My Fair Metadata: Cataloguing Legal Deposit Ebooks at the National Library of France

Mathilde Koskas, Sophie Derrot

My Fair Metadata:
cataloguing legal deposit ebooks at the National Library of France

Mathilde KOSKAS, Sophie DERROT

Abstract:

French law on digital legal deposit covers websites and online content as well as ebooks. It imposes no obligation to produce a bibliography in the traditional form, indexing being sufficient. But despite their innovative characteristics, ebooks are still books, and their bibliographic metadata is closer to that of printed materials than to the indexation of web archives. To set up a complete ebook deposit workflow, the Bibliothèque nationale de France (BnF) benefits from its experience with digital documents and its tradition of legal deposit. This paper aims to present the questions that it faces when dealing with the cataloguing of ebooks and the management of their metadata, and the solutions that are emerging.

Keywords: ebooks; legal deposit; national bibliography; automated metadata retrieval; cataloguing standards.

Legal deposit was established in 1537 in France and it has followed the development of the national intellectual production and the various media used to disseminate it ever since. Through a 2006 law on legal deposit of digital content, the Bibliothèque nationale de France has the legal mission of collecting online content as well as printed books. For a long time, this online content has been mostly in the form of websites, but since 2012, the BnF has been working on a specific legal deposit for ebooks. The nature of ebooks is dual, between electronic documents and traditional books, and their status is ambiguous, as the debate within the European Union about their tax rate reveals. Such a legal deposit falls within the 2006 law, but the workflow which allows a better treatment by the BnF and an easier way for publishers to deposit their documents is very close to the traditional legal deposit for printed books. The major issue is that the digital legal deposit law differs on several points from that on printed legal deposit: there is no obligation of comprehensiveness or of producing a

national bibliography. These differences have to be kept in mind because they draw a line between the treatment of ebooks and that of printed books.

The dematerialisation of the legal deposit process began with the possibility for publishers to fill in their mandatory declaration online, thanks to a dedicated Extranet. More recently, they have also been able to mandate their distributor to send a metadata flow that acts as that declaration. The next step is the processing of ebooks and their metadata arriving at the BnF through an automatic process. To achieve this new evolution of legal deposit, the BnF has to adopt a position on important bibliographic questions and to resolve tensions between the bibliographic theory and the organisation of such an institution. In many ways, the postulate of a proximity with printed books has its limits, which we must take into account when we decide how best to create and disseminate bibliographic records for ebooks.

Working with publishers' metadata

The project of a legal deposit of ebooks began in mid-2012, with a dialogue with our partners from the publishing world, the Syndicat national de l'édition (French Publishers Association). This first phase was important because the ebook market was quite unknown to the BnF in many of its dimensions: how many ebooks are published per year? What formats are the most common? How does the publishing and commercialisation workflow work? All these questions were crucial to the organisation of a viable legal deposit. One of the first questions we faced during these meetings was that of metadata: how does the information about an ebook go from the author and publisher to the final seller and consumer? As our knowledge of the workflow increased, we realised that some of the main actors add metadata during the process and that the need of a standardised way to transport metadata has therefore pre-existed the BnF demand.

The standard widely used within the publishing world (in France and worldwide) is one of the ONIX[1] standards, dedicated to books. *ONIX for Books* was originally developed by EDItEUR and is maintained today by this company and an international steering committee[2]. It is designed to support computer-to-computer communication, in XML language. An ONIX file

[1] **ONIX** (**ON**line **I**nformation e**X**change): "one member of a family of XML-based international standards intended to support computer-to-computer communication between parties involved in creating, distributing, licensing or otherwise making available intellectual property in published form, whether physical or digital." http://www.editeur.org/74/FAQs/#q1

[2]. http://www.editeur.org/8/ONIx/

can describe printed books as well as electronic books, and can express bibliographic, technical and commercial metadata thanks to a great variety of fields precisely described in an up-to-date documentation available online. Within the French ebook market, the ONIX file describing an ebook is created by the distributor based on the information supplied by the publisher.

