Türk Kütüphaneciliği 23, 2 (2009), 337-350

From Gallica to Gallica 2[1] and More: A Survey of the Digitisation Projects at the BnF [2]

Sara Yontan Musnik*

Abstract

This paper will attempt to present a survey of the digitisation project(s) at the French national library in Paris. It will mention the various stages of this venture from its very beginnings until today, highlighting the achievements, but also pointing at the difficulties encountered, both material and theoretical. The talk will close with the question of what digitised collection building may mean for an encyclopaedic and comprehensive heritage library.

Öz

Bu metnin aslı, yirmi kadar slayd ve 3'41 dakikalık bir film gösterisiyle İngilizce olarak Pera Müzesi'nin ev sahipliğinde ve 18 Kasım 2008 tarihinde Fransız Anadolu Araştırmaları Enstitüsü ve Marmara Üniversitesi'nin katkılarıyla gercekleşen Kültür ve Sanat Kurumlarında Bilgi ve Belge Yönetimi Uluslararası Semineri'nde sunulmuş bir konuşmadır.

Yazı, Paris'teki Fransız ulusal kitaplığı Bibliothèque Nationale de France'da gerçekleşen tarama projelerini başından beri özetlemek amacında olup, projeleri gerçekleştirme yönteminde karşılanan teorik ve pratik zorlukları da belirtmek üzere, ansiklopedik ve külürel miras kapsamlı bir kütüphane için dijital koleksiyon geliştirmenin tanımını sorgulamakla noktalanır.

---

[1] Gallica and Gallica 2 coexisted for a while before all became Gallica again. Today as this paper goes to publication almost a year after it was written, many new features have been added to Gallica. Please visit the online catalogue and collection at http://gallica.bnf.fr/

[2] This talk was given on 18th November 2008 at the Pera Museum in Istanbul for the seminar the museum organised together with Marmara University, Information science department and the Institut français d'études anatoliennes (IFEA) entitled "Information and Document Management in Cultural and Art Institutions". It was followed by a 3'41 minute film on Gallica 2 (English version).

(*) Bibliothèque nationale de France, Paris. Türkçe bölümü sorumlusu. E-posta: sara.yontan@bnf.fr

I should like to thank my colleagues Jean-Didier Wagneur, Lionel Maurel and Christine Genin who have kindly shared their thoughts, experience and knowledge without which I would not have been able to see through the complex information necessary for this presentation

Introduction

To collect, to conserve, to organise the written as well as the visual and sound heritage in order to make it available to the public are the four fundamental missions of Bibliothèque nationale de France, henceforth the BnF. Technological progress does not drive these missions but rather serves them. That was the case with the printing press a few centuries ago; it is now the case with the digital revolution.

Before I begin with the core of my talk, and in order to provide some background information concerning the nature of my institution, I should like to emphasize the distinction between electronic resources held by a library and digitisation projects conducted by a library, both of which make up the digital collections.

Bibliothèque nationale de France has indeed a very rich offer in electronic resources. It gives access to 75 CD-Roms and 180 data bases; there are over 30.000 titles of e-journals of which 2000 are current subscriptions and the rest free access periodicals. To this, one may add a significant number of digitised texts such as encyclopaedias and various corpora compiled by publishers and purchased by the BnF. We will, however, leave this offer aside and deal with material digitised from in-house holdings.

A Brief Survey

When back in July 1988 the newly reelected President of the French Republic François Mitterrand had announced the project of a *«library of a completely new type»* as the next episode of the Bibliothèque nationale who was then suffering from lack of space for both its collections and its readers, what his aides had in mind was definitely a *«completely digitised library»*. Digitisation was of course a solution to shortage of space but also to better conservation as well as a means to remote, thus wider, access. It also rhymed with future for an institution which needed to update its image.

Twenty years, that is, almost a generation later, what can be said of this project?

