Des outils automatiques pour le signalement en bibliothèque :

→ EXPÉRIMENTATIONS AUTOUR DU PROJET *DATA.BNF.FR*

Romain Wenz
romain.wenz@bnf.fr

Agnès Simon
agnes.simon@bnf.fr

Bibliothèque nationale de France

Archiviste-paléographe, conservateur des bibliothèques, **Romain Wenz** *a rejoint en 2009 le département de l'Information bibliographique et numérique de la BnF comme expert métadonnées. Il travaille à l'élaboration de l'outil* data.bnf.fr.

Conservateur des bibliothèques, **Agnès Simon** *travaille au département de l'Information bibliographique et numérique de la Bibliothèque nationale de France, où elle est responsable adjointe du projet* data.bnf.fr.

Le développement des outils informatiques, et en particulier des documents numériques, est vécu comme un changement fort par les professions de l'information et les bibliothèques. Ces changements ne concernent pas uniquement les métiers de l'information au sens strict, mais plus généralement toutes les activités de services qui requièrent de manipuler des documents écrits.

Dès que l'informatique a commencé à gagner le grand public, le secteur tertiaire a changé de visage. Dans la grande distribution comme en bibliothèque, on travaille avec un ordinateur. Qu'il s'agisse de vendre des produits alimentaires ou de communiquer des livres, le catalogue se gère derrière un écran.

L'utilisation de l'ordinateur comme outil quotidien a constitué un important changement dans la vie professionnelle. Les apports de l'automatisation dans le traitement des informations sont un vecteur de transformation moins visible mais non moins conséquent. Les capacités des machines informatiques ont évolué dans l'ensemble des professions de services : de plus en plus de possibilités techniques sont offertes, en particulier grâce aux outils dits du « web sémantique [1] », qui permettent d'exposer des données reliées, structurées avec rigueur, et réutilisables par d'autres. Pour les bibliothèques, ces outils peuvent être utiles, en particulier sur trois axes :

- l'identification des ressources dans des masses considérables ;
- la gestion de formats variés ;
- l'échange de données.

Trois questions que la technique est actuellement en train de renouveler.

La Bibliothèque nationale de France déploie progressivement, depuis l'été 2011, un projet qui s'appuie sur les outils du « web sémantique » : *data.bnf.fr*.

Le site permet de signaler les ressources de la BnF et de les rendre plus visibles sur internet. Construit automatiquement, à l'aide d'algorithmes, il a pour but d'expérimenter des possibilités techniques, mais aussi d'observer les usages réels des internautes, au travers de pages web visibles par tous. La bibliothèque intègre ce projet dans une activité prospective, de recherche et d'innovation pour les évolutions futures.

1. Le web sémantique est un ensemble de technologies visant à faciliter l'exploitation des données structurées, en permettant leur interprétation par des machines. Le web de données (Linked Data en anglais) combine les technologies du web sémantique et les principes fondamentaux du web (protocole HTTP, identifiants URI), avec pour objectif la construction d'un réseau d'informations structurées, disponibles en ligne et facilement réutilisables dans de nombreux contextes. www.bnf.fr/fr/professionnels/web_semantique_donnees/s.web_semantique_intro.html

Le travail de conception : *bibliothecarius ex machina*

La diffusion massive des outils informatiques a provoqué un changement rapide des pratiques des lecteurs. Pour s'adapter, les bibliothèques doivent s'insérer dans un monde d'informations beaucoup plus concurrentiel qu'avant le développement de l'internet. Cela implique d'organiser des masses d'informations considérables de façon à ce qu'elles deviennent exploitables sur le web et visibles : à partir de millions de documents (plus de onze millions de notices pour le Catalogue général de la BnF), il faut réussir à regrouper les informations autour de concepts. Les sites d'information qui sont actuellement les plus consultés sur internet, à commencer par Wikipédia, rassemblent par exemple les informations sur les œuvres littéraires ou artistiques, au niveau du « concept » : la page web correspond donc à l'œuvre, au sens intellectuel, autour de laquelle sont regroupées ses différentes versions. Cette organisation de l'information est simple et intuitive, faite pour le grand public ; elle correspond à ce que les professionnels des bibliothèques nomment le modèle « FRBR[2] ».

