Declaration of Suitability According to the Project

We put all the data in a Google spreadsheet, one per library. The files are shared on a website with restricted access that contains all the information about the workflow. The librarian inserts in this file the following information:

- if the book is available
- if its material condition is suitable for digitization process
- book value, calculated by a Sapienza Library community algorithm

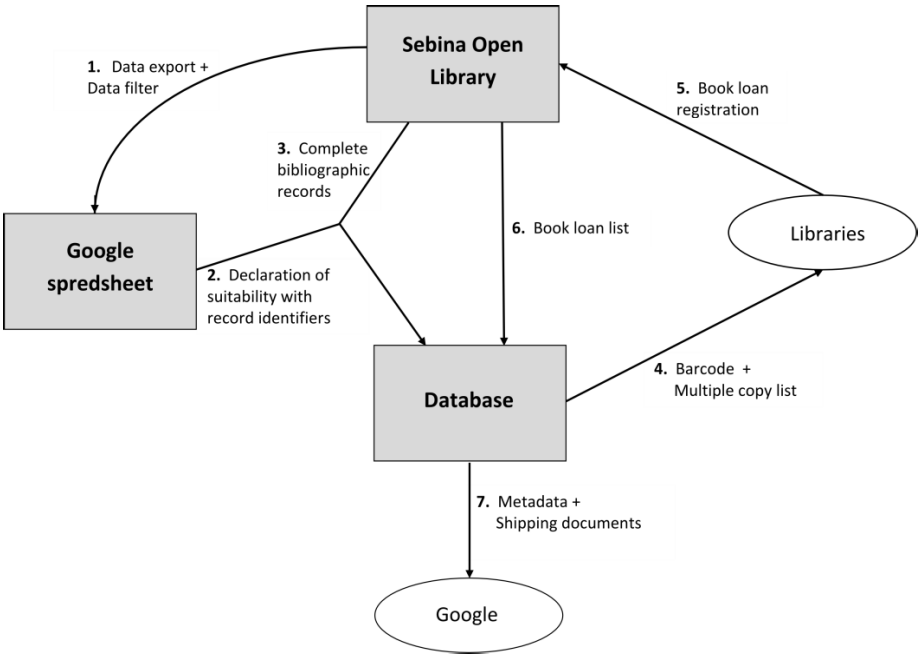

Fig. 1. Diagram of the work flow

4.3 Record Duplication and Barcode Production

All data produced by librarians are imported in a database (MySql) for the next steps. The process that populates the database, developed in Php by Cineca, uses "Google spreadsheet API". Only after the comparison of suitable book data, given by librarians, a unique list can be created choosing just one volume per work (usually the first one inserted in the server, in chronological order); a barcode is linked to the selected

book and all other copies are automatically discarded. The barcode generation phase is therefore also used as a check of multiple copies. The report list management also makes it possible to check if the books have already been digitized in past.

4.4 Books Loan Registration and Sending

When library staff put the books on the cart for shipment, they also register the loan in the catalogue. So, for every cart sent, there is a user ID with a standard code, linked to the library. The temporal sequences of book loans (timestamp information) are used to produce the correspondent lists of volumes ID, separating correctly "ready to go" carts from "come back" carts, for the ten different libraries across la Sapienza and in the city of Rome too. So, the list of volumes registered in Sebina is imported in the database; it is useful to take note of books sent every time, to prepare the necessary documents for each cart and to have a complete check list of all sent books.

5 Results and Developments

For students and professors, especially of this University, the Google Books project means a great improvement of services, both for the access and the quality of checked and enriched data.

The project that we are developing also includes an extension of document typologies for digitization and a better integration with the cataloguing database.

Until now, only monographic volumes were scanned, but we are going to digitize serials too. Serials have some peculiarities: for example, every single issue of a magazine is linked with a bibliographic description and it means a surplus of work is needed to make the digitization management possible and to make the identification of each digital object easy.

The involved libraries are increasing and the direction of next phase is getting clear; so it's necessary, as soon as possible, to align the cataloguing database with all the spreadsheets (now hand-made), maybe by web-service, just as it happened with the alignment of all student databases.