Présentation générale du projet data.bnf.fr

La Bibliothèque nationale a mis en œuvre un nouveau projet, qui a pour but de rendre ses données plus utiles sur le web. Ceci nécessite de transformer données existantes, d'enrichir et de lier son jeu de données avec des ressources externes et internes, ainsi que de publier des pages HTML sur lesquelles les utilisateurs et les moteurs de recherche puissent naviguer. Les données brutes en RDF sont aussi disponibles suivant les principes de l'architecture du linked open data.

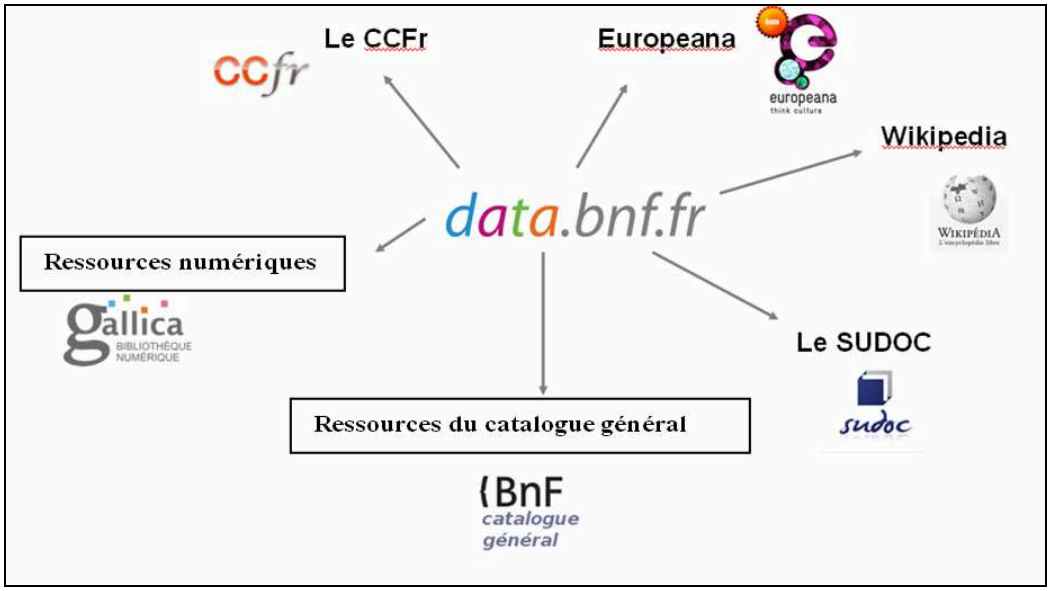
Mots clés : Linked Data ; Web sémantique ; métadonnées ; interopérabilité ; RDF ; URI

1. Mettre les données bibliographiques sur le Web

→ Une mise en valeur des données de la BnF

Les données des bibliothèques peuvent être difficiles à trouver sur le web. Il est naturellement possible, à la BnF, d'accéder à toutes les ressources et à tous les services sur le site de la bibliothèque (www.bnf.fr). Mais pour le moment, peu sont indexés par les moteurs de recherche. Quand c'est le cas, il est difficile de trier les résultats pertinents.
Certains livres numériques, même lorsqu'ils sont gratuits et intégralement disponibles, sont parfois introuvables sur le web, pour quelqu'un qui n'en connaîtrait pas l'existence a priori. Le projet data.bnf.fr est donc un moyen d'ouvrir la bibliothèque numérique de la BnF, Gallica, à un public encore plus large.
De plus, les catalogues de bibliothèques sont, en général, entreposés dans des bases de données relationnelles et, par conséquent, tout simplement inexploitables par des moteurs de recherche. Les utilisateurs ne peuvent donc accéder à ces catalogues (en particulier le Catalogue général et Catalogue BnF Archives et Manuscrits) que par le biais des portails de bibliothèques, dont ils ne connaissent pas forcément l'existence.

Dans la pratique, les internautes ont donc peu de chance de trouver nos ressources à partir d'une interface de moteur de recherche, et doivent connaître déjà notre institution et ses outils.



