Digitization Projects of Documentary Collections in Academic Libraries

Ilaria Giglio
La Sapienza, Università di Roma, ilaria.giglio@uniroma1.it
DIGITIZATION PROJECTS OF DOCUMENTARY COLLECTIONS IN ACADEMIC LIBRARIES
Ilaria Giglio
PhD in Library and Information Science - Sapienza, University of Rome (Italy)
ilaria.giglio@uniroma1.it

Abstract

This paper is focused on digitization projects of documentary collections in academic libraries. The aim of the work is to suggest an evaluation of digitization projects by using a set of parameters deduced by the observation of national and international models. To create this evaluation scheme it has been necessary to look at the recent national and international academic literature and compare different case studies. The parameters were created by thinking about the whole process of digitization and also taking into consideration an user centred evaluation.

The created evaluation scheme has been tested on a sample of digitization projects of Italian, European and American academic libraries. With this kind of analysis it has been possible to check the validity of the evaluation scheme created, to identify points of strength and of weakness within the Italian system and to compare it with the international best practices analyzed.

Keywords:
Digital libraries; Digitization; Academic libraries.

During last year digitization projects have known a great dissemination in the cultural heritage area: libraries, archives, museums have used this technological process in many different ways, even because this kind of activity let them reach public and private capitals. For what concerns the digitization intended as the set of activities needed to convert analogue data into digital form is possible to identify, as known, at least two point of strength:

- by creating a digital copy of a material item it will possible to save it from all the accidents caused by the humans use, from the unexpected ones and from the spoil of time;
- if the digital space, created to host the digitized items, has been realised considering accessibility and interoperability it will be possible to reach a great number of users, probably even higher than potential physical ones.

There are also many relevant weaknesses that is always necessary to consider when it's starting a digitization project:

- digital volatility: while standard conservation procedures are adopted for physical materials, digital items are subject to obsolescence. We can’t be sure that present digital standards will be efficient in future and this is one of the reasons of digitization projects’ management costs: it’s periodically necessary to invest again in the conservation of digital materials and in their adaptation to developing technologies.
- Unplanned events: just like it happens for physical objects, also the digital ones could be damaged from unexpected accidents that humans are not always able to manage: black out, hackers' attacks, digital platforms and used software’s problems.

A library that want to start thinking about a digitization project will surely have to keep in mind these points and will have to try to adopt a strategy for maximize strength changing it in opportunities and minimize weaknesses. Below are shown some general indications, deducted from the observation of standards and of the projects that will be explained later:
- to safeguard and to protect the physical materials during their preparation for the scanning and to choose technologies that are appropriate for the materials' features and their formats.
- To keep more than one of masters’ copy with the best resolution in one or more repositories. As known, the best formats to use are TIFF for masters and JPEG for copies (or, if JPEG is not possible, also PDF).
- Every item has to be followed by a reach set of metadata, more specified as possible.

To guarantee accessibility and interoperability it would be better to take also some precautions:
- to adopt OAI-PMH protocol;
- to use Opens Source platforms;
- to let an open access consultation giving to users the chance to not sign up them self and to not login every time;
- to apply a Common Creative license (if the copyright allows it);
- to give to users the chance to download images directly from the portal.

To be sure to make the most, libraries will also have to consider standards and to benchmark with other cases.

Before to speak about digitization projects in academic libraries, it’s important to nominate some digitization projects in the world that have been pioneer in this kind of activity:

- Europeana [http://www.europeana.eu/portal/en]

For what concerns European projects it’s needed to remember also some portals [Caffo, 2010]:

- Multilingual Inventory of Cultural Heritage in Europe [MICHAEL/MICHAEL PLUS, http://www.michael-culture.eu/]

Finally we could mention also two Italian projects:

- Internet Culturale [http://www.internetculturale.it]
- CulturaItalia [http://www.culturaitalia.it/]

In last ten years digitization projects have known a great diffusion in academic libraries: the chance to create digital files of the materials owned and the possibility to share them with remote users is too much tempting to not consider it. But what are the aspects that a library has necessary to consider when it starts thinking about a digitization project and about the possibility to invest in it?

The aim of this study is exactly to propose an evaluation of digitization projects in academic library focused on the use of a specific set of parameters, created thinking about the whole project of the digitization: from its very start until the moment that the user browses in the digital collections and views an items on his display. To create this set it has been necessary to look to the national and international model.

