Livre numérique accessible et numérisation de masse à la BnF : retour d'expérience

Jean-Philippe Moreux, *Bibliothèque nationale de France*

***Résumé court* — Pour une meilleure diffusion des contenus numériques, la BnF a choisi le format EPUB dans le cadre de ses programmes de numérisation, et en particulier la version 3 qui lui a permis d'améliorer l'accessibilité des contenus numériques. Deux ans après ce choix, quels sont les enseignements à tirer de sa mise en œuvre à grande échelle ?**

INTRODUCTION

Les bibliothèques de prêt s'investissent de plus en plus dans le développement de leurs offres en lecture numérique, et il est légitime pour une bibliothèque patrimoniale d'accompagner cette tendance en travaillant sur ses propres collections.

Depuis 2011, la Bibliothèque nationale de France a donc choisi de renforcer le service offert en matière de lecture numérique grâce à l'utilisation du format EPUB. L'objectif est de tirer profit des avantages de ce dernier en comparaison des formats de diffusion généralement offerts par les bibliothèques numériques (HTML ou PDF), soit notamment la consultation nomade d'un format conçu pour la lecture numérique, sur des appareils dédiés.

Entre 2011 et 2013, cet effort s'est incarné sous la forme de programmes de numérisation de masse avec intégration d'un sous-processus de production EPUB 2, et de programmes de retraitement de documents déjà numérisés avec rétroconversion EPUB 2.

Le programme en cours (2014-2017) s'appuie sur le retour d'expérience des précédents et amplifie leurs ambitions, à travers le choix de la version EPUB 3, pour offrir une meilleure accessibilité aux contenus numériques aux personnes empêchées de lire du fait d'un handicap.

Aujourd'hui, la BnF produit environ un millier d'EPUB par an et Gallica en propose plus de 3000 au téléchargement.

I. POURQUOI EPUB ?

Dès 2010, EPUB s'est imposé comme le format ouvert standard pour la diffusion d'ouvrages numériques. Basé sur les formats techniques du Web, il offre une excellente garantie en matière d'interopérabilité et de préservation. Comparé au format PDF, il présente l'avantage d'un meilleur confort de lecture, du fait de son mode opératoire par flot de texte, qui permet d'adapter les contenus tant au lecteur (grossissement des caractères), qu'aux caractéristiques du dispositif de lecture (taille d'écran).

En 2013, afin de répondre aux obligations légales en matière d'accessibilité aux contenus numériques, la version EPUB 3 a été choisie pour ses nombreux mécanismes dédiés :

- table de navigation enrichie,
- structuration sémantique des contenus,
- description du niveau d'accessibilité avec des métadonnées ONIX.

Le risque lié à la lecture de contenus EPUB 3 sur un parc de logiciels de lecture et de liseuses encore en partie compatibles EPUB 2 a été évalué comme étant faible. En effet, les EPUB 3 patrimoniaux n'utilisent pas de mécanismes (mise en page de type *fixed layout*, interactivité) ou de contenus (multimédia) à risque. De plus, une transition générale des dispositifs de lecture d'EPUB 2 vers EPUB 3 est engagée depuis 2014, (même si elle demeure incomplète).

Parallèlement à la production d'EPUB 3, il a été décidé de générer un fichier adaptable XML DAISY afin de faciliter la production d'autres formats accessibles et d'accompagner la transition DAISY vers EPUB. Ce fichier s'appuie sur la DTD XML DTBook 2005-3, mais ce choix pourra évoluer selon l'actualité technique du domaine (par exemple ZedAI).

II. PRODUIRE DES EPUB EN NUMERISATION PATRIMONIALE

La mise en œuvre de ce nouveau format a nécessité une étude détaillée et de nombreuses interactions entre toutes les équipes de la bibliothèque impliquées dans la numérisation des contenus patrimoniaux.

A. Sélection documentaire

Pour des raisons de coût, tous les ouvrages numérisés ne peuvent être convertis en livre numérique. Un choix doit donc être fait, outillé par des critères de sélection : le conservateur devient un éditeur. Concilier la diversité d'un fonds patrimonial, les limitations techniques du format comme des dispositifs de lecture et le manque de souplesse d'un programme de numérisation de masse peut s'avérer un exercice délicat…

Des critères techniques sont également définis, liés par exemple à la qualité de l'OCR d'origine, à la langue (français contemporain, classique, ancien français) ou encore à la qualité des images source, afin d'orienter les documents vers des filières de production adaptées (et des coûts associés).

B. *Ingénierie des contenus*

Un modèle « EPUB patrimonial » apte à recevoir une grande variété de genres documentaires a été élaboré, ainsi qu'un référentiel de production EPUB [1], qui définit les règles de transformation d'un document numérique patrimonial « classique » (composé d'un manifeste, d'images et d'OCR) en EPUB :

- *mapping* des métadonnées bibliographiques vers les métadonnées EPUB,
- *mapping* des « objets » OCR vers le balisage HTML,
- règles de création des éléments constitutifs de l'EPUB : couverture, page de titre, table de navigation.

C. *Contrôle qualité*

Le système de contrôle automatique des documents numériques reçus par la BnF a dû évoluer, afin d'analyser ce nouveau format relativement aux exigences définies dans le référentiel de production.

Par ailleurs, une équipe dédiée au contrôle qualité EPUB a été mise en place :

- contrôle visuel sur échantillonnage, à l'aide d'un parc d'appareils et de logiciels de lecture,
- évaluation de la qualité des textes retranscrits.

