DigItalia

Anno III, Numero 1 - **2008**

Rivista del digitale nei beni culturali

ICCU-ROMA

Une contribution française à la bibliothèque numérique européenne: Europeana et la Bibliothèque nationale de France

Elizabeth Freyre – Emmanuelle Bermès
Bibliothèque nationale de France

La Bibliothèque nationale de France (BnF) a été créée en 1994 à partir de la fusion de l'ancienne Bibliothèque nationale, héritière de la bibliothèque des Rois et de l'Etablissement public pour la Bibliothèque de France.
C'est une bibliothèque encyclopédique. Elle reçoit le dépôt légal; elle acquiert et échange des documents et bénéficie de dons. Sur place, elle met à la disposition de son public, selon la volonté du président François Mitterrand, promoteur de ce nouveau site, une bibliothèque de recherche et une bibliothèque d'étude. À distance, tout internaute peut accéder à de nombreux services qui vont de l'accès au catalogue à la consultation d'expositions virtuelles et de dossiers pédagogiques en passant par la réservation de places ou de documents par les chercheurs sans oublier la bibliothèque numérique: Gallica[1].
Avec Gallica, conçue au début des années 1990 et mise en ligne en 1997, la BnF a acquis une avance reconnue dans le domaine du numérique en offrant en ligne plus de 100.000 documents (livres et périodiques) et 90.000 images, libres de droit. Elle a pris une importance encore plus grande à partir de 2004. En effet, lorsque à la fin de l'année, Google annonça son intention de numériser quinze millions d'ouvrages, Jean-Noël Jeanneney, alors président de la BnF, ouvrit un débat[2] qui suscita de multiples réactions des médias, des internautes et du grand public, relayées par plusieurs déclarations au plus haut niveau de l'État français et de l'Union européenne.
En France, un Comité de pilotage «pour la création d'une Bibliothèque numérique européenne» fut créé afin de dresser un état des lieux, de confronter les différents points de vue, de dégager des orientations et d'énoncer quelques propositions consensuelles regroupées dans un Livre blanc[3].

[1] http://gallica.bnf.fr.
[2] Jean-Noël Jeanneney, *Quand Google défie l'Europe*, «Le Monde», 24 janvier 2005, p. 13; Jean-Noël Jeanneney, *Quand Google défie l'Europe Plaidoyer pour un sursaut*, Paris: Mille et une nuits, 2005.
[3] http://www.bnf.fr/PAGES/dernmin/pdf/bnue.pdf.

En mai 2006, le gouvernement a confié explicitement à la BnF la responsabilité opérationnelle du projet pour la France et l'a dotée d'un budget spécifique de 3,5 millions d'euros.

La contribution française à la Bibliothèque numérique européenne

Pour remplir cette mission, la BnF s'est appuyée sur l'expérience qu'elle avait acquise avec Gallica et sur ses équipes afin d'accomplir:

- la réalisation d'une maquette puis d'un prototype appelé Europeana donnant un accès plein texte à des documents de la BnF, des bibliothèques nationales de Hongrie et du Portugal et explorant des services collaboratifs qui pourraient être mis à la disposition des internautes. Le nom Europeana a ensuite été cédé début 2008 à la Fondation European Digital Library (EDL) pour le développement de son propre prototype réalisé dans le cadre du réseau thématique européen EDLnet (European Digital Library network);
- la modernisation technique de sa propre bibliothèque numérique, Gallica: passage du mode image au mode texte, mise en ligne de plusieurs titres de la presse quotidienne numérisés, consultables de manière plus aisée grâce à un outil adapté;
- l'intensification des programmes de numérisation d'œuvres issues du fonds de la BnF: le rythme de numérisation passe, à partir de 2007, d'environ 6.000 à 100.000 ouvrages par an.

Pour mener à bien la réalisation de ces trois objectifs, la bibliothèque nationale de France a adopté une démarche orientée utilisateurs et s'est dotée d'une infrastructure technique favorisant l'interopérabilité et la préservation des données numériques produites.

Une démarche orientée utilisateurs

La réalisation de la maquette puis du prototype Europeana par la BnF, de fin 2006 à mi 2007, s'est appuyée sur une démarche fonctionnelle mettant l'usager au premier plan. Il s'agissait de développer un site grand public, ancré dans les pratiques du Web d'aujourd'hui telles que la recherche simple à la façon des moteurs de recherche du Web, la personnalisation ou encore l'annotation collaborative à travers des mots-clefs (*tags*) ou des commentaires qu'il est possible de partager dans des groupes d'intérêt. Pour autant, il ne fallait pas laisser de côté les aspects plus spécifiques à une bibliothèque: l'organisation de l'information suivant une ou plusieurs classifications, les métadonnées structurées, la recherche avancée suivant des critères bibliographiques (auteur, titre, etc.) Tous ces aspects étaient représentés dans la première maquette Europeana, qui a été présentée en novembre 2006. Au moment de les mettre en œuvre, des priorités ont toutefois dû être définies.

