Preservation of ebooks: from digitized to born-digital

Sophie Derrot, Jean-Philippe Moreux, Clément Oury, Stéphane Reecht

Preservation of ebooks: from digitized to born-digital

Sophie Derrot
Legal Deposit Department
Bibliothèque nationale de France
sophie.derrot@bnf.fr

Jean-Philippe Moreux
Department of Preservation and Conservation
Bibliothèque nationale de France
jean-philippe.moreux@bnf.fr

Clément Oury
Legal Deposit Department
Bibliothèque nationale de France
clement.oury@bnf.fr

Stéphane Reecht
Department of Preservation and Conservation
Bibliothèque nationale de France
stephane.reecht@bnf.fr

ABSTRACT

The scope of digital curation at the BnF covers documents digitized from BnF collections as well as born-digital material bought by the BnF or collected under its legal deposit mandate. It is therefore critical for the library to investigate if common approaches may be adopted for similar document types, whatever their origin may be. This paper proposes to focus on the case of electronic books (or ebooks), by comparing the way BnF teams intend to ensure the long term preservation of those directly digitized by the library and those that will enter through legal deposit.

Data and metadata formats are different, even though EPUB appears as the reference format for both kinds of ebooks. Acquisition procedures are necessarily specific. However, for the other steps of the treatment process, digitized and born digital books should follow similar and parallel workflows: indexing in BnF General Catalogue, access through the digital library Gallica and preservation in SPAR, BnF's digital repository. Common validation tools, characterization schemes and preservation metadata will be used in order to preserve both faces of French digital heritage.

General Terms

design, documentation, experimentation, legal aspects, performance, security, standardization, verification

Keywords

digital library, legal deposit, born-digital archives, digitization of heritage content, accessibility, ebook, DRM, EPUB, ONIX, PDF

1. INTRODUCTION

The scope of digital curation at the BnF – i.e. the set of processes intended to acquire, index, give access and preserve digital resources – covers two distinct kinds of digital material:

- on one hand, documents digitized from physical media (books, engravings, maps, etc.) held in BnF's collections;
- on the other hand, born-digital material bought by the BnF or collected under its legal deposit mandate.

Even though these resources are distinct in terms of acquisition method, legal status and heritage value, and even though they may depend on different organizational entities, we find in both cases similar types of documents: books, periodicals, images, audio and video.

It is therefore critical for the library to investigate if common approaches may be adopted for similar document types, whatever their origin may be. This paper proposes to focus on the case of ebooks, by comparing the way BnF teams intend to ensure the long term preservation of those directly digitized by the library and those that enter through legal deposit. It intends to see if they present similar characteristics and issues regarding long term preservation, if the same tools and the same workflows can be used, and where expertise can be shared. This paper does not take into account questions related to ebooks acquired for payment, as the topic is not yet mature enough. However, when the procedure for handling them is eventually designed, it will benefit from the double experience of digitization and digital legal deposit.

2. EBOOKS AT THE BIBLIOTHÈQUE NATIONALE DE FRANCE

2.1 Digitization: from image files to ebooks

The BnF has been involved in the field of digitization since 1992 through a digital library, Gallica, with nearly 2 million documents (books and periodicals, newspapers, engravings, manuscripts, objects, sound recordings, audiovisual material).

The BnF has recently decided to enhance the public dissemination of its digital contents through the production of electronic books, in addition to image and text modes. The aim of this new delivery format is to take benefit from some of the advantages of dedicated electronic book formats in comparison with the standard delivery formats usually offered by digital libraries (web, PDF):

- nomadic reading outside the digital library's website, on dedicated devices in a dedicated ebook format, EPUB;
- better dissemination of contents and better accessibility of digital contents for visually impaired people.

During the period 2011-2013, this effort has been embodied in two separate digitization programs: integration of an EPUB production process in a mass digitization program; and reprocessing of documents previously digitized, with the production of tables of contents and EPUBs. Digitization is either performed in-house or by a private contractor.