The BnF already has experience of processing distributors' ONIX files for printed books. Since late 2014, publishers have the possibility to make their mandatory legal deposit declaration directly through an ONIX flow provided by their distributors to the BnF. Books are sent separately and when they arrive at the BnF, their digital declarations are checked and then automatically transferred to the cataloguing application for treatment by the cataloguers. After this last step, some specific metadata created by the BnF (RAMEAU indexing for example) can be sent automatically to the publishers who need it, to complete their own database. This experience was very helpful in the implementation of the ebook legal deposit workflow.

All of our partners use ONIX 3.0 files to transfer ebooks metadata from the distributor to the online sellers. They provided us with samples from their production, so we could have an idea of the quality of these ONIX files. Meanwhile, we conducted a reflection on the form that an ebook record should take within the BnF's General Catalogue[3]. We compared the information from both sources and thus made an ideal ONIX model, with mandatory, optional or wished fields. This model was sent to our partners; it tends to evolve according to the ongoing reflection on the subject and the information provided by distributors.

Samples of the BnF ONIX model, concerning the ebook format and its title:

ONIX 3 fields	Values and description[4]	ONIX standard[5]	BnF wishes
<DescriptiveDetail>		M	M
<ProductComposition>	00: Single-item retail product	M	M
	10: Multiple-item retail product		
	11: Multiple-item collection, retailed as separate parts		

[3]. The BnF already buys ebooks (and access to ebooks). However they are not described within the General Catalogue but within a specific application dedicated to "Electronic resources", such as databases or e-journals.

[4]. These values refer to code-lists attached to the standard.

[5]. M is for Mandatory; R for Repeatable; O for Optional.

	<ProductForm>	EA: Digital (delivered electronically)	M	M
		EB: Digital download and online		
		EC: Digital online		
		ED: Digital download		
	<ProductFormDetail>	E101: EPUB		M
		E107: PDF		
		E108: PDF/A		
		E200 (if E101): Reflowable		
		E201 (if E101): Fixed format		

[…]

<TitleDetail>			R	R
	<TitleType>	01: Distinctive title	M	M
		03: Title in original language		
		Other values are optionnal		
	<TitleElement>			
	<TitleElementLevel>	01: Product	M	M
	<TitleText>	Text of a title element	M	M
	<TitlePrefix>	Text at the beginning of a title element	O	M if prefix
	<TitleWithoutPrefix>	“No prefix” indicator	O	M if prefix
	<Subtitle>			M if existing
	</TitleElement>			
</TitleDetail>				

Thanks to the ONIX format, the BnF and all its partners have a common way to handle metadata. The ONIX file describing an ebook is mandatory for the legal deposit of that ebook. Both metadata and ebook files will be uploaded to the BnF sFTP platform by the distributor. The automatic treatment begins by checks (presence of both files, format compliance checks of EPUB, PDF and ONIX files) and once the files are certified to be well-formed and matching the BnF’s requirements, the metadata is converted for the next step, cataloguing.

Organisation of the cataloguing workflow

The aim of this agreement with the publishers was to fit a workflow by which metadata enters automatically the cataloguing channel of the library. The BnF already has experience dealing

with ONIX files describing printed books, as mentioned earlier. This work is intellectually still quite close to the traditional method: printed books arrive in the usual way and are handled by cataloguers and the ebooks workflow is quite similar in terms of the treatment of metadata.

As it arrives on the BnF server, the metadata sent by the distributor is controlled both automatically (compliance with the ONIX format but also compliance with the BnF model) and manually. This manual procedure is still under construction, but librarians will have to check whether the incoming ebook comes under our definition of a book or should be sent to one of the BnF's other legal deposit sections (periodicals, maps, music, audiovisual, prints and photographs) to receive the proper bibliographic treatment. After this first step, the ONIX file is converted into an INTERMARC record in the same way whether it describes a printed or an electronic book. The result of this conversion is a rather full basis for a bibliographic record, already mentioning a majority of the needed information (title, contributors, publisher, etc.). The cataloguers then have to check this information and enrich the record to reference level, the highest-quality level of description, as defined by French cataloguing rules, with all authorities, indexing and correct bibliographic description.