Pour cela, des études d'usages ont été réalisées auprès de chercheurs, d'internautes et d'usagers de Gallica: la première étude, suivant la méthodologie des *focus groups*, réunissait des groupes d'usagers types pour commenter la maquette, tandis que la seconde étude a été réalisée sous forme de questionnaire mis en ligne en même temps que le prototype, en mars 2007[4]. Il a ainsi été possible de délimiter des fonctionnalités essentielles, qui ont été développées dans le prototype Europeana (accessible de mars 2007 à janvier 2008) puis dans la nouvelle version de Gallica, Gallica 2[5].

Les aspects techniques: d'Europeana à Gallica 2

Tandis que le prototype Europeana était fermé en janvier 2008 au moment où la Fondation EDL reprenait le nom "Europena" pour sa propre maquette qui porte le projet à une dimension européenne et l'élargit à d'autres acteurs que les bibliothèques (archives, musées, etc.), la BnF a poursuivi son effort de réalisation technique pour faire évoluer Gallica. Le nouveau site actuellement en version bêta, Gallica 2, prend en compte les paramètres techniques définis dans la phase précédente et s'enrichit progressivement de nouvelles fonctionnalités. Fin 2008, il sera appelé à remplacer complètement l'ancien site Gallica.

Le protocole OAI-PMH (Open Archive Iniziative – Protocol Metadata Harvesting) a été utilisé pour récupérer et indexer l'ensemble des métadonnées, aussi bien dans la phase du prototype que dans Gallica 2. En effet, ce protocole permet de mettre à disposition des moteurs de recherche du Web des descriptions de documents dans un format bibliographique simple et largement partagé par les bibliothèques en Europe: le Dublin Core. Pour les documents de Gallica, les métadonnées au format Dublin Core simple, sont extraites du catalogue de la BnF pour être versées dans l'entrepôt OAI de Gallica[6]. Dans le prototype, c'étaient également des métadonnées au format Dublin Core moissonnées en OAI qui étaient utilisées pour indexer les documents des bibliothèques nationales de Hongrie et du Portugal. Le protocole OAI permet de récupérer automatiquement ces métadonnées avant de les intégrer dans l'interface commune. Ensuite, la consultation du document s'effectue sur le site d'origine via l'utilisation d'un lien hypertexte pérenne.

L'indexation, réalisée grâce au moteur de recherche open source Lucene, porte à la fois sur les métadonnées descriptives des documents et sur le plein texte[7]. Le moteur Lucene est un moteur plein texte qui permet de faire de simples recherches par mots-clefs, mais aussi d'exploiter la puissance des métadonnées structu-

[4] Les résultats de ces études sont accessibles en ligne: http://bibnum.bnf.fr/usages/index.html.
[5] http://gallica2.bnf.fr.
[6] Voir: http://bibnum.bnf.fr/oai/index.html.
[7] Sur ce sujet, voir Emmanuelle Bermès, *Les moteurs de recherche: petit précis de mécanique à l'usage des bibliothèques numériques*, «Bulletin des bibliothèques de France», tome 52, 2007, n. 6, p. 5-10.

rées par un affinage par facettes ou par une interface de recherche avancée: l'utilisateur peut ainsi préciser, pour une recherche, si elle porte sur l'auteur, le titre, la table des matières, etc., mais il peut également, après avoir effectué une première recherche, préciser sa demande en sélectionnant parmi les facettes proposées un auteur, un thème, un type de document, une période ou un autre critère de son choix. Cette possibilité, appelée navigation à facettes, a été utilisée pour proposer un accès thématique aux collections: basé sur l'indexation des documents suivant la classification décimale Dewey, largement répandue dans les bibliothèques, cet accès thématique permet aux utilisateurs d'accéder directement à des listes de documents classés par grands thèmes de la connaissance, sans avoir à formuler au préalable une requête précise. Cette fonctionnalité est essentielle pour permettre aux internautes, qui découvrent la bibliothèque numérique, de se faire une idée de la teneur des collections disponibles.

L'une des nouveautés importantes, par rapport à l'ancien site Gallica, était également l'indexation plein texte portant sur le contenu même des documents. Pour la production du texte à indexer, le choix a été fait d'utiliser une technologie de reconnaissance optique de caractères (OCR). Celle-ci a été réalisée par l'intervention d'entreprises prestataires de service, soit de façon rétrospective sur une partie significative de la collection de Gallica déjà numérisée en mode image, soit de façon courante, l'OCR étant produit conjointement avec la numérisation image pour tous les nouveaux documents numérisés dans le cadre des nouveaux marchés de numérisation de masse. Le résultat obtenu, encodé en XML grâce à un format spécifique dénommé ALTO, est utilisé pour l'indexation mais aussi pour l'affichage du mode texte lorsque le niveau de qualité atteint est suffisant, et la génération d'une version téléchargeable au format PDF multicouche[8].