En bibliothèque, il est particulièrement intéressant de faire ce type de regroupement lorsque l'on conserve à la fois des manuscrits, des éditions imprimées et des versions numérisées d'une même œuvre, qui peuvent être décrits dans des bases et des formats différents : leur regroupement évite à l'utilisateur d'avoir à connaître et interroger plusieurs catalogues.

La numérisation, avec la diffusion de documents consultables partout et à tout moment, amène une nouvelle vision du signalement. Le patrimoine numérisé actuellement disponible sur internet a atteint un volume tel que la difficulté principale devient l'accès aux fonds numérisés. Pour permettre aux lecteurs de trouver ce qu'ils recherchent, et même pour éviter de numériser en double, la qualité de la description bibliographique des documents devient un enjeu primordial. Il s'agit de faciliter l'accès à un public plus large, qui utilise les ressources des bibliothèques de manière imprévisible. Par exemple, l'enluminure médiévale peut intéresser non seulement les lecteurs de manuscrits, mais aussi un public de curieux, dans un contexte de loisir.

À la BnF, le projet *data.bnf.fr* regroupe et expose des données issues de différents catalogues (livres, archives, manuscrits) et de la bibliothèque numérique Gallica. Ces données permettent de créer des pages sur les auteurs, les œuvres et les thèmes, qui rassemblent les liens vers toutes les ressources disponibles à la BnF. Tout en inventant une manière de mettre en application les principes « FRBR » de façon innovante, ce projet rejoint donc des principes fondamentaux du métier : fournir des contenus, des liens vers des documents, et offrir un service simple à trouver et à utiliser.

Les outils et formats, créés spécialement pour les bibliothèques, ont désormais à s'intégrer dans le monde plus vaste de l'internet, régi par les standards mondiaux du W3C[3]. La bibliothèque doit désormais être non seulement *sur* le web, mais aussi *dans* le web. Autrement dit, si les catalogues et bibliothèques numériques sont déjà en ligne, il s'agit aujourd'hui d'intégrer leurs données dans l'écosystème du web. *Data.bnf.fr* rend exploitables par des machines les données de la bibliothèque, jusque-là cloisonnées et spécifiques. Cela a une double conséquence : d'une part, l'internaute n'est plus obligé de connaître, *a priori*, les différentes bases de recherche de la BnF, mais retrouve directement la ressource pertinente sur le web, en passant par les moteurs de recherche. D'autre part, les données, exposées sur le web sémantique, peuvent être récupérées, liées, et réutilisées de manière inédite. Car l'ouverture technique a été confortée par une ouverture juridique. Pariant sur l'« Open Data[4] », la BnF a placé les données de *data.bnf.fr* sous Licence ouverte de l'État[5], ce qui les rend utilisables librement, à condition de mentionner la source BnF.

Ainsi exposées sur le web, les bibliothèques sont au cœur d'un environnement concurrentiel, en particulier en ce qui concerne les documents numériques. La BnF souhaite donc valoriser son offre, pour une part unique ou rare sur le web : une photographie d'Eugène Atget, une édition du XVIe siècle d'un ouvrage de Christine de Pisan. *Data.bnf.fr* permet aussi de reconstituer le lien entre les documents numériques et les descriptions bibliographiques et de fournir les informations sur les ressources non numérisées de la BnF. La consultation d'un document est enrichie et contextualisée. Si, par exemple, l'internaute, depuis la page sur la *Divine comédie* de Dante Alighieri[6], consulte une édition du XVe siècle, il peut trouver des informations sur l'auteur, complétées par des informations de ressources extérieures comme Wikipédia, mais aussi sur l'auteur du commentaire Marcile Ficin, puis trouver la version numérisée du manuscrit du XIVe siècle, consulter facilement différents volumes d'une édition du XIXe siècle en plusieurs volumes, ou encore rebondir vers la page sur la date « 1472[7] ». La notion de lien hypertexte vient donc servir le signalement des documents d'une façon inédite, en plaçant les informations pertinentes sur le parcours de l'utilisateur.