The biggest problem has been the lack of a unique standard centred on the whole process of digitization. Indeed there are different standards and every time libraries have to search for them and have to try to understand the way for their application. Shown below the most important ISO documents for this area:

In front of all these materials each library has to identify singularly the standards needed and to find best practices to follow. For this reason there are millions of projects with different features, with no standardization and no interoperability. Platforms can’t communicate and can’t share their materials between one another.

For all these reasons this study would like to identify all the parameters that libraries have to consider starting a digitization project and to put this parameter inside an evaluation scheme that can be used to analyze single projects existing.

The parameters set has been created by looking to the recent national and international academic literature and comparing different case studies. In their definition has been considered the whole process of digitization and also the users’ point of view trying to define an user centred evaluation. To identify them it has been necessary to check all the ISO standards mentioned before, and also to compare them with another important European standard: Technical Guidelines for Digital Cultural Content Creation Programmes [MINERVA eC Project, 2008].

Below are explained the evaluation’s parameters.

Stakeholders
All the parts that in a directly or indirectly have contributed to create and to realize the project.

Costs
Where it’s been possible, I’ve tried to know the total cost of all the projects and the nature of the investment (sponsorship, partnership, fundraising).

Collections
For every projects analyzed I’ve described digitized collections considering many topics:

- kind of material digitized;
- items’ content;
- description’ features relating the single item and relating the collection belonging;
- chosen formats;
- access and consultation’s way.

Resources’ displaying
Focus of this parameter is the way in which the digitized items are displayed inside the referring collection, so for example: layouts’ features, the chance to view the image directly on the web without the downloading, etc.

Long term preservation
For all the libraries that want to start define a digitization project is necessary to consider how they will be able to guarantee the preservation in the time. So, when it’s been possible I’ve tried to describe the all conservation system thinking about the presence of masters with related formats, the presence of many servers in which keep more copies of the created archives. For every case I’ve specified used protocols, chosen metadata and their visualisation way.
Privacy and intellectual property rights
For the privacy all the international cases examined declare to respect the referring local law, while for the Italian ones considered law is the d.lgs. n. 196/2003, Codice in materia di protezione dei dati personali.

For what concerns the intellectual property rights, for every projects it has been specified the kind of license used.

Usability
According to ISO 9241-11:1998 the definition of usability is “The effectiveness, efficiency and satisfaction with which specified users achieve specified goals in particular environments”. According to the standard, the usability of an interactive system is given by the relation created between the system and its users and it is characterized by seven element that are: suitability for the task, self descriptiveness, conformity with the user expectations, controllability, error-tollerance, suitability for individualization [Polillo, 2010]. All these features have been used to analyze the usability of the portals where digitized collections are hosted, together with others considerations related to traceability of the same portals in the web, starting from many terms for the search and also in the universities' web pages. I've also tried to describe the interface by analyzing the position of the different elements on the display and overloading's presence. Finally I've also tried to give opinion respect my personal experience of navigation and of dialogue with the interface.

Accessibility
In the web the accessibility is the capability of an information system to give the same information to all the kind of users, without discrimination and with tools that can help people with disabilities. Accessibility is the access key for the system.

Platform’s architecture / Project Management
Regarding portals' architecture, usually they've been structured in line with the Open Access philosophy. Where it has been possible, I’ve tried to analyze deeply the features of each portal, focusing on their inner articulation. When the achievement of these kind of information has not been possible, I've focused the attention on the Project Management, trying to understand all the steps of the projects.

Interoperability
“Interoperability is a characteristic of a product or system, whose interfaces are completely understood, to work with other products or systems, present or future, in either implementation or access, without any restrictions” [http://interoperability-definition.info/en/].

Every project should consider interoperability as the most important goal to achieve. The chance for two different systems to communicate and to share their contents is a very considerable aspect to take under control and according to the MINERVA’s Technical Guidelines [p. 10] is possible to guarantee interoperability in two main ways:

- the use of standards, general rules, best practices in the creation, management and the organization of digital contents;
- use of protocols and of API (Application Programming Interface) in the publication on the web of the digital collections.

Metadata
I’ve tried to describe the specific kinds of metadata used in the materials’ description, the way in which users can view them (it's not always been possible).

These parameters have been used to test a group of Italian and international digitization projects with some goals:

- to understand the validity of the indentified parameters;
- to identify the main features of Italian and international digitization projects (strength and weaknesses);
- to understand the difference between Italian situation and the international one, considering that for the international I've chosen some best practices.
Shown below there are the institutions and the evaluated projects.