As you may all know well, the new library was officially established in 1994, merging the centuries old heritage library with the newly launched institution towards the present Bibliothèque nationale de France. What was new about the BnF was its site built within a short lapse of time -only three years- a fruit of contemporary architecture. More so were the reorganisation of its holdings, the scope of its wider encyclopaedic collections, the policy of developing new media

and finally the ambition of more efficient services such as the on-line integrated catalogue.

It may be necessary at this point to take the time of an extra slide or two in order to introduce the BnF as it stands today, before we move onto talking about its digital activities and services so that we have a better grasp of its accomplishments during the last two decades.

A Useful Flashback

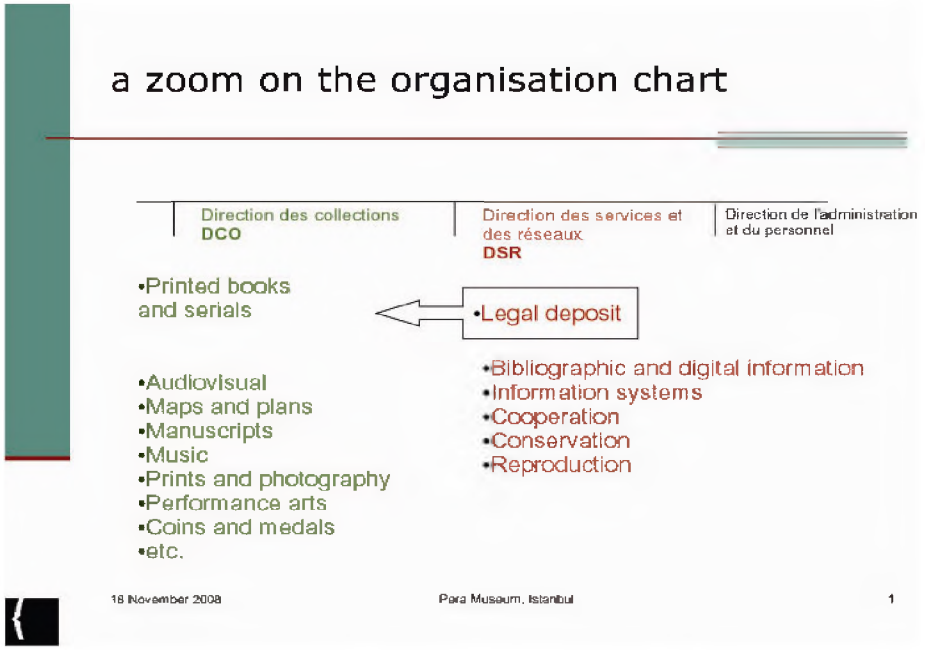
The BnF is the heritage library of France since the 14th Century. The origins of the holdings go back to Louis XIth's library which ceased to be the king's personal collection and became part of the kingdom's possession. However, it was François I who, in 1537, instituted the legal deposit with a royal decree whereby a copy of every printed book put on sale in France was to be given to the royal library. Even if this law was not necessarily enforced as expected in the early years, it set a milestone in the history of the institution and launched the spectacular expansion of its collections to be continued by other kings and kings' librarians all along the next two subsequent centuries.

The French Revolution too considerably enriched the collections via confiscations from church and other private libraries. Later, the 19th century increased further the number of items especially due to the explosion of the printed book and the dawning of journalism.

The Present Context

TToday the BnF holds over 13 million printed items, and the same amount of documents in various other media such as manuscripts, prints, music scores, maps, sound recordings, coins, etc. that have joined the collections via legal deposit but also via acquisitions by purchase, and to a lesser degree by exchange or gifts. To give an approximate estimate of one medium only, the printed book, the annual *«increase»* is around 60.000 volumes by legal deposit and the same amount by acquisitions.

The BnF is organised in three main units two of which will be of importance to our topic: Direction des Collections, hereafter DCO, and Direction des Services et réseaux -services and networks-, DSR.