Le choix de protocoles et de formats normalisés, comme l'OAI-PMH, ALTO ou encore le TIFF pour les images numérisées, est une des briques essentielles de la contribution de la BnF à la bibliothèque numérique européenne. En effet, ces formats et ces standards sont la garantie que les documents produits seront réutilisables dans un contexte plus large. La problématique de la conservation sur le long terme de tous ces fichiers numériques n'a pas non plus été laissée de côté. Au contraire, elle est prise en compte dès l'amont par le choix de formats adaptés (TIFF non compressé pour les images, XML ALTO pour les fichiers issus de l'OCR) et la production de métadonnées spécifiques qui seront ensuite réexploitées dans le système de préservation numérique de la BnF. Celui-ci, nommé SPAR (Système de préservation et d'archivage réparti) est en cours de développement et permettra d'archiver de manière pérenne dès 2009 l'ensemble des documents numériques produits au titre de la numérisation.

[8] Le PDF multicouche est un format numérique qui permet d'inclure dans le même fichier les pages numérisées en mode image, et leur transcription en OCR associée par transparence.

La coopération nationale avec les autres partenaires de la chaîne du Livre

Au niveau national, la BnF a mis l'accent dans deux directions complémentaires: en sollicitant les grandes bibliothèques universitaires ou territoriales, notamment le réseau des pôles associés de la BnF, pour contribuer à la réflexion autour des priorités documentaires guidant la numérisation; en engageant avec le Syndicat National de l'Edition (SNE) et quelques grands éditeurs, des discussions afin de permettre l'accès à des ouvrages contemporains dans le respect du droit d'auteur.
Un groupe conjoint constitué de la BnF et du SNE a confié une étude à la société NUMILOG[9] afin de proposer de nouveaux modèles économiques. La faisabilité du modèle économique retenu est actuellement testée dans le cadre d'une expérimentation d'un an, au cours de laquelle les documents sous droits sont intégrés à Gallica 2 par l'intermédiaire de e-distributeurs. Ceux-ci mettent à disposition de la BnF des métadonnées moissonnables par le protocole OAI-PMH. Il est ainsi possible de récupérer non seulement les descriptions des ouvrages, mais aussi leur première de couverture qui sert d'illustration, et une version plein texte utilisée exclusivement pour l'indexation de tous les mots dans Gallica 2. Ainsi, un utilisateur qui fait une recherche dans Gallica 2 a la possibilité de trouver dans les listes de résultats aussi bien des documents libres de droits, issus de la numérisation réalisée par la BnF, que des documents d'éditeurs accessibles sous conditions via les sites des e-distributeurs. Bien que la recherche s'effectue dans Gallica 2, la consultation des documents eux-mêmes se fait sur le site des e-distributeurs, ce qui permet un accès à l'information souple et respectueux des droits d'auteurs. Actuellement, une dizaine de e-distributeurs sont partenaires de la BnF pour cette expérimentation.

La coopération européenne

Au niveau européen, la BnF met à disposition son expérience et sa bibliothèque numérique Gallica afin d'œuvrer aux côtés de l'ensemble des institutions culturelles européennes (bibliothèques, archives, musées et institutions audiovisuelles) à l'édification d'Europeana, la Bibliothèque numérique européenne qui permettra d'accéder en ligne à plusieurs millions de documents numérisés représentatifs du patrimoine historique et culturel de l'Europe.
La BnF participe ainsi à différents projets européens:

- EDLnet qui va développer au cours de 2008, une maquette puis différents prototypes d'Europeana;
- TELplus (The European Library plus) qui étudie les questions d'OCR, d'indexation plein texte de bibliothèques numériques ainsi que de nouveaux services collaboratifs à offrir aux internautes;

[9] http://www.bnf.fr/pages/catalog/pdf/EUROPEANA-NUMILOG2007.pdf.

– IMPACT (IMProving ACces to Text) qui se concentrera dans les 4 années à venir sur une meilleure reconnaissance des textes grâce à de nouvelles pratiques d'OCRisation;
– ARROW, projet conjoint bibliothèques/éditeurs qui développera une base des œuvres orphelines et étudiera l'accès des œuvres numériques contemporaines via Europeana dans le respect des droits d'auteur.

Grâce à ces différents projets, la BnF poursuit son effort de contribution à la bibliothèque numérique européenne, en partageant son expertise et en mettant à disposition les documents numérisés pour Gallica et Gallica 2. La démarche de développement orientée utilisateurs, l'utilisation de formats et de protocoles standards pour les métadonnées, et le souci de la préservation à long terme, partagés au niveau européen, constituent le ciment de cette bibliothèque numérique européenne dont la BnF est partie prenante.