The organisation of the cataloguing of ebooks is still in its prefiguration stage, but the major elements are already in place. A taskforce of five cataloguers and a coordinator was created in the unit of the library which already catalogues all the printed books arriving to the legal deposit. As an integral part of this unit, they must be able to catalogue printed books as well as ebooks. During the experimental phase of the project, only a few cataloguers will work with ebooks, defining new ways of working. But the aim is to spread this expertise and to make every cataloguer of this unit able to deal with both ebooks and printed books.

To prepare the arrival of ebooks within the workflow of this unit, a large training and communication program was put in place: several general and specific information sessions took place for both the dedicated taskforce and the rest of the unit. Cataloguers were mostly very positive and keen on information about the project. Formation sessions are also planned between the newcomers and the existing team, so the knowledge about cataloguing at the BnF can be directly shared between experts. Several working groups were created on subjects related to the arrival of ebooks in the unit: cataloguing, training and communication. Dynamic communication media were created internally, namely thematic "digital coffee sessions" and

an internal monthly newsletter dedicated to the project. The new cataloguers are fully involved in these initiatives and it made their integration within the unit easier.

The aim of this training and communication program was to help the whole cataloguing team to make this new object and working principles their own. However, the cataloguing of the first ebooks to arrive will be assigned only to those in the taskforce; it will enable us to study and experiment a new way of working and answer questions like what kind of work station these cataloguers need (*e.g.* two screens or a single, larger one). As the whole workflow is automated, the cataloguers can access the ebooks only when they are available in the digital library, after the preservation step[6]. So their work on the record really finishes and closes the process. Thus they will have the possibility to catalogue with the book "in hand" – or, more precisely, before their eyes. But some questions remain open, such as the suitability of the public interface of the digital library to a consultation in order to catalogue.

We have already learned from our first experience: the taskforce's offices have been brought closer to those of the rest of the unit and they take part in technical tests. Working sessions are organised on the reading interface and they make remarks which are taken in account for further developments. A "virtual book cart" was created to enable us to distribute incoming ebooks thematically for cataloguing and indexing (mirroring the organisation we have for printed books), to access the ebook and do all the management operations we would do with shelves and book carts in the physical world. The thematic distribution is based on the metadata from the publishers' ONIX, namely the CLIL subject classification system[7] (now aligned with THEMA). Due to its very practical use, the virtual book cart was first developed as a very simple interface, which will evolve as needed.

Besides these very pragmatic points, we had to think on an almost conceptual subject: the definition of an ebook's bibliographic record. This definition determined the mapping of the conversion from the publishers' to the library metadata and thus the information the cataloguers will have to deal with.

What could an ebook record within the catalogue of the BnF be like?

---

[6]. On the global workflow, see Sophie Derrot and Clément Oury, "Ebooks: rather electronic or book? Extending legal deposit to ebooks at the Bibliothèque nationale de France", *IFLA WLIC 2014* (16-22 August 2014, Lyon, France). [http://library.ifla.org/830/].

[7]. Commission de liaison interprofessionnelle du livre, http://clil.centprod.com/index.html.

To help answer this question, a specific working group on the bibliographic aspects was created from the beginning of the project, to study the needs of the ebooks workflow in this matter. Establishing the bibliographical context has indeed been useful to support other aspects of the project (human resources for cataloguing or preservation, for example). This working group still meets regularly, to follow evolutions on the subject.

The people participating in this working group reflect the complexity of the nature of ebooks: members of the Legal Deposit department (leader of the project and department of the cataloguers of deposited books), the Audiovisual department (which takes care of the legal deposit of multimedia documents) and the Digital and Bibliographic Information department (which is in charge of metadata consistency and standards compliance).