(Fig.1): A simplified version of the organisation chart of the BnF zooming on the units that contribute to collection development

For our purposes, I shall highlight only some of their activities: the DCO, as its name implies, develops the collections and encompasses various departments to that effect, four of which collect material by subject matter and the rest by type of medium. The DCO teams also catalogue the material collected by the acquisitions librarians. The DSR, on the other hand, organises and provides cooperative networks and technical services, which include the collecting and cataloguing of legal deposit that end up at the DCO as well as the preservation and digitisation of material collected previously by the DCO.

These two units are thus very much involved in the achievement of the virtual library.

From Gallica…

To go back to the *«completely new library»* announced in 1988, indeed the teams that were preparing towards this promising new institution began picking titles, buying books and digitising them as early as 1992. The purpose was to build a general collection, useful to both researchers and laypersons. Consequently, the content included works covering literature from Antiquity to early 20th century,

periodicals of reference, tools such as dictionaries and bibliographies, all of which were in public domain. The material was, then, stocked.

However, when, five years later, in 1997, the website, later baptised Gallica, was available and ready to offer access to digitised material, various legal issues came up, some of which involved commercial interests of existing publishers and copyright of authors or their heirs. After a series of discussions and negotiations, important adjustments were to be made in this area.

This led to a number of new programmes that were launched in 1999 on the basis of thematic files, aiming at exhaustive collections on a given topic to be completed over a number of years. Themes still available on Gallica as *"files"* or *"dossiers"* include such topics as *Gallica Classiques*, *France in America*, *Voyages en France*, etc. In order to achieve such clusters, cooperative action was necessary with BnF's several partner libraries known as the "pôles associés" network[3], as well as other heritage institutions. As the projects progressed, the need for drafting and publicising a digital collection policy was inevitable. It was finally written in 2004.

The next year, that is 2005, was also the beginning of a very important component of Gallica, namely the digitisation of daily newspapers which deserves to be highlighted: it is an ongoing project supported by an exceptional subsidy from the French Senate that covers 31 titles of national and regional dailies that have been widely read during the 19th and the first half of the 20th century in France or in French. This amounts to some 3,5 million pages in both image and text format. The technical aspects of this project have served as a laboratory and enabled improvements in Gallica.

By the end of 2006, close to a hundred thousand text items and almost as many images were already available on Gallica, mostly in image mode.

To Gallica 2[4] via Europeana…

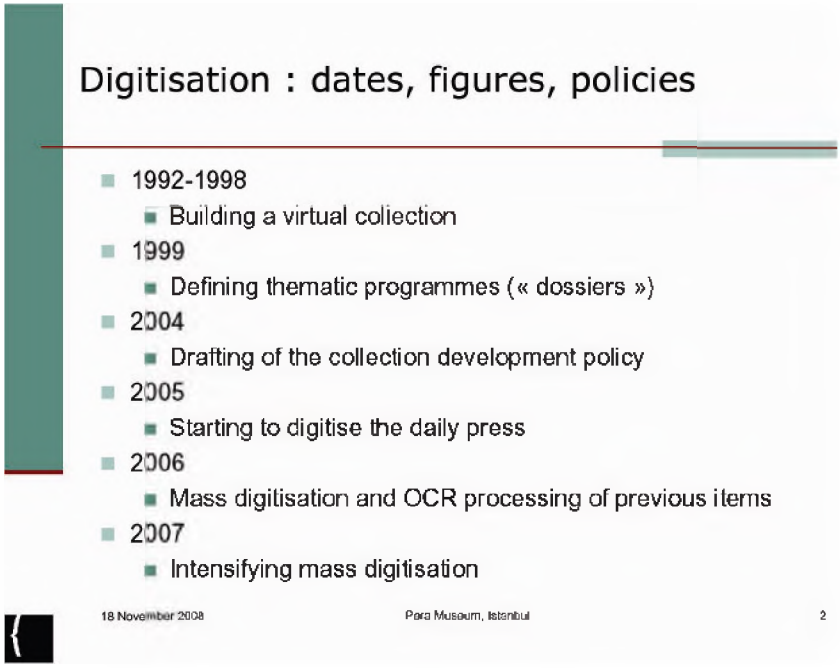
While Gallica, proud of its site, updated in 2000, was growing at a decent rate of 5 to 6.000 items per year, offering an encyclopaedic digital library of various formats, Internet was progressing at a giant's pace. At the eve of 2005, Google announced that it would digitise some 15 million books - held mainly by major American university libraries- and render them accessible worldwide. The president of the BnF at the time, Jean-Noël Jeanneney, immediately answered back to this challenge, not to say counter attacked, with a European project whereby EU