Data.bnf.fr « *permet la consultation à distance en utilisant les technologies les plus modernes de transmission des données*[8] », en d'autres termes, répond à la fois aux missions traditionnelles de la BnF et aux nouveaux usages du public.

2. Groupe de travail Ifla sur les Fonctionnalités requises des notices bibliographiques. En ligne : www.ifla.org/publications/functional-requirements-for-bibliographic-records

3. Le World Wide Web Consortium, ou W3C est un organisme de normalisation du web. www.w3.org

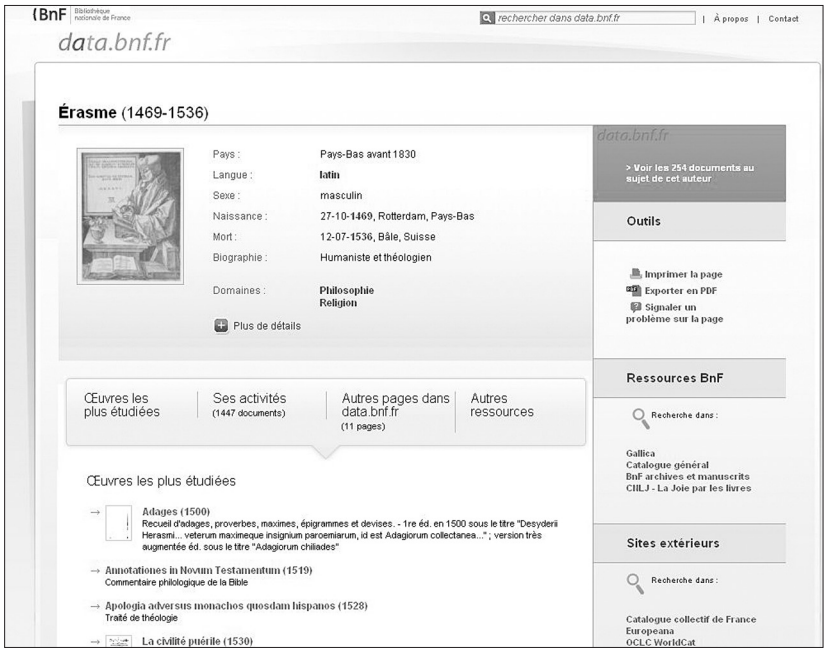
4. L'Open Data désigne le mouvement d'ouverture sur le web des données publiques ne relevant ni de la vie privée ni de la sécurité et collectées par les organismes publics.

5. http://data.bnf.fr/docs/Licence-Ouverte-Open-Licence.pdf

6. http://data.bnf.fr/ark:/12148/cb11952658b

7. http://data.bnf.fr/what-happened/date-1472

8. Décret n° 94-3 du 3 janvier 1994 portant création de la Bibliothèque nationale de France. En ligne : www.legifrance.gouv.fr



Page *data.bnf.fr* de l'auteur Érasme

« Passage en production »

Le projet est complémentaire des autres formes de médiation numérique de la BnF, comme l'éditorialisation autour des documents de Gallica ou l'ouverture sur les réseaux sociaux. Certains aspects du site sont traditionnels à dessein (interface épurée, listes simples, pas de présence autonome sur les réseaux sociaux), car l'objet principal demeure les apports d'outils automatiques de traitement des données.

« *On voit des ordinateurs partout, sauf dans les statistiques de productivité* », disait Solow[9] à propos du monde marchand. Dans les bibliothèques, l'informatique permet traditionnellement la gestion et le signalement des documents. Mais son apport est potentiellement plus important, comme le montre l'essor des « Digital Humanities[10] ».

Le logiciel de *data.bnf.fr* (CubicWeb), permet d'extraire des données des différentes bases de données (BnF catalogue général, BnF archives et manuscrits, Gallica), de les fédérer autour de concepts (les œuvres, les auteurs et les sujets), grâce à des algorithmes d'alignement et de regroupement, et d'en donner plusieurs vues : des données brutes utilisables et lisibles par des machines et des vues classiques dans des pages HTML.