The implementation of this new format in the library has required much interaction between all the BnF teams involved in heritage digitization and some radical changes in the way the library considers its digitization activity:

- Intellectual selection of documents: for cost reasons, all the digitized books can't have an EPUB version. A choice must be made and selection criteria have to be defined: the librarian is turned into a publisher. In addition, the EPUB format is not suitable for all types of documents and some difficulties can arise in reconciling individual intellectual selection and the lack of flexibility of a mass digitization

program, which needs to be fed with thousands of documents every year.

- Publishing: a "heritage EPUB" template suitable for accommodating a variety of types of documents has been defined, as well as an EPUB production charter. Again, the library must change its habits: it no longer produces facsimiles of heritage documents but a totally new editorial product.
- Quality Assurance: the BnF automatic input control system has evolved, in order to analyze this new format (metadata, technical requirements, etc.). A specific EPUB quality assurance team has been set up within the quality assurance section of the digitization service, to perform visual checking of EPUBs on reading devices and assessment of text quality.
- Archiving and long-term preservation: section 4.2 below describes the ingest process of the digitized books in the BnF long-term archiving system, SPAR.

In 2014, the Legal Deposit of ebooks is an opportunity to leverage the EPUB expertise acquired from the digitization programs.

2.2 Digital legal deposit: from web harvesting to direct deposit

In August 2006, an extension of the law on legal deposit mandated the BnF to collect, preserve and provide long-term access to all French online publications. Until recently, the BnF digital legal deposit team mainly focused on setting up a complete web archiving workflow[1]. Legal, heritage and technical reasons explained this choice. On one hand, websites were considered the most at-risk documents. Besides, tools and best practices were already available thanks to international cooperation. On the other hand, the ebook market was not very dynamic, both in terms of production and sales, and the production and distribution workflow were still maturing on publisher's side.

Finally, in the absence of a decree enforcing the law, the BnF was not able to ask for content distributed under payment and was limited to harvesting freely available resources.

When the decree on digital legal deposit was published, in December 2011 – it is now part of the "Heritage code" – the BnF started designing its ebooks deposit system.

This entry track is still under development, but some critical decisions have already been taken. First, a deposit system was preferred to web harvesting for ebooks. The BnF could have chosen to crawl ebooks hosting platforms, as it currently does for news websites [3]. However, direct deposit, via an FTP platform, was more appropriate to allow unitary treatment and cataloguing of ebooks. Moreover, in several cases, ebooks distributed online are not directly hosted on the website of the online bookstore, which is only the place where financial transaction occurs; the document itself or the link to download it is then sent to the purchaser by another mean (e.g. by email). In that case, web harvesting would not have been efficient.

Second, it was decided to work with ebook "distributors". In the French ebook market, the publisher is in charge of creating the book (both the intellectual content and the digital document); the digital bookseller is in charge of promoting and selling the ebook to end-users; the role of the distributor is to make a bridge between these two stakeholders: it receives the ebook, checks its format, verifies and enhances the quality of its metadata, and sends it to the bookseller.

Working with distributors appeared very quickly as the best solution as:

- there are few distributors compared to hundreds of publishers;
- the BnF benefits from a first set of quality controls performed by distributors, both for the ebooks and their metadata;
- distributors receive ebooks without DRMs; they are therefore able to send them to the library without DRMs.

2.3 Parallel workflows: using common tools for digitized and born-digital books

Finally, an internal workflow has been designed. Ebooks deposited by distributors along with their metadata will be received on a dedicated FTP platform. A first set of checks will be performed by the library, in order to ensure that all declared documents are available, to verify that data and metadata are consistent, and to validate the format of the ebooks and metadata. If the package that has been delivered passes the check, the ebook receives a legal deposit number and the distributor is informed that the document has been deposited. If not, the library requests a new deposit from the distributor.