[3] *Pôles associés* : Network of libraries in France that began with shared acquisitions on given fields of excellence and that pursue partnership on other fields of action such as digitisation.

[4] Please refer to footnote 1

national libraries would unite and cooperate towards a common digital library. The main argument that convinced the twenty-three potential partners was obviously a political one, that of an alternative cultural policy. Google seemed to promise bulk, rather homogeneous in content while the European project defended a varied and organised offer. Thus was conceived Europeana, the prototype of which is to be launched officially the day after tomorrow, on November 20th

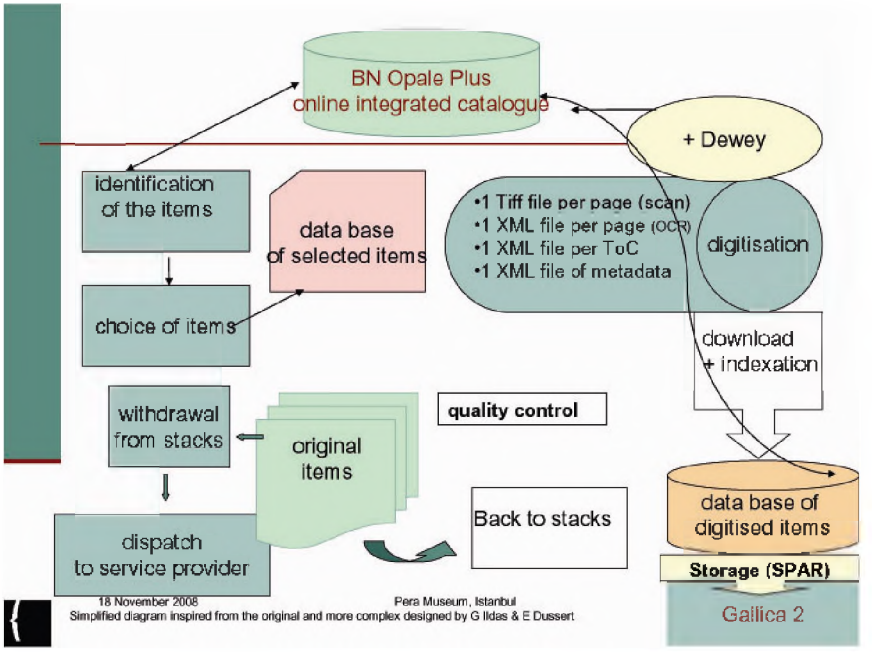
To set a model to the partner libraries in Europeana, to show the way so to speak, the BnF engaged in a mass digitising project as of 2006, speeding up from 6.000 items annually to 30.000, from simple image format to text mode. However the significant turn in strategy was announced the following year whereby mass digitisation more than tripled in volume, aiming at some 100.000 additional items per year on the virtual shelves of Gallica.