Cependant, l'emploi de techniques automatiques ne doit jamais faire perdre de vue les réalités du métier et les besoins du public. La programmation des robots est donc au service de la bibliothèque, dont les utilisateurs sont représentés par le chef de projet, qui au quotidien est l'interface avec les services informatiques. Pour rester au plus proche des besoins des utilisateurs et des objectifs des bibliothèques, un projet web doit garder une certaine souplesse dans son évolution. Aussi *data.bnf.fr* est-il construit selon la méthode de gestion de projet informatique dite « méthode agile[11] », qui permet de travailler en concertation avec les développeurs, et qui structure, selon un rythme de trois semaines, les évolutions de l'application (on parle d'« itérations »). Les fonctionnalités demandées sont priorisées, puis complétées ou transformées, pour rester au plus proche de la demande du chef de projet et des besoins de la bibliothèque. Travailler sur *data.bnf.fr* implique donc un suivi « métier » de toutes les opérations de développement, d'intégration, de validation, jusqu'aux mises en production du site.

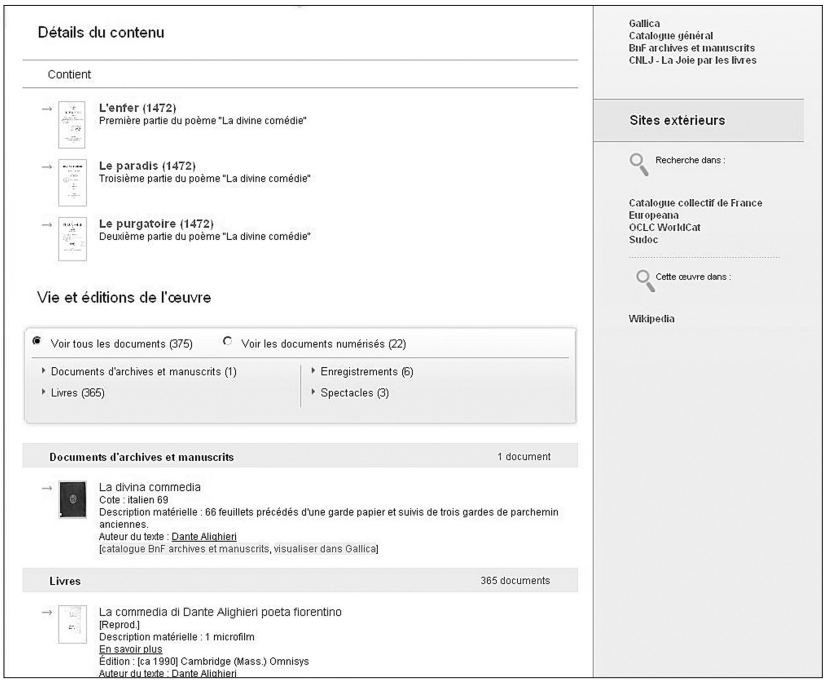
D'autre part, l'automatisation implique une politique de long terme de l'établissement sur la structuration des données et la construction de notices d'autorité. Grâce aux liens effectués par les catalogueurs entre les notices d'autorité et les notices bibliographiques, il est possible, sur la page *data.bnf.fr* d'Érasme[12], de trouver automatiquement toutes les notices descriptives liées à la notice d'autorité personne d'Érasme, avec la mention de la fonction de cet auteur sur chaque document (auteur, préfacier, illustrateur...). D'autre part, la gestion d'identifiants pérennes et fiables par la bibliothèque est capitale. En dotant les documents de Gallica et les notices du Catalogue général d'identifiants pérennes (les identifiants ARK[13]), la BnF a favorisé une réutilisation confiante de ses ressources.

Le passage des fichiers papier au catalogue sur le web, au-delà de la simple conversion des notices existantes, transforme donc en profondeur notre conception du catalogage. Les liens hypertextes permettent de pointer vers des informations sans les répliquer. L'utilisation d'identifiants pérennes dans les métadonnées permet d'organiser l'information et de l'enrichir.

