

Guillaume Fau,
Archivist, Department of manuscripts, BnF, Paris
Digitizing manuscripts at the Bibliothèque nationale de France :
technical and legal issues

These are the themes I would like to address to you in my paper : first the inadequacies of microfilm (MF), then how digitization resolved those inadequacies, thirdly the BnF's approach to digitization, and finally the technical and legal challenges faced in digitization

1 Over the past 60 years, the manuscripts Department has built a microfilm collection which amounts today to 25, 000 copies for the western division of the Department and 70 % for the oriental division.

This has been possible because MF has proved to be the librarian's best friend ... and the reader's worst enemy...

In fact, MF has proved to be reliable media for conservation, permitting easy and satisfactory transfer of data from one copy to another with a minimum of legal constraints (reproduction for consultation purposes, limited to the Department reading room, is permitted by the French copyright system).

But MF often provides a poor legibility (both for richly illuminated medieval manuscripts and contemporary manuscripts that contain many deletions and additions). Moreover, color MF is fragile. Finally, MF provides very limited public access to documents (currently, 11 desks are equipped with machines in the manuscripts Department reading room).

2 . This situation has led us to take up the challenge of digitization for several reasons :

- we want to ensure better image quality through the possibilities offered by digital photography
- we want to provide wider, including international access to our collections
- we want to take advantage of the possibilities offered by XML archival description and electronic metadata
- we want to experiment new techniques of interactive electronic annotation of manuscripts.

3 . So if we turn to the digital collections of the BnF, we must go back to the early 1990's when President F. Mitterrand launched the idea of a French national library of a completely new kind, which included the utopia of a comprehensive digital library.

In fact, *Gallica* (which is the name of our current virtual library), available since October 1997, was very quickly restricted to the digitization of printed books in the public domain, within a mass production context and in the image mode.

Then, cautiously, the BnF began to consider the digitization of manuscripts.

Today, digital manuscripts can be found in *Gallica* and in the *Mandragore* database.

In *Gallica*, digital manuscripts are available in the « Anthologie » section or in the thematic sections (Proust/Zola).

The « Anthologie » section aims at presenting on line fragmentary images of the treasures of the BnF to a wide public, so it has a very simple structure as you can see on the screen, starting from the home page, which provides search fields that you can restrict to a particular department. Today, you can have access to a list of 91 hits for the western division of the manuscripts department, ranging from medieval to mid-XXth century manuscripts.

Two digitized manuscripts can be found in the thematic folders of *Gallica* : draft copies of *Le Temps retrouvé* by Proust and *Le Rêve* by Zola. Both folders provide access to images of the manuscripts linked to brief records, transcriptions and original publication texts.

Finally, *Mandragore* is a richer and far more complex achievement : it is an illustrated digital catalogue of illuminated manuscripts, presenting 17, 000 images and 110, 000 records or rather iconic descriptions. It was designed to give access to the painted decorations of manuscripts to art historians : this is the home page, then you have the search screen with my « tristan » search, then the list of hits giving access to a close up image of the painting.

Now that you have an overall view of the digital collections of the manuscripts Department, let me tell you about our on-going digitization projects.

Our goal at the manuscripts Department, inspired by our wish to improve the results which I've just presented, can be described as follows :

- we would like to digitize complete collections
- we would like to give digital access to genetic literary sources
- all this, within an international context, using new standards in archival description and in collaboration with scholars and other institutions.

This is what we are currently working on with 2 projects : the Dunhuang project (for the oriental division of the manuscripts Department) and the *IDA* project (for the western division).

4 . The Dunhuang project or *MIDA* (Mellon International Dunhuang Archives), which is presented on the Internet on the *ARTstor* web site, is a multinational and multi-institutional effort to present on line the Dunhuang archaeological site in China, through the *MIDA* and the BnF web site, funded by the Mellon Foundation.

The project aims at recreating a 3-dimensional image of some of the caves, linked to digital images of the manuscripts and liturgical objects found in cave 17. These documents are scattered around the world, so we want to provide links with other archaeological sources such as photographs and mission reports.