For Italy:

- DigitUniTo - Università di Torino
- Impronte digitali - Università degli studi di Firenze
- Sapienza Digital Library - Sapienza, Università di Roma
- Salernum - Università degli Studi di Salerno

International case:

- Harvard College Library Collections Digitization Program
- Cambridge Digital Library
- Bibliothéque numérique patrimoniale - Université de Strasbourg

To fill the fields of the evaluation scheme it has been necessary to find all the needed information. I’ve used two kinds of sources. In first time I’ve been searching for documents concerning specifically every single project. Unfortunately it’s been very difficult to find on the web this kind of materials, or, when I’ve found them, they were containing limited information. So, it has been necessary to study extensively the web sites and the digital portals. I’ve tried to analyze and to highlight all the steps: from the founding of the digital platform in the site of the academic library to the resources’ display and its features.

**Italian projects evaluation**

<table>
<thead>
<tr>
<th>Sapienza Digital Library (University of Rome)</th>
</tr>
</thead>
</table>
| **Stakeholders** | Inner: DigiLab, Sistema Bibliotecario della Sapienza (Sapienza’s librarian system), InfoSapienza.  
  **External**: CINECA. |
| **Costs** | 500,000 euro (more or less)  
  [http://www.cineca.it/it/progetti/digital-library]. |
| **Collections** | 20, formed by texts, images, video, audio and maps. |
| **Resources’ displaying** | Images: IA Bookreader (open source).  
  Video: HTML 5 or Flash Player.  
  Geographic maps: Djakota (open source). |
| **Privacy and intellectual property rights** | D.lgs. n. 196/2003, Codice in materia di protezione dei dati personali.  
  [cc-by-nc](https://creativecommons.org/licenses/by-nc/). |
| **Usability** | Highly usable system. |
| **Accessibility** | Highly accessible system.  
  Four kinds of access to resources:  
  - access Sapienza community  
  - open access  
  - closed access (copyright)  
  - reserved access.  
  Languages: Italian, English. |
| **Platform’s architecture/Project Management** | Content Management System: Drupal (open source: www.drupal.it).  
  Digital Asset Management platform: MediaMosa.  
  Jpeg treatment: Djakota. |
| **Interoperability** | SDL tries to guarantee interoperability with the goal to communicate with national and international digital projects. That’s why it uses open source and open access tools and also it’s in compliance with OASIS protocol and with international standards ISO-16363, METS, MODS, |
**Impronte digitali (University of Florence)**

| Stakeholders          | Inner: academic librarians (resources' selection, monitoring of the operation of materials' transport, control during the digitization).  
|                       | **External**: Space S.p.A. (external contracting company, for digitization). |
| Costs                 | 72,000 euro plus IVA. |
| Collections           | 5, formed by incunabula, ancient books, manuscripts, pictures, periodical. |
| Resources' displaying | Three kinds of visualization:  
|                       | - Detail  
|                       | - Web surfing with index  
|                       | - Sequential web surfing. |
| Long term preservation| For every items three copies:  
|                       | 1) File master for the long term preservation (in XML and in compliance with MAG 2.0).  
|                       | 2) Picture for the web visualization.  
|                       | 3) Thumbnail image.  
|                       | Collection organized in a tree structure, created by using directories. All files has to be given to University in double copy and also the external company has to serve it for three years on a own server. |
|                       | CC-BY-NC (Non commercial use). |
| Usability             | Good usable system. |
| Accessibility         | "A" level according to the WCGA (Web Content Accessibility Guidelines - http://www.w3.org/TR/WCAG10/).  
|                       | Use of accessibility software:  
|                       | - Cynthia says Portal, Torquemada, Color Blindness Check, JAWS screen reader.  
|                       | Languages: Italian, English, French, Spanish. |
| Platform's architecture/Project Management | No information found. |
| Interoperability      | In compliance with OAI-PMH Protocol. |
| Metadata              | Administrative: MAG 2.0.  
|                       | Descriptive: Dublin Core. |

**DigitUniTo (University of Turin)**

| Stakeholders          | An inner digitization project group specifically created, choosing 10 members from academic librarians and archivist.  
| Costs                 | No information found. |
| Collections           | 15, formed by modern books, archival documents, Turin’s Theatre programs. |
| Resources' displaying | Before the selection every items has a little miniature with the bibliographic description. After the selection the user can choose between different web surfing options:  
|                       | - preview;  
|                       | - complete description;  
|                       | - download. |
| Long term preservation| In compliance with OAI-PMH protocol. |
| Privacy and intellectual property rights | Every users can browse in the portal, without registration.  
<p>|                       | There are many licences used for sharing the items: |</p>
<table>
<thead>
<tr>
<th>property rights</th>
<th>Creative Commons Attribution Unported License; Creative Commons Attribution Unported License (non-commercial 3.0), for Turin's Theatre programs.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usability</td>
<td>Highly usable system.</td>
</tr>
<tr>
<td>Accessibility</td>
<td>Open source platform. Language: Italian.</td>
</tr>
<tr>
<td>Interoperability</td>
<td>Open source and open access contents. OAI-PMH protocol.</td>
</tr>
<tr>
<td>Metadata</td>
<td>Dublin Core. For books and texts UNIMARC/DC; for archival documents EAD/DC. Chance to export descriptions in XML.</td>
</tr>
</tbody>
</table>