Ces évolutions, que *data.bnf.fr* rend bien visibles, ont des conséquences sur le métier de catalogueur : d'abord, elles valorisent ce travail de fond et en démontrent la légitimité et la force, ce qui n'est pas sans importance dans un contexte de resserrement budgétaire ; elles suscitent aussi des questionnements sur la qualité des données, issues de l'histoire longue et complexe

9. Robert Solow, « We'd Better Watch Out », *New York Review of Books*, 12 juillet 1987.
10. Les humanités numériques sont une discipline transverse, qui étudie l'apport des outils informatiques à la recherche en sciences humaines et en littérature.
11. http://commons.wikimedia.org/wiki/File:Agile-Software-Development-Poster-En.pdf
12. http://data.bnf.fr/11886243/erasme
13. www.bnf.fr/fr/professionnels/s_informer_autres_numeros/a.ark_autres_numeros.html

du catalogue de la BnF, et sur le choix des données à exposer sur le web ; elles révèlent encore l'importance de l'intégration de liens, à l'étape de la production, au sein même des notices, en particulier dans les nouveaux formats archivistiques comme l'Encoding Archive Description (EAD) ou la Text Encoding Initiative (TEI). Enfin, le développement du « web de données » est un moyen de rationaliser le travail et les outils, en évitant la redondance de l'information. À long terme, il serait envisageable, par exemple, qu'une bibliothèque municipale constitue son catalogue en pointant vers des ressources exposées dans *data.bnf.fr*, par des liens hypertextes, complétés par des données locales.



Extrait de la page *data.bnf.fr* de l'œuvre *La divine comédie*

Perspectives : « *business as usual* »

Aujourd'hui, *data.bnf.fr* continue d'évoluer, dans le cadre d'un nouveau marché public. Il arrive à une nouvelle étape : après avoir été conçu comme une expérimentation, il a grandi et largement fait ses preuves. Il peut prétendre devenir un objet bibliothéconomique durable, à part entière, dont le développement appelle maintenant la définition d'une politique documentaire. En effet, le site comprend 200 000 pages, soit environ 20 % des ressources du catalogue général. Il doit s'agrandir pour intégrer progressivement de nouvelles ressources : nouveaux auteurs, nouvelles œuvres, mais aussi nouveaux types de documents et nouvelles bases (bibliographies, catalogues, expositions virtuelles de la BnF...). *Data.bnf.fr* pourrait être, à terme, un pivot au cœur d'un écosystème de données de la BnF.

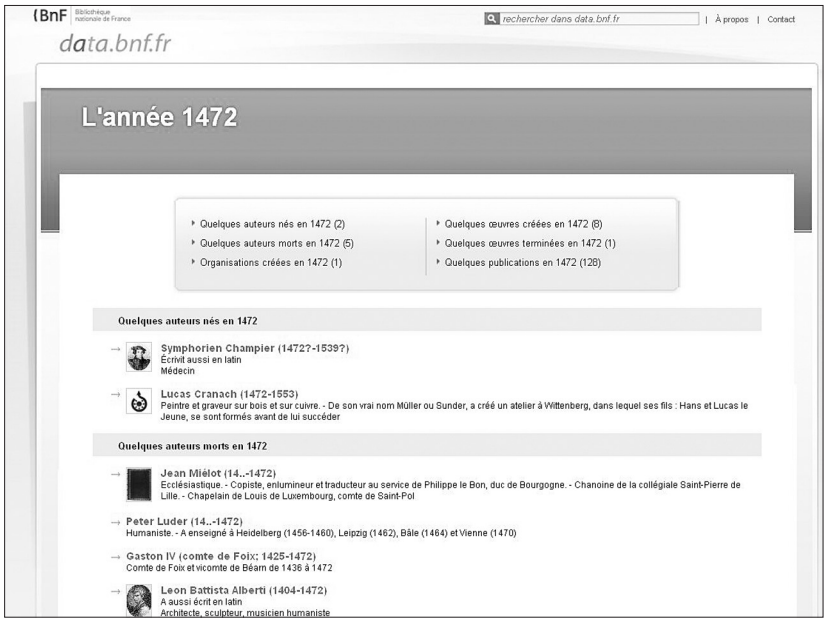
Il faut donc donner un sens à cette extension massive du site : en fonction de l'offre de la BnF, en mettant en avant les œuvres numérisées, par exemple ; mais surtout en fonction des besoins des utilisateurs et des réutilisateurs. Les statistiques du site montrent que de nombreuses pages sont consultées, mais que chacune d'elles l'est peu. Cet éparpillement des consultations sur des ressources, appartenant à des niches de savoir, est caractéristique de la « longue traîne[14] » sur le web. Cependant, l'observation de ces publics reste difficile et le lecteur virtuel, méconnu.