D. *Archivage et préservation à long terme*

Les fichiers EPUB et XML DTBook sont contrôlés et caractérisés avant ingestion dans SPAR, le système de préservation numérique de la BnF [2].

E. *Diffusion*

Les livres numériques sont mis à disposition des internautes dans la bibliothèque numérique Gallica. Un critère de recherche sur la disponibilité de ce format a été ajouté.

F. *Coûts de production*

La montée en qualité du texte est un facteur de coût majeur : les contenus patrimoniaux océrisés doivent être amenés à un taux qualité de type éditorial (99,9 % à 99,95 % au mot, selon les filières de production liées à la difficulté de la conversion, cf. §II.A).

On peut estimer que cette nécessaire correction du texte et le surcroit d'ingénierie conduisent à un facteur ×3 à ×4 entre une production OCR qualité garantie et EPUB ; et ×10 entre OCR brut et EPUB.

Enfin, des enjeux de budget imposent certaines limitations quant au périmètre des contenus en conversion EPUB :

- pas de documents multilingues,
- pas de contenus scientifiques,
- pas d'index avec renvois actifs.

III. Produire des contenus accessibles en numerisation patrimoniale

A. *Ingénierie des contenus*

La mise en œuvre des mécanismes d'accessibilité offerts par EPUB 3 a été l'objet d'une étude spécifique, en regard des contraintes spécifiques de tout programme de numérisation de masse :

- ajout d'une couche de structuration logique à l'aide du balisage HTML 5 et de l'annotation sémantique EPUB 3 (`epub:type`),
- création de la table de navigation enrichie (liste des pages, liste des repères),
- mais…
  - pas de description des illustrations,
  - pas de typage de la langue des mots isolés.

La production du format adaptable s'appuie sur un nouveau référentiel de numérisation [3], qui exprime les règles de *mapping* des contenus EPUB 3 vers le format XML DTBook 2005-3. On pourra d'ailleurs noter que les vocabulaires DTBook et EPUB 3 ne sont pas parfaitement alignés et qu'ils sont lacunaires pour certains genres documentaire (théâtre, poésie, sciences, etc.).

B. *Contrôle qualité*

Le contrôle automatique des EPUB a été mis à jour, afin de permettre la gestion des deux versions actives du format et de leurs caractéristiques propres.

L'équipe dédiée au contrôle qualité EPUB a pris en main les nouveautés EPUB 3 : contrôle visuel des nouvelles tables de navigation et de la structuration logique.

Le format DTBook 2005-3 est vérifié à l'aide de l'outil `pipeline`.

C. *Diffusion*

La recherche de contenus accessibles sera possible dans la prochaine version du site Gallica (septembre 2015).

D. *Coûts de production*

On peut estimer à environ 5% le surcoût lié à la mise en œuvre des mécanismes d'accessibilité. Il s'agit essentiellement du surcroit d'ingénierie et de travail opérateur nécessaires à la structuration logique des contenus : typage fin des éléments de contenu au niveau macro (avant-propos, remerciements, achevé d'imprimer) ou micro (épigraphe, poème, etc.) et la détection de la structure de l'ouvrage (liminaires, corps, annexes).

Le format adaptable DTBook étant généré par *mapping*, il ne suscite donc pas de coût variable.

Au final, le prix moyen à la page est de 0,4 à 0,8 € pour une prestation en contenus accessibles incluant :

- OCR,
- correction du texte,
- livraison EPUB 3, XML DTBook, XML ALTO (OCR avec texte corrigé).

Conclusion

Avec ce programme, la BnF a prouvé qu'il est possible de produire des contenus numériques nativement accessibles avec un surcoût quasi négligeable. Contenus parfaitement lisibles par les deux générations de dispositifs de lecture (seuls les dispositifs compatibles EPUB 3 profitant des mécanismes

d'accessibilité).

Les futurs programmes de numérisation menés par la BnF ou par sa filiale BnF-Partenariat ont vocation à s'appuyer sur le savoir-faire acquis en matière de livre numérique accessible. La plupart des ebooks produits par ces programmes seront donc des EPUB 3 accessibles et accompagnés d'un format adaptable. C'est notamment le cas de la production du projet Relire/Indisponibles (2014-2024) : ± 500 000 ouvrages, majoritairement au format EPUB.

BIBLIOGRAPHIE

[1] « Référentiel EPUB 3 », version 1.
http://www.bnf.fr/documents/ref_num_epub3.pdf
[2] Sophie Derrot, Jean-Philippe Moreux, Stéphane Reecht, et Clément Oury, "Preservation of ebooks: from digitized to born-digital", in *Proceedings of the 11th International Conference on Digital Preservation* (iPRES), Melbourne, Australia, 2014.
[3] « Référentiel DAISY », version 1.
http://www.bnf.fr/documents/ref_num_daisy.pdf

BIOGRAPHIE

Jean-Philippe Moreux est l'expert OCR et formats éditoriaux du service de la numérisation de la BnF (département de la Conservation). Il travaille sur tous les programmes de numérisation patrimoniale de la BnF. A ce titre, il participe notamment aux actions de production de livres numériques, ainsi qu'aux projets de recherche dont la bibliothèque est partenaire. Il est membre du comité éditorial ALTO. Ingénieur de formation, il a été chef de projet dans une SSII, éditeur scientifique et consultant (ingénierie éditoriale, édition numérique).