**Salernum (University of Salerno)**

<table>
<thead>
<tr>
<th>Stakeholders</th>
<th>Only inners: a group of three different offices of the academic library.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Costs</td>
<td>No information found.</td>
</tr>
<tr>
<td>Collections</td>
<td>28, formed by Salerno’s picture postcards, archival documents, ancient and modern books, manuscripts, local periodicals.</td>
</tr>
<tr>
<td>Resources’ displaying</td>
<td>Users can choose to view the items on the web or download them (in PDF).</td>
</tr>
<tr>
<td>Long term preservation</td>
<td>In compliance with protocol OAI-PMH.</td>
</tr>
<tr>
<td>Privacy and intellectual property rights</td>
<td>The collections that constitute Salernum are included in the EleA institutional repository. It is in compliance with the Berlin Declaration regarding the open access in academic literature [Berlin Declaration on Open Access to Knowledge in the Sciences, 2007].</td>
</tr>
<tr>
<td>Usability</td>
<td>Moderate usable system.</td>
</tr>
<tr>
<td>Platform’s architecture/Project Management</td>
<td>Institutional repository: DSpace (open access).</td>
</tr>
<tr>
<td>Interoperability</td>
<td>In compliance with standard for institutional archives: protocol OAI-PMH, SWORD (Simple Web-service Offering Repository Deposit), technologies Opens Search and RSS.</td>
</tr>
<tr>
<td>Metadata</td>
<td>Dublin Core.</td>
</tr>
</tbody>
</table>

**International projects evaluation**

<table>
<thead>
<tr>
<th>Stakeholders</th>
<th>Harvard College Library Collections Digitization Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>Costs</td>
<td>Only inner: for every step of the projects (projects’ management, conservation, cataloguing, images and audio treatment, websites’ management) are assigned to specific department of academic librarians.</td>
</tr>
<tr>
<td>Collections</td>
<td>29 (more 4 in progress), formed by medieval manuscripts, historic pictures, music scores, pamphlets.</td>
</tr>
<tr>
<td>Resources’ displaying</td>
<td>Each digitized collection has an own website, accessible from the homepage of the project. For specific materials, there are different ways of displaying: pictures, stamps, objects: full record and zooming without loose of definition. Manuscripts and books: choice between the visualization of</td>
</tr>
</tbody>
</table>
single pages ad one image, or with a full reader that gives the possibility to leaf through the images.

<table>
<thead>
<tr>
<th>Long term preservation</th>
<th>The items are inserted in the Digital Repository Service: it is access and conservation repository together. In compliance with protocol OAI-PMH.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Privacy and intellectual property rights</td>
<td>Collection’s access is free, without registration. Regarding property rights, there’s a specific copyright statement, in compliance to Digital Millenium Copyright Act (DCMA) Pub. L. 105-304, for which every item is shared only for teaching or individual research.</td>
</tr>
<tr>
<td>Usability</td>
<td>Very high usable system.</td>
</tr>
<tr>
<td>Accessibility</td>
<td>In compliance with Web Accessibility Initiative, but the websites are only in English.</td>
</tr>
</tbody>
</table>

### Platform’s architecture/Project Management

| Interoperability | Open platforms, OAI-PMH protocol, in compliance with referring standards. The university has also created a specific web page [https://wiki.harvard.edu/confluence/display/HLSLibraryInteroperability/Home](https://wiki.harvard.edu/confluence/display/HLSLibraryInteroperability/Home) focused on library interoperability, that hosts a dedicated wiki (totally open, where everyone can write for proposal or feedbacks). |
| Metadata | MARC, MODS. |