Quelques liens depuis data.bnf.fr

Data.bnf.fr est un pivot documentaire qui rassemble des données numériques et des données descriptives de différents catalogues de la bibliothèque, et permet à l'utilisateur de retrouver facilement des informations pertinentes.
Nous voulons donc que les ressources de la BnF soient aussi visibles sur le web que la bibliothèque dans la ville.

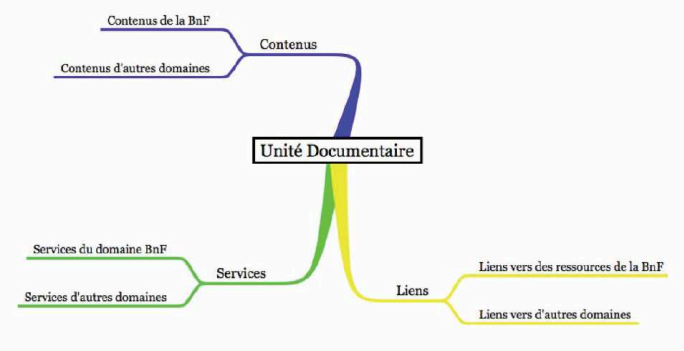
→ Les données structurées ont une valeur

Les données typées, normalisées et référencées sont le fondement de la recherche sur Internet. Depuis longtemps, les bibliothécaires ont identifié leurs ressources de façon uniforme. Ils font déjà, en quelque sorte, du « linked data », en créant des liens entre les œuvres, les auteurs et les sujets. Notre catalogue de bibliothèque comprend ainsi **plus de 12 millions de notices, toutes structurées et reliées ensemble**. Il repose sur deux millions de notices d'autorités fiables et exactes sur les auteurs, les organisations, les œuvres et les sujets (noms communs, géographiques et subdivisions chronologiques RAMEAU[1]), qui sont maintenus via des URIs permanents, qui sont des **identifiants ARK** à la BnF[2].
Les données mises à disposition par une institution de service public telle que la BnF ont une valeur particulière, n'ayant pas d'autre but que d'offrir des informations utiles, des sources fiables et des liens déréférençables . Ils nous permettent non seulement de citer et identifier les ressources, mais aussi de les rassembler. Nous pouvons ainsi travailler sur l'alignement des ressources, en se basant notamment sur les modèles FRBR, dans l'objectif constant d'améliorer les services au public.

Nous voulons donner aux machines les moyens d'indexer l'accès au **contenu, aux liens et aux services**, rassemblés dans chaque page (ou Unité documentaire) autour d'un concept au sens large du terme. Par « contenu », nous entendons des données descriptives, valides et exactes, élaborées par un service non lucratif. Par « liens », une façon de naviguer et d'aller vers des ressources plus pertinentes si nécessaires, notamment vers des versions en ligne d'une œuvre, ainsi que l'intégration dans un graphe de données. Enfin, « services » peut signifier d'autres services de la bibliothèque, tels qu'un service de question-réponse, le téléchargement ou l'impression.

---

[1] http://rameau.bnf.fr
[2] Bermes, Emmanuelle. (2006). *Les identifiants pérennes à la BnF*. URL : http://bibnum.bnf.fr/identifiants/identifiants-200605.pdf