The entry system is necessarily different for digitized and born digital books. However, for the other steps of the treatment process, the library intends that both kinds of documents follow similar and parallel workflows:

- Descriptive metadata will be ingested (and potentially corrected by human cataloguers) in the BnF General Catalogue, which indexes most published resources hosted in the Library.
- Access to digitized books will be given via Gallica, the BnF digital library; access to deposited ebooks will be given via Gallica intra muros. This is a specific version of Gallica, which is only accessible within the Library premises, and which gives access to content protected by intellectual property rights (as recent documents entered through legal deposit are).
- Preservation will be ensured by the BnF digital repository, SPAR, which is described more thoroughly in section 4 of this paper.

This choice has been made in order to avoid reinventing the wheel and redeveloping already existing tools. However, the reader's perspective and needs were also taken into account: readers would probably have been lost if forced to use two series of tools and applications. In short, BnF readers should not need to know BnF internal systems, and should not wonder if they are looking for a digitized or a born digital document before accessing it.

3. DIGITIZED AND BORN-DIGITAL EBOOKS: TECHNICAL CHARACTERISTICS

Management of ebooks, from entry to access, has thus become a strong issue across the whole of the BnF. Questions related to preservation have been particularly taken into account, as both digitization and digital legal deposit channels are intended to deliver documents that will be accessible over the long term. From this point of view, do both types of documents present the same characteristics? This raises two series of questions especially

[1] See [1] for questions related to digital legal deposit legislation; and [2] about the web harvesting workflow set up by BnF.

important in a preservation perspective, the one related to ebook formats, the other related to their metadata.

3.1 Formats

3.1.1 Digitization track

In 2011, the BnF chose EPUB 2 as support of its digitized books program, because it was the de facto technical standard for digital reading. The alternative "fixed layout" was not used because it was not yet specified at this time.

BnF ebooks have been designed to present as few problems as possible in terms of preservation:

- EPUB is based on standards and formats already mastered: XHTML, CSS, Dublin Core, etc.
- They are produced under a BnF charter: their content and structure are well known, and remain consistent over digitization programs.
- They do not include contents or formats that are potentially "risky" for preservation: multimedia files, programming code for interactivity, etc.

The next mass digitization program (2014-2017) will foster the accessibility to digital contents with the EPUB 3 format. This new version offers a wide range of accessibility mechanisms based on the semantic annotations of content. These mechanisms are all based on markups (HTML5 markup and EPUB 3 specific markup). Consequently, the risk for preservation is considered sufficiently low.

3.1.2 Legal deposit track

Historically the jungle of commercialized ebook formats is very dense, as every content or device producer has tended to create its own format (PDB, LRF, LIT, MOBI, etc.). Over the past couple of years though this density has tended to reduce, to the advantage of standardized formats (EPUB) or closed-source formats of the market leaders (such as Amazon's KF8). Even if it does not solve every issue, this simplification is quite a relief from a preservation point of view, especially in a legal deposit context which theoretically extends the perimeter of the objects concerned to all produced and commercialized formats.

Above all, it is necessary to determine the limits of the scope of legal deposit. We tried to concentrate our efforts on files which can be easily defined as "books" in comparison with printed books. Therefore, ebooks on formats such as TXT or DOC are initially excluded, as they are closer to production formats than to diffusion formats. These formats are never to be found on commercial markets and distributors do not work with them. At the other end of the ebook channel, physical reading devices and their software won't be concerned either by this deposit track.

Once these exclusions have been made, what is left? During the initial discussions, publishers and distributors made clear that the most frequently produced and sold formats were EPUB 2 and PDF; MOBI and EPUB 3 were in minority. Distributors add DRMs to these files or send them to international online booksellers (Apple, Amazon and Google); these online booksellers take then care of the migration into their own format.

It was agreed that ebooks with DRMs will be excluded, to permit manipulations of the files during the deposit process. Frequent copies are necessary for deposit, access and preservation processes, yet they are often prevented by DRMs. Setting up a legal deposit system of this kind of protected material didn't seem feasible[2]. Our publisher partners agreed to that pragmatic position, which simplifies the whole process. Closed source formats won't be deposited for the same reasons: preservation systems will probably be unable to deal with them, especially in the long term. In spite of the ambiguity of this position – in contradiction with the traditional objective of comprehensiveness of legal deposit –, it seemed more reasonable to proceed in this way in order to ensure long-term access and preservation – which are also part of the objectives of legal deposit.