Au-delà de la seule consultation, nous avons connaissance de plusieurs projets réutilisant les données de *data.bnf.fr* : par exemple, If Verso[15], plateforme du livre traduit de l'Institut français, exploite des données FRBRisées pour le regroupement des traductions autour d'une même œuvre ; Isidore[16] (portail d'accès aux données numériques des sciences humaines et sociales) et le projet de logiciel pédagogique AbulEdu[17] réutilisent les données Rameau en SKOS, un des vocabulaires principaux du web sémantique ; l'application mobile CatBNF[18] pour consulter les données de *data.bnf.fr* a été développée par un particulier. Nous devons non seulement encourager ces types de réutilisation, mais aussi mieux les évaluer, ce qui est problématique quand le principe de l'Open Data est : « *Prenez nos données et faites-en ce que vous voulez.* » La question de la diffusion des références est ainsi posée d'une nouvelle manière : si on souhaite diffuser les informations auprès du plus grand nombre, comment en observer l'utilisation ?

Les évolutions fonctionnelles du site posent aussi la question de l'articulation à long terme d'un site, conçu comme un pivot entre les bases de la BnF et dont l'objectif est d'inciter à consulter les applications existantes de la BnF. En effet, *data.bnf.fr* n'a vocation à remplacer ni les catalogues ni Gallica, mais bien de permettre une première approche des ressources de la BnF.

En revanche, au-delà de l'application et des services qu'elle rend, les technologies de *data.bnf.fr* peuvent être réemployées dans d'autres applications de la BnF. Les études menées dans le cadre du projet donnent des outils pour faciliter la création de notices d'autorité œuvre et leur lien aux notices bibliographiques. À l'heure où l'adoption du code Resource Description and Access (RDA) est discutée et suscite des réserves, *data.bnf.fr* est un

14. Chris Anderson, *La longue traîne*, Pearson, 2009.
15. http://ifverso.com
16. www.rechercheisidore.fr
17. www.abuledu.org
18. http://itunes.apple.com/fr/app/catbnf/id501048946?mt=8



Page de date 1472 dans *data.bnf.fr*

moyen d'avancer vers la FRBRisation des catalogues.

D'autre part, le logiciel CubicWeb, utilisé pour *data.bnf.fr*, est un outil intéressant pour publier d'autres bases. Son infrastructure permet en effet de relier des données de nature différente, de publier les pages sous différents formats, conformes aux standards du web et, surtout, de les ouvrir sur le web sémantique. À titre d'exemple, ce logiciel vient d'être choisi dans l'objectif de publier une base sur les anciennes reliures de la Réserve des livres rares, en liant leurs descriptions aux documents numérisés, et en respectant les standards du web.

À plus long terme enfin, *data.bnf.fr* pose la question du devenir d'un projet de « recherche et développement ». Doit-il conserver un aspect expérimental, par définition incertain ? Le web étant par nature versatile, le site est-il amené à évoluer continuellement ? Entre maintenance et évolutions, un juste équilibre sera à trouver.

L'informatique est-elle responsable du progressif recul de la lecture et du livre, dans les pratiques culturelles des Français ? « *À un moment où plus de la moitié des Français disposent chez eux d'une connexion à haut débit et où plus d'un tiers d'entre eux utilisent l'Internet tous les jours à des fins personnelles*[19] », *data.bnf.fr*, en associant les données bibliographiques et numériques et en les exposant sur le web, fait le pari inverse : inciter le public, grâce au web, à consulter les documents numériques de la bibliothèque, voire à venir sur le site physique de la BnF. L'utilisation d'outils automatiques de signalement et de médiation peut ainsi transformer profondément l'offre des bibliothèques, son rapport avec le public, et le métier de bibliothécaire. ●

Août 2012

19. Olivier Donnat, *Les pratiques culturelles des Français à l'ère numérique. Éléments de synthèse 1997-2008*. En ligne : www.pratiquesculturelles.culture.gouv.fr/doc/08synthese.pdf