Articulation de data.bnf.fr autour des Unités Documentaires

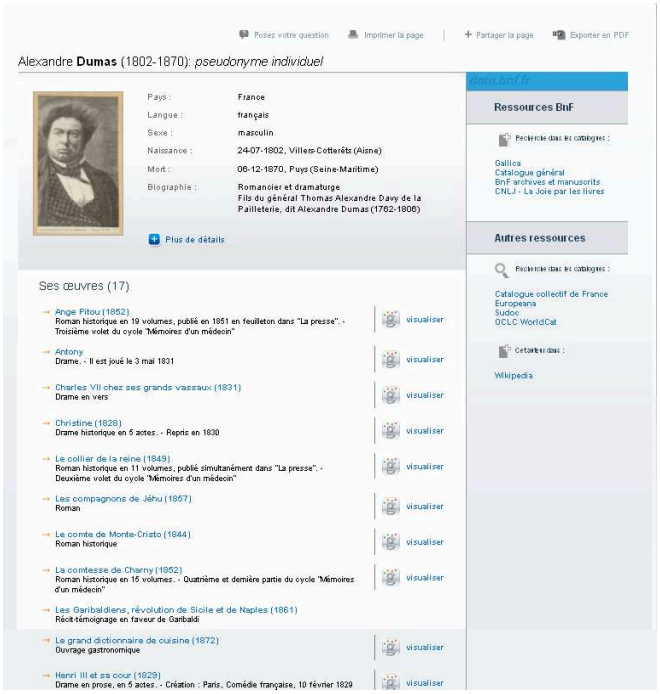
Chaque page rassemble des liens, des services et du contenu autour d'un concept informatif. Le site est basé sur de nouvelles de techniques de modélisation et sur le langage RDF (Resource description framework).

2. Des pages web sur les auteurs, les œuvres et les sujets

Nous avons construit une interface web, dont les pages HTML regroupent des informations autour des concepts d' « œuvres », d' « auteurs » et de « sujets », destinées au grand public. Chaque concept correspond à une unité documentaire. En parallèle, nous publions des données brutes avec un modèle construit autour de « concepts » et des données interopérables, qui sont exposées sur le web de données. L'enjeu majeur est :

- d'un côté, de s'assurer que nous pouvons répondre aux besoins à court terme, à des requêtes spécifiques ou à des questions stratégiques, relatives à des ressources populaires, au graphisme ou à des outils en vogue, par exemple.

- de l'autre, de prendre en compte les missions traditionnelles et à long terme de la bibliothèque, comme de proposer des modèles de données, des solutions techniquement avancées ou encore des informations valides et correctes.



Alexandre Dumas dans data.bnf.fr

→ Le lien aux modèle FRBR (Functional requirements for bibliographic records)[3]

Cette façon d'articuler les données bibliographiques sur le web nécessite plusieurs choix. D'un point de vue pratique, pour publier des pages HTML, notre modèle de données devait se fonder sur des concepts pertinents. Nous avons choisi les concepts d'œuvres, d'auteurs et de sujets, qui sont des entités du **modèle FRBR**, avec lequel notre modèle veut être compatible.

Cette interface web est à la croisée des chemins entre les différentes ressources que nous mettons sur internet. Elle rassemble différents types de données au bon niveau : œuvre, expression et manifestation. Pour un auteur, les internautes retrouvent tous les liens vers les pages web des œuvres de et au sujet de l'auteur dans deux sections distinctes de la page. Pour une œuvre, un lien est effectué vers la page de l'auteur, mais aussi vers les différentes manifestations de cette œuvre (ressources bibliographiques, ressources en ligne, fonds d'archive).

Pour créer ces pages, nous devons rassembler des données issues de différentes bases de données, dans différents formats : EAD[4] (Encoded Archival Description) pour les manuscrits et les fonds d'archives, MARC (Intermarc) pour le catalogue général, et Dublin Core[5] pour les livres numérisés de Gallica[6] et pour les expositions virtuelles[7]. C'est pourquoi le travail de modélisation est en lien direct avec celui d'alignement et d'enrichissement des données qui ont été extraites et traitées.

Enfin, ces pages proposent plusieurs fonctionnalités : exporter en PDF, exporter et envoyer, citer sur les réseaux sociaux…. En outre, il existe des liens vers d'autres services en ligne où l'utilisateur peut trouver l'information pertinente, si la page sur laquelle il était ne suffisait pas. Nous récupérons des données de différentes bases de données ouvertes présentes dans telles que Wikipedia, pour lier leurs données aux nôtres et fournir d'autres types d'informations.

Les pages de data.bnf.fr doivent :
- être facile à comprendre et à parcourir pour l'utilisateur ;
- développer de nouveaux modèles, tels que le modèle FRBR, et proposer des alignements vers des jeux de données extérieurs.