To define what an ebook record could be, this group had to face ontological questions: should the bibliographical description of an ebook be closer to that of a printed monograph or of an electronic document? One of the major elements of context was that these ebooks will be treated by cataloguers who deal with printed monographs. For the time being, the decision was made to treat these documents more as books than as electronic documents.

Once this choice had been made, the next question was how to integrate new metadata in the BnF model. On the one hand, some information doesn't yet exist within the General Catalogue of the library (a file's format and its versions, for example). On the other hand, some bibliographical uses already exist to describe the very diverse collection of the BnF and had to be taken into account: for instance, the notion of the weight of a document is declared in a specific field of our format, but that field is currently used to record the weight of a coin or medal, so another field had to be found for the "weight" of a file (its size).

Meanwhile, this reflection on the ideal ebook record allows innovations that are a driving force for printed books. For instance, when working with the publishers on the ONIX data model, we asked them to provide us with the European Article Numbering (EAN) of other versions of the same book, a field that exists in ONIX (`<RelatedProduct>`). Thanks to this information, a link is created between the different records during the automated conversion of ONIX into MARC records, if the catalogue already holds a record with the said EAN. The same will also be possible with authority records for authors, by means of the International Standard Name Identifiers (ISNI). Using the links between records for the same book in different formats, it will then be possible to retrieve all the relevant information, ranging from the original title to the link(s) to authority records and to the content type, from the most

complete record to automatically enrich the other one(s). Depending on which version of the book is deposited and catalogued first, this means not only that a record created for a printed book can be used to enrich a record for a PDF or an EPUB, but that the reverse will also be possible.


000 nam 22 3 450
001 FRBNF44241987 0000007
003 http://catalogue.bnf.fr/ark:/12148/cb44241987t
010 .. $a 978-2-258-10890-5 $d 14.99 EUR
035 .. $a 3012166100000978225810890 5
073 .0 $a 9782258108905
100 .. $a 20160212d2014 m y0frey50 ba
101 1. $a fre $c eng
102 .. $a FR
105 .. $a ||||z 00|a|
106 .. $a s $a r
135 .. $a dru|||||||||
200 1. $a Les filles du prophète $b Texte électronique $e roman $f Peggy Riley $g traduit de l'anglais (États-Unis) par Mélanie Blanc-Jouveaux
210 .. $a Paris $c Presses de la Cité $d 2014
215 .. $a 1 ressource dématérialisée
225 1. $a Romans Domaine Etranger
230 .. $a 2263128 octets
300 .. $a Notice rédigée d'après les métadonnées fournies par l'éditeur
307 .. $a Pagination restituée par l'éditeur : 290 pages
330 .. $a Au nom du père. Après un incendie criminel dans leur communauté, Amaranth, première des cinquante épouses du gourou d'une secte mormone fondamentaliste, décide de prendre la fuite avec ses deux filles, Amity et Sorrow. Elle est convaincue que son mari, dans un acte de folie, a tenté de tuer tous ses disciples en mettant le feu au domaine. Leur cavale connaît une fin abrupte au beau milieu de l'Oklahoma après un accident. Les trois fugitives échouent chez Bradley, un fermier désoeuvré et solitaire, qui accepte de les héberger. Ce qui ne devait être au départ qu'une solution de dépannage prend des allures de nouveau départ au fur et à mesure qu'Amaranth et Bradley se rapprochent. Si Amity, la cadette, s'accommode à merveille de la situation et découvre avec des yeux ébahis le monde moderne, Sorrow, quant à elle, vient de faire une fausse couche et n'a qu'une obsession : retourner dans sa communauté et retrouver sa place de prophétesse et préférée de son père? $2 éditeur
452 .1 $0 43784968 $t Les filles du prophète $b Texte imprimé $y 978-2-258-10352-8
454 .1 $t Amity and Sorrow
610 0. $a Anglais Américain Anglophone
610 0. $a Femme
610 0. $a Secte
700 .| $3 16753875 $a Riley $b Peggy $f 1965-.... $4 070
702 .| $3 14410717 $a Blanc-Jouveaux $b Mélanie $4 730
801 .3 $c 20160212 $b 3012166100000 $h 9782258108905 $c 20160212 $g onix/3.0
930 .. $5 FR-759999999:44241987001001 $a LNUM-107 $b 759999999 $c Document numérisé $d N $v 2016/02/12