(Fig. 2): Milestone dates in the process of building digital collections at the BnF

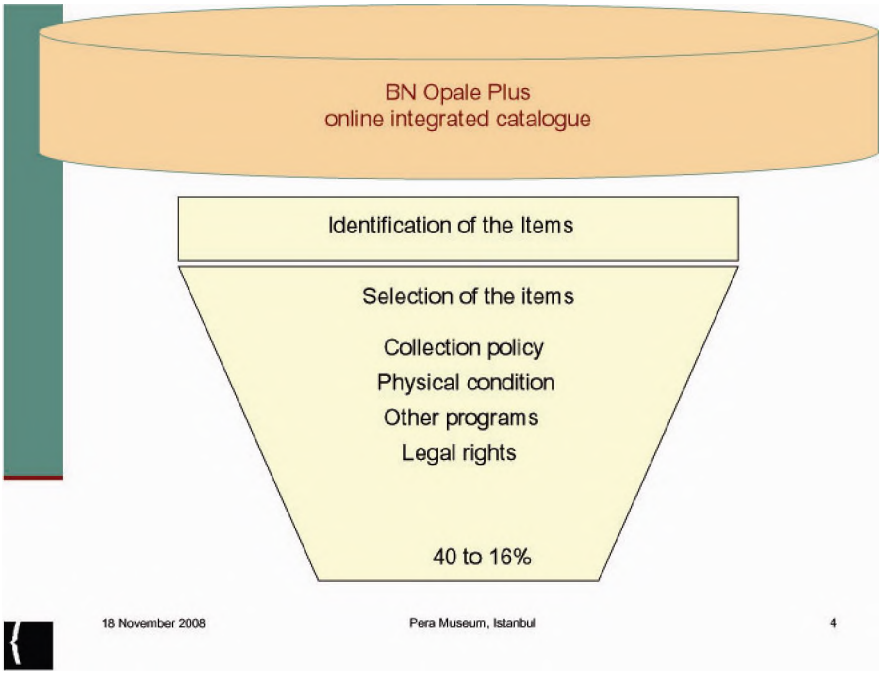
This ambitious venture which is meant to last three years, requires a different approach : it implies the sweeping across or the combing of whole shelves of books that have been grouped together by subject matter in the past. Unlike today where books receive subsequent shelf numbers as they arrive, up until the mid nineties, according to the old system they were grouped under a letter of the alphabet, each

of which corresponded to a discipline. "A" to theology, "Y" to poetry, "F" to law, etc. In other words, while the logic of corpus is at work again, the scopes are much wider than the "files" mentioned previously.



(Fig. 3): Simplified diagram of how Gallica is "fed"

Yet, in spite of the quantities involved and the rhythm to be kept, the volumes whose records are downloaded on a separate database are examined by a scientific team for the content. Attention is paid, for instance, to avoid similar or poor editions. This selection is then followed by a logistics team that controls the physical condition of the items to determine if and how they can undergo scanning and OCR processing. The criteria include the quality of the paper or the nature of the binding, the margins and the typography. After the approval of both teams, a third party checks the selection against existing digitised material either announced or already comprised in other projects by institutions that are part of the "pôles associés" network mentioned above. Ultimately, the legal feasibility is confirmed. After sifting through all these criteria, the amount of the selection that ends up being digitised varies between 40% at the most and 16% minimum of the initial selection.



(Fig. 4): Records and items funnelled through the process of various selection criteria

New Momentum

Consequently, this new momentum called for the adjustment of the collection development policy as well as numerous cooperative projects especially on a national level.

Indeed the scope of the heritage digital collections points at three directions, which, in fact, merge and spread where France's cultural legacy meets:

- **National heritage programmes**, namely milestones of French history, French literature, major French periodicals, etc.
- **Internationally oriented programmes** such as French translations of major European works, French journals abroad, etc. and
- **Programmes related to Europe's influence** such as ideas that have spread through and from the continent.

Cooperation

With regard to cooperative agreements, for the time being there are three basic models:

The simplest one involves sharing digital holdings. Since 2006, Gallica hosts digitised resources of its network partners, either by archiving their material on the BnF server or by harvesting their dematerialised resources via Open Archive Initiative protocol. Examples are many. Access, both remote and on-site, is naturally free.