### Cambridge Digital Library

| Stakeholders | **External**: The Polonsky Foundation for the best part of needed funds; the Jewish Manuscript Preservation Society, the Friedberg Genizah Project Inc., the ‘Arts and Humanities Research Council for the digitization of the Taylor-Schechter Cairo Genizah collection; the Arts and Humanities Research Council also for supporting the digitization of the Sanskrit Collection and of the Spanish Chapbooks; the National Science Foundation for the Darwin Correspondence Project. **Inner**: Digital Library Team; Digitisation Team; Conservation Team, for the practice works with the contribute of chosen professors and researchers. |
| Costs | 1,5 million of pounds from the The Polonsky Foundation, but it has not been possible to know total costs. |
| Collections | 37, formed by books, manuscripts, objects, archival documents. |
| Resources displaying | The platforms uses some JavaScript libraries: Seadragon Ajax, ExtJS, JQuery and Glow that guarantee an open visualization without plug-in needed. |
| Long term preservation | There is a specific Conservation Team with the rule to manage the digitized items and to preserve them in the time. |
| Privacy and intellectual | All the items are property of the university. So their use is allowed just for study and research. |
property rights
Usability Very high usable system.
Accessibility It’s guaranteed among the compliance with some statements:
- university disability and employment policy;
- equal opportunity policy;
- quality assurance agency (QAA) quality code;
- disability discrimination acts [1995, reviewed in 2005];
- special education needs and disability act [law Senda, 2001].
Platform’s architecture/Project Management Projected by the Digital Library Team, using Java language. It’s based on an open source platform, XTF (eXtensible Text Framework) and it uses Goobi for the data computerization.
Interoperability Open source and open access. XFT supports OAI-PMH protocol.
Metadata METS, MODS, TEI.

| Stakeholders | Only inners. The projects originates from the joined work of the Service Commun de la Documentation and the Direction Informatique of the university. These two divisions have also used academic and scientific partners for the selection and the creation of the digital collections. |
| Costs | No information found. |
| Collections | 17 formed by different kinds of materials. |
| Resources’ displaying | TIFF, JPEG, PDF, OCR. |
| Long term preservation | The project uses the ContenDm software, that guarantees the management of digital collections, but also masters’ repository and conservation. It supports OAI-PMH protocol. |
| Privacy and intellectual property rights | Users can view the collections without needing of registration. All the items are without copyright, so their use is free, except for commercial use (for this is required a specific application). |
| Usability | High usable system. |
| Accessibility | In compliance with WAI. |
| Platform’s architecture/Project Management | ContentDm, open access software projected expressly for digital libraries. |
| Interoperability | The items are also indexed in many catalogues: the University of Strasbourg one, SUDOC, CCFR and Worldcat. Thanks to OAI-PMH protocol they can be also shared in others digital libraries or web sites such as Medic@, Gallica and Isidore [Signalement et visibilité des fonds, www.docnum.unistra.fr/cdm/about]. |
| Metadata | No information found. |

At the end of this evaluation work is possible to identify some features of Italian and international projects, that are important to explain the main trends in this specific area. For Italian cases, could be also interesting to understand which is their own gap respect the international cases chosen, that have to be considered best practices.

**Italian projects features:**

- Collaboration between librarians and IT experts.
- Specific digital platforms (except Salerno – repository).
- National and international funding, no fundraising.
- Bibliographic and archival description using UNIMARC, ISAD and ISAAR, UNIMARC/DC.
Masters in TIFF, copies in JPEG or PDF.
OAI-PMH Protocol.
Privacy managed according the Italian law (D. lgs. 30/06/2003 n°196) - except Salerno (Berlin Declaration, 2007).
Creative Commons licenses.
Low traceability in the web and inside Academic Libraries web sites.
Care for the overloading.
Low attention for users' feedbacks.
Accessibility policies.
Use of open Content Management Systems.
More used metadata: METS, MAG (structural and administrative); Dcsimple (descriptive). XML exportation possible.

**International projects features:**
- Work functional division, different operating units for the same project.
- Specific web space for each collection.
- Attention to connections between items and referred resources.
- Diplomatic transcription and manuscripts’ translation.
- High traceability in the web.
- Medium traceability inside Academic Libraries websites (as in Italy).
- Modern interface, care for social media.
- Focus on users’ feedbacks.

Considering the comparison between these two different realities is possible to highlight how “Italian situation” is on the good way: by following the standards and the best practice it will be surely possible to get the best from the own projects.
Considering the proposed evaluation scheme is possible to note that it could be an useful instruments for academic libraries in two different times:

- During the first phase (projecting) when libraries could use it to understand which are the aspects to take under control.
- In the last phase (check), when it could be used to evaluate what has been done and to highlight best and wrong practice.

This work is deduced from my final paper for the higher degree in Library at Information Science [Giglio, 2014]. The paper hasn’t ever been published, but in it there is a deeper analysis about every chosen case, with the specification of all the aspects of the different projects, that are the base for the synthetic opinions expressed in this text.
References