3.2 Metadata

3.2.1 Digitized ebooks

The EPUB format embeds metadata (Dublin Core) to provide information about the digital publication. These metadata are exported from the BnF catalogue. Some of them have particular values in a library context:

- ID: ark[3] of the digital document in the BnF digital library (Gallica).
- Source: HTML link to this digital document.
- Relation: catalogue entry of the heritage document.

EPUB 3 version offers a richer description of the bibliographical metadata and enables the inclusion of accessibility compliance metadata in an ONIX[4] message.

But the EPUB file, as every digital object in the library, must also be characterized within the BnF IT systems:

- Version: EPUB 2 or EPUB 3?
- Format: standard EPUB or fixed layout EPUB?
- Quality: Bronze and Silver are heritage EPUBs produced by mass digitization programs, with two text quality levels; Gold are editorial EPUBs (enrichments, editorial works, etc.).
- Accessibility: does the EPUB embed accessible features?
- Production information: service provider, tools used, date of production, etc.

This information is relevant for various uses: diffusion, preservation, production, etc.

3.2.2 Deposited ebooks

The main idea when putting in place an ebook legal deposit was automation, including re-exploitation of metadata created by publishers and distributors for their own needs. The most used format for carrying this metadata in the book trade is ONIX for Books, a XML standard designed to support computer-to-computer communication of bibliographic information. ONIX files are generally completed by distributors from information provided by publishers and then sent to online booksellers along with ebook files.

- Advantages of ONIX files attached to ebooks by the distributors are their richness and precision: this information has to be exact because of its commercial purpose.

---

[2] About issues of preserving digital resources with DRMs, see [4].

[3] ARK (Archival Resource Key) is a persistent identifier system created and managed by the California Digital Library. See https://wiki.ucop.edu/display/Curation/ARK.

[4] ONline Information eXchange. See http://www.editeur.org/8/ONIX.

- On the other hand, quite a few of the data are useless from a librarian point of view (for example: prices for every country where the ebook is commercialized) and some important bibliographic metadata is often lacking (ISBN of the printed version is not always provided).

To make existing ONIX files as useful as possible for us, an ONIX model including the BnF specifications was defined by librarians and proposed to publishers and distributors. In the meantime, an ONIX-to-Intermarc[5] conversion was developed to allow an easy and automatic transformation of the trade information into bibliographic notices. This conversion also enables the use of this metadata for preservation needs: it will be reused within the METS file attached to the EPUB or PDF file into the Submission Information Packages (SIP).

4. COMMON APPROACHES FOR INGEST?

4.1 SPAR in a nutshell

SPAR, the Scalable Preservation and Archiving Repository, is the BnF preservation system [5]. It has been developed since 2005, and seeks to conform to the OAIS model. Its initial scope was to automate all entities that could be automated, and to offer a wide range of functions, in order to preserve various types of asset. Up to now, development was mainly concentrated on the Ingest, Storage, Data management and Administration modules.

The sets of documents to be ingested are grouped into tracks and sub-tracks (channels), according to their nature (digitized books, audiovisual files, web archives, administrative records…), to their legal frameworks, and to the way the BnF plans to manage their life cycle and apply preservation strategies. At the present time, SPAR ingests objects in four tracks: Digitized documents and associated files (except audiovisual), Audiovisual objects, Web legal deposit (ARC or WARC files), Third party storage (various kinds of files, from partners outside the institution); several others are in progress, including the Negotiated legal deposit track that will be presented in 4.3.

Each track needs a specific preingest module, because no producer[6] is able yet to deliver well-formed SIPs according to SPAR's requirements. These modules build SIPs depending on specific settings and send them to a generic ingest module, which transforms them into Archival Information Packages (AIPs).