Another model is that of conditional access -the condition being that of registering as a research reader with the library and viewing the content on site. An example to this case would be the full contents to 150 titles of periodicals in social sciences digitised by CAIRN, a group of academic journal publishers.

Finally a third and rather innovative agreement is one that is presently being experimented with publishers and retailers that enables partial access to books under copyright. This venture will be discussed in more detail below.

It is of course impossible to have national borders to cooperation when one talks about digital resource sharing. While the BnF pours the contents of Gallica 2 to Europeana, it also federates other French institutions' electronic input to the European digital library. Various other agreements exist such as the mutual access to a series of collections with the Library of Congress or shared digitisation of French language newspapers.

Indeed, I would like to say a few extra words in particular with regard to this last enterprise.

The idea of a French-speaking library network germinated in 2006. The countries involved then were Belgium, Switzerland, Luxembourg, Canada, Quebec and France. Soon Egypt via its Bibliotheca Alexandrina joined the group. Today, it includes nine more countries all from Asia and Africa (a full list is available on www.rfbnn.org).

The creation of a common digital collection through a unique portal seemed the next natural step to be taken. Supported by the Organisation internationale de la Francophonie, the first bulk of the digitisation activity covers the newspapers. As mentioned earlier, the BnF has a very ambitious project in this field, which, besides the French national and regional press, and thanks to this ongoing international project, includes newspapers in French published in other countries. Thus, thanks to BnF's impulse and know how acquired through its own programmes, a single gateway leads the reader to copyright free Tunisian, Vietnamese, Haitian etc. newspaper heritage in French.

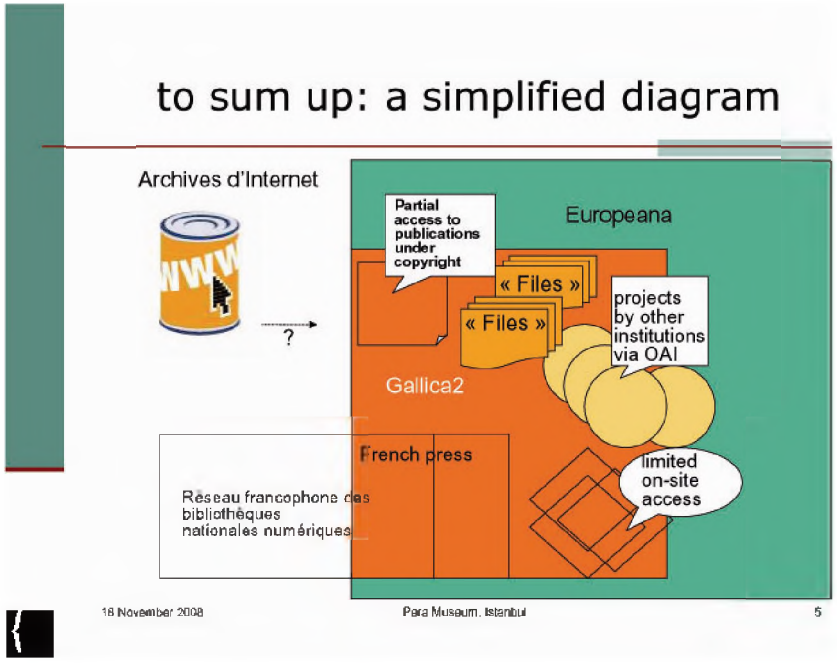
Features

The main feature of Gallica 2 besides its scope of documentary offer thus is the possibility of organising this huge mass of information via OCR (optical

recognition of characters) processing. This enables sophisticated semantic search and extraction of significant data, a fundamental *"added value"* to material that may be found on the Web otherwise. In the long run, search will be possible in a multilingual context. The performance of the new technological search tools is such that the organisation of the results will be of utmost importance.