Example of an ebook record in UNIMARC format with information retrieved from a printed book record: elements of the Title (200), Original title (454), Author(s) (200$g, 700, 702) and Summary (330) fields come from the printed book record. The following fields, when present, can also be retrieved: Dissertation note, Uniform title, General note, Reproduction note, With note, Audience characteristics, Subject access fields and Summary, as well as part of the Control field.

Consequently, records of different levels of completeness and standard compliance will coexist within the catalogue. It is a common situation, but it will be particularly intricate because it won't be a simple case of printed and electronic books having different quality records. Some ebooks will have richer records because of the data retrieved from a pre-

existing printed book record, and the reverse might be true too. Thus, the very process will induce such a symbiosis between records that we won't be able (should we want it) to separate workflows for printed and electronic books.

The overlap also exists in terms of standards: the French standard applicable to ebooks is NF Z 44-082 *Catalogage des documents électroniques*, which dates back to 1999. It only gives indications about recording the specific characteristics of electronic documents, the rest of the resource being described under the specifications of the relevant standard (printed books, maps, serials, etc.). Of course, that standard, which predates the rise of ebooks, is now outdated. That is another point we had to take into account to devise the cataloguing framework for ebooks.

"Within a catalogue and beyond"[8]

Our first response to the question "How do we let ebooks into the catalogue?" was expressed in terms of format, because it is how we catalogue, in a most concrete way. But the answer lies with standards.

There is no set of standards specifically dedicated to ebooks in France today. The rules are derived from those for printed books and for electronic material, both French cataloguing rules and ISBD. The issue of updating French cataloguing rules is broader than that of ebooks, especially in the global context of switching to RDA. France has had a working group on the adoption of the latter cataloguing rules since 2010. In late 2014, both national bibliographic agencies[9] decided on a policy of converging towards RDA: adopting as much of the code as possible, without compromising our national practices where they are richer or more conform to FRBR than RDA currently is.

Under the new name "Bibliographic transition", the group is now tasked with updating French standards into what we call RDA-FR. As it is a very long process, parts of the new standard are published and gradually replace the corresponding parts of French standards. It means that we are currently working with two standards at the same time. But RDA's consolidated

---

[8]. International Federation of Library Associations and Institutions, Barbara Tillett and Ana Lupe Cristán, *IFLA cataloguing principles: The Statement of International Cataloguing Principles (ICP) and its glossary : in 20 languages* (München: K.G. Saur, 2009). [http://www.ifla.org/files/assets/cataloguing/icp/icp_2009-en.pdf]

[9]. The Agence bibliographique de l'enseignement supérieur = Bibliographic Agency for University Libraries (ABES), and the Bibliothèque nationale de France.

approach, with general rules that encompass all types of documents, could be useful toward a convergence of records.

For instance, inside the National Library, parallel to our project on the legal deposit of ebooks, there is a distinct project for ebooks coming into the collections by other acquisition modes, and the Audiovisual department also handles digital documents. We share the need to describe system details for an electronic resource. We worked on the in-house INTERMARC format while discussing at a national level on RDA-FR. As fast answers are needed for these projects, the standard is discussed and the format updated in very quick succession, with to and fros because, contrary to the way standards tend to be written, the use cases happen almost simultaneously and we can put our questions directly to those who write them. That is how the INTERMARC field 257 evolved over the course of one year. It used to be specific to the audiovisual collections, to describe the type and size of an electronic resource and had only three subfields, for size, type and linking words. During the past year, we added eight subfields, cancelled one and completely changed the field's structure, so as to use it for every type of electronic resource. It took months, because it was not a case of creating a new field just for one purpose, but taking an existing one and evolving it to the point where it could be used for all types of documents and cataloguing practices.