Four levels of formats are distinguished in SPAR, corresponding to four levels of risk: stored (the most unsafe), identified, known and managed [6]. A "managed" format has published documentation, at least one reference tool and a characterization scheme. Besides, the BnF may define use restrictions depending on the producer.

Metadata for package and preservation information are contained in METS files, with PREMIS elements. Metadata for data management are expressed in RDF.

4.2 Ingest of digitized books

EPUB files aren't considered preservation copies of digitized books, but a medium for dissemination. Though, their cost and their value explain that the BnF intends to preserve them in the long term.

When they enter SPAR in the "Preservation digitization track", EPUB and adaptative[7] files are controlled and characterized. It was necessary to find a characterization tool and a characterization scheme for EPUB files. This difficulty was solved by using and adapting Epubcheck 3[8], in order to improve this software and make it a basic characterization tool. This solution is not yet completely satisfactory, and we are still looking for a characterization scheme in order to record the preservation metadata extracted by the tool. This is the reason why it can't be said yet that EPUB is a "managed" format in SPAR.

EPUB 2 and 3 files are accepted, in both standard and fixed layout for EPUB 3. This corresponds to the three kinds of ebook formats produced or to be produced soon in our digitization process. The quality level (Bronze, Silver or Gold, see 3.2.1) will not be checked, but the information declared by the digitization contractor will be preserved, as well as other production information.

Ebooks are at the moment considered as associated objects of books digitized in image mode (TIFF and now JPEG2000 files). They are described in the METS manifest with specific fileGrp use (*epub* or *adaptative*) and structMap type (*ebook*)[9]. So ebooks can't be ingested alone: they are delivered either with new digitized books or while reworking digitization (pictures are re-delivered with new OCR and EPUB files). The possibility to deliver an isolated ebook and then create a new completed version of an existing AIP is yet to be investigated.

4.3 Ingest of deposited ebooks

As explained in 4.1, a dedicated preingest module will be developed for the "Negotiated legal deposit track". The legal deposit is considered "negotiated" as the form of the delivery is negotiated between the distributor and the library, so that for example not all formats will be accepted.

This module will get the package ready for ingest, joining together the ebook file (EPUB or PDF), the original ONIX file, possibly a picture file of the book cover, and finally the METS manifest.

The role of the SPAR system in the legal deposit workflow is critical. On one hand, the preingest module will provide fundamental information to the General Catalogue, such as the ARK number of the ebook file, i.e. its persistent identifier. On the other hand, SPAR will deliver the deposited ebooks to the access platform and application.

Within SPAR itself, two channels will be set up, according to the ebook format and its components:

- Ebooks whose format is considered "managed" after all quality checks and characterization will be ingested in the first channel.

[5] Intermarc, in the family of MARC formats, is the BnF format for bibliographic metadata.

[6] In the OAIS model, the producer is the external or internal entity that produces the resource and transfers it to the Archive with the mandate to preserve it.

[7] DAISY format, used to create text or audio books for visually impaired people.

[8] https://github.com/IDPF/epubcheck.

[9] See BnF's METS profile for SPAR: http://www.loc.gov/standards/mets/profiles/00000039.xml.

- Files that are valid from a format specification point of view but which present some preservation threats will be ingested in the second channel. For example, EPUB files containing Flash or JavaScript elements are not considered managed as BnF does not have sufficient confidence in its capability to preserve them in the long term.

Thanks to the digitization track, Epubcheck has already been chosen to perform a new format check on EPUB files when they enter SPAR. But the choice has still to be made for PDF. It is currently investigated if Apache™ Tika 1.5[10] may be used in addition to Jhove: the first one to characterize the files; the second one to validate them against the results of Tika's characterization and against pre-defined profiles. This will also be an opportunity to improve other tracks containing PDF files (particularly administrative records), where Jhove is the only and imperfect tool for validation and characterization.

For the legal deposit track, XMP will probably be used as a characterization scheme for both formats PDF and EPUB. If this characterization format is considered relevant, it will in turn likely be used for the EPUB files produced by digitization. Thus, files and formats analyzes would be performed in a consistent manner. Every file of each format will be handled with the same tools and schemes, regardless of the channel or the track it belongs to. Only the application rules will differ.