3. Le web de données

Nous avons construit une architecture de publication qui permet à la fois, d'avoir des pages HTML et d'exposer les données brutes sur le web de données.

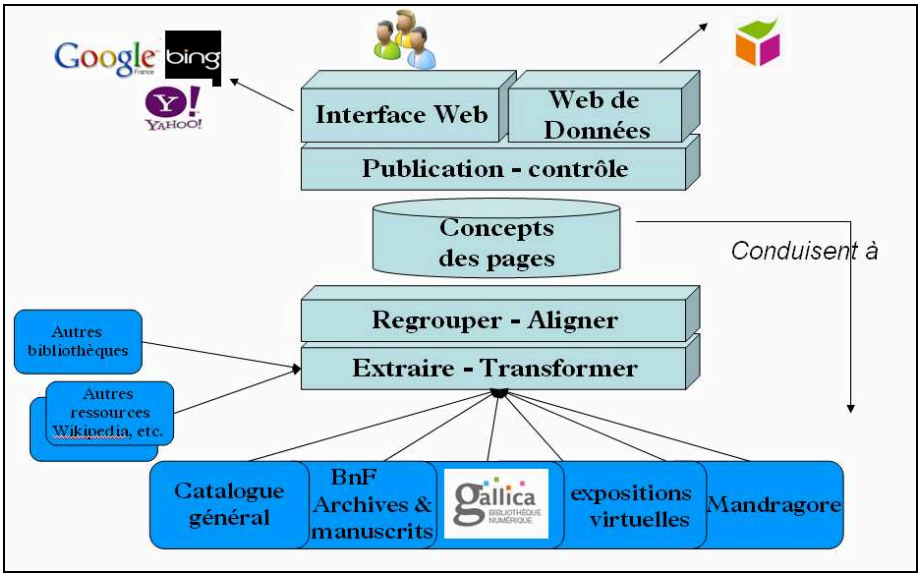
Notre objectif est d'utiliser des standards communs et de créer ce service au travers d'un modèle répondant aux bonnes pratiques du web sémantique. Ceci nous permet de mettre nos ressources et nos données dans le « linked data », afin de les rendre aussi utiles que possible, tant pour le grand public que pour les professionnels.
L'utilisation des standards du web sémantique permet d'exposer des **données structurées exploitables par des machines,** en se fondant sur **l'interopérabilité non seulement avec les ressources extérieures, mais aussi au sein même de nos bases de données internes**, puisque nos ressources viennent de différents catalogues dans différents formats.
Nous exposons notamment les notices de sujets de la BnF RAMEAU (Répertoire d'autorité-matière encyclopédique et alphabétique unifié). Elles ont été converties dans le langage RDF SKOS (Simple Kowledge Organisation), dans le cadre du projet européen TELplus. **Ce référentiel a été mis à jour avec la base de données totale et actualisée de la BnF**.
Nous utilisons un logiciel nommé Cubicweb, un schéma d'application du web sémantique, sous licence LGPL.

→ Les principes du Web sémantique[8]

[3] http://www.ifla.org/files/cataloguing/frbr/frbr-fr.pdf
[4] http://www.lcweb.loc.gov/ead/
[5] http://dublincore.org/
[6] http://gallica.bnf.fr/
[7] http://expositions.bnf.fr/
[8] W3C Incubator Group Report. Library Linked Data Incubator Group Final Report. 25 October 2011. Disponible en anglais : http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/