Experiences like this made us realise early on that one of the biggest hurdles we librarians face is the nature of ebooks. We tend to see them as books first and foremost, and are tempted to deal with them as we do with printed legal deposit material: create reference-level records for all ebooks, and publish them in the National Bibliography.

However it is not what the law tell us. Ebooks fall within the scope of the law on digital legal deposit, which was created with the web in mind. It means no obligation for the National Library of instituting a comprehensive legal deposit for ebooks, or of publishing a national bibliography. But our whole reflection on ebooks led us to emphasise the convergences rather than the differences between printed and electronic books, and to create a descriptive model closer to printed books than the web.

Furthermore, we know that there is a need for a comprehensive, open national register or database for ebooks, with reference-level, trustworthy records; this was expressed in the Lescure report to the minister of Culture in 2013[10]. The commercial database Electre began

[10]. Pierre Lescure, *Contributions aux politiques culturelles à l'ère numérique* (Rapport à la ministre de la Culture, 2013). [http://www.ladocumentationfrancaise.fr/rapports-publics/134000278/index.shtml]

providing metadata for ebooks lately. But the BnF has a unique position of observer of the publishing world, both printed and digital. Moreover, the very principle of producing individual records for ebooks in the library's catalogue seems an argument in favour of taking the next step and publishing them in the French National Bibliography.

The question of course is how?

In terms of quality, the National Library is a service provider to a national community, especially through the National Bibliography, which, unlike the General Catalogue, traditionally holds only records of the highest quality. But the *Guidelines for bibliographies in the electronic age*[11] have identified the problem and accepted that, with ebooks, different levels of description are acceptable. It remains for us to determine what level of description to adopt, bearing in mind that our process makes the records for ebooks and printed books tightly interdependent.

In terms of display, should records be displayed in the books section of the bibliography, mixed with printed books, with only a dedicated title index? In terms of IT developments it would be the quickest and simplest solution. It might be a temporary one though, until a new section of the bibliography or a new website for the whole publication can be created. If we decide on updating the current website, what form would it be better for it to take? Once we have (potentially) several versions of the same title, a "FRBR" display might be considered. A look at the other websites the BnF uses to display metadata would be the first step of that particular study.

Conclusion

Once we look past the first impulse of doing just the same as we have "always" done, we must ask ourselves what service is expected from us. With ebooks, the context of legal deposit changes: ebooks carry with them formatted and standardised metadata. Even if the standards are not those we librarians use, we are provided with more than basic records that already provide a service to users. But is it the service we need to provide? At this point, it looks like a very good step in the right direction, but with still some way to go to be quite up to standards. So how do we close this gap? Automation will be a great help, retrieving data to

[11]. Maja Žumer (ed.), *National bibliographies in the digital age: guidance and new directions* (München: Saur, 2009). [http://public.eblib.com/choice/publicfullrecord.aspx?p=454007]

have the same level of description between records for the same title, but it has its limits. That is where human intervention is necessary. This human intervention will probably be necessary during the early phase of the treatment as well, to control the incoming ebooks and their metadata. We are still in the testing phase, but we now deal with real metadata provided by our publisher partners. Our reflection gains from observing these metadata in our cataloguing environment. New questions come out almost every week and we need to think creatively, which makes the building of ebooks cataloguing very stimulating.

This project is highlighting the convergence of questions libraries are confronted with: the adoption of RDA, the revision of standards, the automation of cataloguing. To answer these questions, we have to take a more holistic view of our activity, transcending the library's sometimes rigid structure.