Besides the well-known *"advanced search"* criteria, many other navigation features and interactive services are either already available or on the agenda. Among the *"customerisation"* functions are downloading, flagging and marking, printing, mailing or even creating one's own virtual library through personal files. The item's bibliographic record or metadata will naturally be available as well as a short abstract and the table of contents.

To Sum Up



(Fig. 5): The digital collections of the BnF in a (square) nutshell

To define Gallica 2 in a sentence then, one would say that it is an on-line digital library service *and* library collection, respectively provided *and* compiled by the BnF mainly but not solely from its own holdings that vary from printed materials to graphic works to sound recordings. Since the BnF is an encyclopaedic

heritage library, its digital collection inevitably reflects these two specific aspects. However, in building its digital encyclopaedic heritage collection, the BnF pays attention to offer both primary source material and research tools, bearing in mind that spreading knowledge is not necessarily a purely academic activity. In other words, Gallica 2's ambition is to serve both the scholarly community and the wider public to research and/or to discover France's cultural legacy.

By the end of the first decade of our century, by 2010, 400.000 items and over 14 million pages mostly extracted from BnF holdings will be accessible via Internet. The first results, in both image and text mode, have already joined Gallica 2, the updated version of Gallica.

Some Perspectives

Bibliothèque nationale de France, as a state run public institution, has recently pronounced its strategic plan for the next three years that is for the period covering 2011. The first of the six aims listed -and not only *one* of the six aims listed- clearly states that BnF's priority is *"to become a digital library of reference"*. To quote further in an unofficial English translation by myself, *"BnF will develop... a digital offer, both original, diverse and rich, in order to hold on to its pioneering position in this field and become the library of reference, thanks to the variety of material and services it will provide"*.

- The three actions identified to reach this aim are
- To complete the offer of "mass digitisation" with material in complementary formats, namely sound and image
- To develop quality access to books under copyright, in collaboration with publishers

To lead the project of a French-language digital library [with private and public partners] on behalf of the State.

The first step is a matter of widening the scope of the heritage holdings. Indeed lately major map collections and sound recordings have been heavily, if I may use the term, digitised. There are also additional programmes that foresee the digitisation of the immigration, colonial and clandestine newspapers as well as smaller size projects involving precious bindings. This is accompanied by regular preservation digitisation.

As for the third step towards the fulfilment of BnF's primary rôle, namely that of becoming a reference digital library, it requires acting on behalf of the French state and cooperating with a number of other institutions towards the development as well as the organisation not only of the contents, but also of the technologies necessary for wider and longer access to collections.

The second "action" because of its innovative nature deserves a zoom:

This experiment was mentioned only briefly earlier: it implies no commercial counterpart for the BnF and relies on pure cooperation with publishers, namely that of giving partial access through Gallica 2 to several thousands of recent books available in digital format via e-retailers' websites. The working group made up of BnF and the national union of publishers (SNE) commissioned an economic model to a consultant, expert in digital libraries (Numilog) that required

- The involvement of authors, publishers, retailers;
- The guarantee of reading practices similar to those for texts free of copyright;
- The respect of intellectual property and just payment to holders of rights.

The consultant's report concluded that the model had to involve payment for the end users. Technical and legal issues were then settled and the first specimen submitted to public appreciation during the 2008 Book Fair in Paris last March.

BnF plays no commercial role in this partnership. The CNL, Centre national du livre, a public administration receives proposals, examines them against the digital collection building policy of the BnF and if they are suitable, subsidises the commercial partners, potentially all French publishers and e-retailers.

At present, the number of participating publishers is over 140, including the most prestigious names of the sector. The aim is to have 10 000 titles available within a year's time, that is until the upcoming Paris Book Fair in March 2009 when an evaluation by all participating parties will decide of the future of this experiment.