Some critical choices are thus still to be done. The BnF intends to proceed on these questions during the current year, and to ingest the first deposited ebooks at the end of 2014 or the beginning of 2015.

4.4 From a digital strongbox to digital library stacks

In the current situation, that is for the digitization track as well as other tracks (e.g. web archiving), there is a fork in the document management workflows between access and preservation. SPAR is not a step between entry and access, but only one branch of the fork, separated from the access branch. This current solution is not really satisfactory, as SPAR still appears as a digital strongbox, not as BnF digital stacks.

The ebook legal deposit tracks will represent a chance to improve this situation. In this workflow, the SPAR system will play the role of the central application, as it will receive the documents from the entry step, send information to the catalogue and provide the books to the access application.

However, this architecture decision implies some challenges:

- First, SPAR will need to show a better ability to communicate with other library applications.
- Second, it should develop its capability to provide the expected files according to defined rules (for example if only one format for a specific book is requested).
- Third, the response time of SPAR must be guaranteed, because every slowdown or interruption will increase the delay between the entry of a document and its visualization.

In this way, the BnF will be able to ensure that it gives access only to documents that are already ingested in the repository. There won't be any difference anymore between what is preserved and what is offered to readers.

[10] https://tika.apache.org/1.5/index.html.

5. CONCLUSION

Ebooks from legal deposit and digitization tracks differ in various aspects: they have different legal statuses; they were not acquired for the same goals and for the same audiences; and BnF's preservation mandate for them is different. In one case, the BnF (or its contractor) is the producer of the documents; in the second case the BnF is only the depository.

Moreover, even though both kinds of ebooks are available in the same formats (EPUB and PDF), their technical characteristics may differ (use of JavaScript, of embedded content, etc.).

It is nonetheless possible to adopt common approaches and to leverage developments performed for one track to improve another track. Systems originally built for digitized books (Gallica/Gallica intra muros for access, SPAR for preservation) will be used for ebooks received through legal deposit. Common tools (Epubcheck, Tika, Jhove) and common characterization schemes (XMP) are applicable in both cases.

Benefiting from expertise of various teams with different backgrounds has actually been a strength: crossing points of views brought a global vision considering all aspects of ebooks preservation.

6. REFERENCES

[1] Illien G., Sanz P., Sepetjan S. and Stirling P. 2012. The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future. In *IFLA journal*, 2012, vol. 38, n° 1. [http://www.ifla.org/files/hq/publications/ifla-journal/ifla-journal-38-1_2012.pdf]

[2] Le Follic A., Stirling P. and Wendland B. 2013. Putting it all together: creating a unified web harvesting workflow at the Bibliothèque nationale de France. [http://www.netpreserve.org/sites/default/files/resources/Putting%20it%20all%20together.pdf]

[3] Oury C. 2012. When press is not printed: the challenge of collecting digital newspapers at the Bibliothèque nationale de France ». In *Proceedings of the IFLA Preconference, newspaper section*, (Mikkeli, Finland, August 2012). [http://www.ifla.org/files/assets/newspapers/Mikkeli/oury_clement.pdf]

[4] APARSEN. 2013. *Report on DRM Preservation.* [http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2014/02/APARSEN-REP-D31_1-01-1_4.pdf]

[5] Derrot S., Fauduet L., Oury C., and Peyrard S. 2012. Preservation is Knowledge: A community-driven preservation approach. In *Proceedings of the 9th International Conference on Preservation of Digital Objects* (Toronto, Canada, October 2012). [https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf]

[6] Ledoux T. 2012. *SPAR: From Design to Operations*. Presentation at the Preservation and Archiving Special Interest Group (PASIG) (Austin, USA, January 2012). [https://lib.stanford.edu/files/pasig-jan2012/12B4%20Ledoux%202012_01_11-BnF-SPAR-FromDesignToOperations.pdf].