The participating institutions are listed on the screen now, including the manuscripts Department of the BnF where about 8, 000 documents have been kept since 1910 when the library acquired the collection gathered by Paul Pelliot in his 1906-1908 mission along the Silk Route.

Here are a few figures : this is a 4 year project (which began in July 2001), with a \$ 1, 100, 000 budget granted by the Mellon foundation. At the manuscripts Department, 40, 0000 images are to be digitized by a team of 12 including project managers, technical advisors, curators and language specialists.

We thought that this project was also a good opportunity to improve the traditional digital presentation of manuscripts : we prescribed specifications for the shots that automatically included the restitution of the physical aspects of the documents that MF does not provide (for example, we asked for reconstructed virtual views of very long scrolls using pixel by pixel image stitching ; and we asked for views of physical details such as booklet binding and folded scrolls).

For the time being, 4/5 of the Pelliot Chinese collection has been digitized : that's 2, 986 documents, 21, 815 images, 918 CD.

That's as far as we've gone with the Oriental collections.

5 . What about the western division of the manuscripts Department, where we keep, apart from ancient and historical collections, extensive collections of modern and contemporary papers ? What can digitization do for the Rousseau, Hugo, Flaubert, Pasteur, Curie, Proust, Artaud, Cixous archives, to mention but a few among numerous others ?

The *IDA* project can be described as an on-line interactive annotation of literary manuscripts (« Annotation collaborative de l'archive manuscrite », in French).

Its goals are :

- to link the digital images of manuscripts to electronic records and to an electronic infrastructure of interpretations
- to provide the reader with consultation tools that individual or MF consultation does not make possible

IDA is a 2 year-project, which began in June 2003, including the following partners : BnF, ITEM and INRIA.

Each of the partners has a specific role to play : the BnF ensures a stable reference system for the description of the manuscripts (folio number), provides records and digitizes the manuscripts on site ; ITEM elaborates the required functionalities (zone description of the manuscript, possibility of an improved classification of the folios, typology of links available between zones of the text and external sources, construction of reading aids...) ; INRIA, in collaboration with ITEM, builds the interactive annotation structure using workgroup computing techniques.

The corpus was selected because of the interest presented by each manuscript regarding its, genetic approach, writing process and expected results. The physical characteristics were of course very important factors too : all 3 manuscripts were like tests for digitization (we wanted to show the different colors of the paper, the various sizes of the pages ; Flaubert's and Proust's manuscripts are written recto/verso and we wanted to be able to digitize both sides of the paper. Proust added « paperoles » and we thought it was essential to digitize and index them as precisely as possible too).

Trois Contes by Flaubert was selected for digitization because this manuscript is written like a script (successively, from notes, to drafts, to corrected publisher's copy). So what's interesting is to be able to identify text blocks (that is : the successive versions of the same passage in notes/drafts/fair copy versions) in order to establish a new virtual classification of the manuscript as opposed to the order adopted in the 1920's when the manuscript was bound. Now on the screen is the order of the folios in the actual physical manuscript (fair copy/draft copy/summary/reading notes) ; and now on the screen is the genetically correct order made possible by digital reconstruction (Flaubert first wrote the summary, then the reading notes, then the draft copy, finally the fair copy).

The second selected manuscript is notebook 54 by Proust which contains first drafts for « Sodome et Gommorrhe II » and *Albertine disparue*. The writing process is continuous, and scholars want to be able to draw a cartography of the text. Proust uses lots of cross references, such as the one shown on the screen now, to link two or more passages separated by numerous pages : this is the case with the Latin word « Mors » (mort/death), which Proust writes in red to indicate two complementary passages related to Albertine's death. Digitization makes it possible to read these 2 passages together.

I also wanted to show you the possibility of linking the image to the transcription of the passage.

The third and last selected manuscript for the *IDA* project is notebook 78 A by Paul Valéry. The writing process used by Valéry is a thematic one which means that digitization can help us to compare varying themes scattered throughout the text : a general abstract context can thus be built such as the one shown on the screen now, related to 2 different meditations on time and death, for example.