Towards a Conclusion

To repeat a fundamental principle, all public domain items digitised from BnF collections in Gallica 2 may be screened, downloaded or printed free of charge. Contemporary publications, which may be searched (advance search) and found via Gallica 2, however, are labelled as such with a tag that alerts the potential reader of the conditions of access. A click on the title provides more information on the item in the form of a particularly rich record, including table of contents, the blurb on the back cover, a description of the book and a brief excerpt. If the patron wishes to find out more, there is a link to the e-retailer which in turn offers a few pages of the book for free and various possibilities of paid full access that vary from temporary reading to downloading to even purchasing the printed version!

This experiment requires three successive agreements or frameworks for

agreements, the first of the two being a prerequisite for the last step which is of our concern:

- Between authors and publishers,
- Between publishers and e-retailers,
- Between e-retailers and the BnF.

The reason for BnF's participation in this new venture is to offer the products contemporary edition and widen the scope of digitised material available through the library.

And more?

More quantity, more variety, more cooperative enterprises, more services…all this will be inevitable and probably the topic of other lectures in the future. However, in the meantime I would like to mention, if only briefly, another important component of BnF's electronic collection, which is not -at least not yet- part of Gallica 2, but a separate venture. Although not a pioneer in this field, BnF has decided to take the decisive step in collecting websites as part of its five centuries old legal deposit mission.

Unlike the traditional legal deposit procedure whereby the publisher or printer entrusts the BnF with the printed, electronic or audiovisual item, in the case of websites, it works the other way around. Here the BnF collects and the publisher agrees to supply the technical facilities, if necessary. Although the decree concerning Internet dates from August 1st, 2006, the BnF has started harvesting since 2004, all .fr domains as well as .org or .com domains produced or hosted in France, some of which go back as far as 1996. The capture is done in two ways:

a) through annual wide harvesting that obviously collects only a sample of the material available on the Web. The robots copy pages, images, films, audiovisual files which are then dated and indexed according to their original publication context and,
b) by more frequent focus crawls of web librarians who in so doing develop specific thematic collections such as personal diaries, blogs or official political party websites that invaded the Internet during the presidential and parliamentary elections and provide their findings to the robot.

It is important to note that some national or regional French newspapers have almost immediately volunteered to support this experiment by depositing their e-versions, thus contributing to the scope of the digital library without waiting for the robot nor the librarian, through the traditional procedures of legal deposit.

In fact legally speaking Archives d'Internet concerns Internet publications that are addressed to a larger public via electronic means such as institutional or individual newsletters, periodicals normally accessible through payment, blogs,

etc. The law applies to websites whose contents have something to do with the national territory, i.e., France. Further amendments with regard to specific details are being drafted through the year.

For obvious legal reasons meant to protect the rights of the organisations or persons producing the collected websites, access to this material is possible only on-site in the research reading rooms of the BnF. Six months ago BnF's Internet archives already comprised around 13 billion files.

The technical means necessary to preserve all this electronic data « forever », that is, permanently, as part of the heritage mission BnF is entrusted with for the future generations, involve an extremely powerful digital stack called SPAR, a distributed archiving and preservation system.

Information about all these achievements and projects may seem difficult to seize as it is developing and moving constantly at an incredibly fast speed. However much of it is available on the Internet in different forms and via different foci, depending on the web page of the institution that presents it. To enable a global perspective, the French Ministry of Culture performs a valuable and important task in identifying and checking all the initiatives, some of which are carried out on a national level and some of which inevitably end up being part of international enterprises such as Europeana or La Grande bibliothèque numérique francophone.

To Conclude

I should like to finish this presentation where I began, namely by listing the four basic missions of the BnF as a national heritage library: to collect, to organise, to conserve and to communicate. Obviously, these tasks will undergo mutation in the digital environment. Collection building will have to appeal to creativity; for while organising knowledge, librarians will inevitably contribute to creating its content as well.