6 . Now that you know about the achievements in digitization at the manuscripts Department, it is time for me to consider technical and legal issues attached to them.

What are the technical choices made by the BnF ?

- manuscripts are digitized at the BnF (for security reasons) by an outside service provider
- formats, definitions and resolutions are as shown on the screen now. I'd just like to draw your attention on the 4 versions produced for each digitized document : archive/migration/Internet/Intranet version. These last two versions are important because they have legal implications that I'll explain later.
- As far as metadata are concerned, there are 2 points which must be considered :
 - the metadata retrieved from the printed catalogues (in other words, records and indexing) have been restructured using EAD which, as an international standard, is particularly useful for exchanges and interactivity with foreign institutions ; moreover, XML DTD, EAD gives us the possibility to illustrate the digital catalogue through links to images (what the MARC format cannot do). In the majority of cases, in fact in all cases for the IDA project, this has meant a considerable addition of descriptive information. Here is an example on the screen using the Flaubert manuscript. In comparison to printed records, EAD has made a folio by folio description possible. The EAD encoded record reflects as precisely as possible the intellectual structure of the manuscript. Of course, indexing possibilities have also been greatly enriched. Finally, in the Dunhuang project, another advantage of using XML is that UNICODE can be used to encode Chinese characters (that means transliteration problems are solved).
 - the metadata that derive from the digitization process (information on the digitization, its different versions, etc.) will be structured through METS (an XML schema). 2 aspects of METS are

particularly interesting for us : first, the descriptive and administrative metadata of the EAD record can be embedded in METS ; second, the structural map of METS offers the possibility of matching the logical description of the manuscript and its digital image.

Now what about legal issues of digitization ?

According to the French copyright system, owning a manuscript (for example, in a public collection or library) does not mean holding the copyright attached to it. That means that each time we want to digitize a manuscript we need to first check the property rights (limited to 70 years after the author's death in most cases, including reproduction rights) and then the moral rights (unlimited in time).

In fact, there are only very few cases where a library holds the copyright of manuscripts : this is the case when authors bequeath their copyright by will (for example, Romain Rolland to the BnF). But that is also the case of posthumous unpublished works for which publication rights belong for 25 years to the institution that keeps them and decides to publish them for the first time.

Sometimes, when texts have been published, publishers have acquired copyright by contract. Consequently, when libraries want to digitize these manuscripts, they have to negotiate with publishers as well.

Finally, there is the question of illegal copying on the Web : the BnF has chosen not to tattoo its digital images ; protection is ensured through a lower resolution quality that makes commercial exploitation impossible. There is also the possibility of higher resolution access on the Intranet only : in this case, each digitization is negotiated with publishers or authors so that copyright is respected.

7 . I would like to conclude this presentation by making a few additional observations.

Exhaustive digitization is very difficult and expensive to achieve. For the *IDA* project alone, digitization meant 3, 000 images at the end of the process, for just 3 manuscripts !! And we keep hundreds of manuscripts and notebooks for the Proust collection alone !! For the Dunhuang project, digitizing a simple folio costs 5.9 euros and digitizing a scroll nearly 8 euros !!

However, digitizing manuscripts is important for us because it is a way to give wider on line access to our collections : the Proust and the Zola folders in *Gallica* had 2, 301 and 506 on line visitors respectively last september, which is much more than would be possible or desirable in the manuscripts Department reading room.

But the challenge was also important for us curators as digitization has provided us with a new and more efficient tool to understand our manuscripts (reading and analyzing the original with the help of the digital manuscript offers new possibilities : we can experiment a chronological classification of the manuscript or we can have very helpful close ups that solve many deciphering problems).

As you can see, digitization adds value to our collections and it has enabled BnF to meet the challenges of preserving a national and international heritage in a way that benefits scholars and the public alike. However, with so much material to preserve, our major issue for the future will be mass digitization.

guillaume.fau@bnf.fr

monique.cohen@bnf.fr