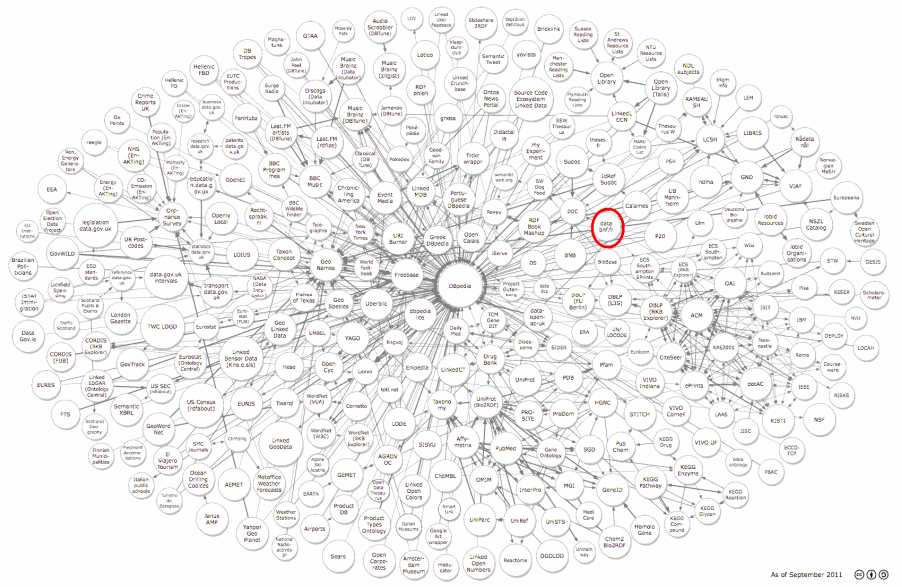
Pour les pages décrivant les ressources, nous souhaitons :

- maintenir des **URIs permanents**, qui doivent rester compréhensibles pour l'utilisateur, faire référence à des ressources pertinentes et être intégrées ;
- créer un système de **négociation de contenu** ;
- **utiliser un modèle de données compatible avec le RDF**, avec des vocabulaires standards (principalement SKOS, RDA et FOAF) ;
- utiliser, autant que possible, des **vocabulaires existants** ;
- utiliser un **vocabulaire spécifique** uniquement pour les classes et les objets spécifiques à la bibliothèque ;
- aligner nos données avec des données extérieures de la Bibliothèque du Congrès, la Deutsche Nationalbibliothek (Bibliothèque nationale allemande), Geonames et le Thésaurus W.



Principes fonctionnels de data.bnf.fr

Créer un hub de données significatif dans le nuage du « linked data » soulève questions très stratégiques. Premièrement en termes de **descriptions bibliographiques**, en particulier sur le modèle FRBR, sur la norme Resource Description and Access (RDA) et sur l'intégration de l'EAD. Cette approche est particulièrement innovante dans les bibliothèques. Si, celles-ci ont indubitablement une forte tradition de standardisation autour de l'interopérabilité et de l'efficacité du partage des données, elles doivent néanmoins faire face aux nouveaux enjeux du « linked data ». Deuxièmement, les **enjeux juridiques** autour de la dissémination des données sont importants. Data.bnf.fr est un projet ouvert **:** les données peuvent être récupérées et réutilisées, selon les besoins, intégralement ou partiellement, notamment par des développeurs et des professionnels des bibliothèques La licence d'utilisation permet ainsi la récupération et la réutilisation gratuites et libres des données en RDF, sous réserve du maintien de la mention de la source BnF.



Le nuage du linked open data par Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

Nous tentons de construire un site fondé sur un modèle compatible avec les principes du web sémantique.
La BnF a rendu ses données en RDF librement et gratuitement disponibles et réutilisables.

→ Les données embarquées dans les pages html

Dans l'objectif de rendre nos pages aussi utiles que possible, nous prenons les métadonnées les plus pertinentes pour les intégrer dans le code source du HTML sous forme de balises RDFa. Par ailleurs, nous utilisons le vocabulaire partagé Schema.org pour structurer nos pages avec des microdonnées et favoriser le référencement par les moteurs de recherche. Nous insérons aussi des balises de l'Opengraph Protocol (OG).

Nous pensons que cette approche « open data » doit rester cohérente avec le travail de construction des pages web, non seulement parce que la modélisation des données autour des concepts entraîne le regroupement d'informations pertinentes autour d'un URI unique, mais aussi parce qu'une des bases du web sémantique est de fournir, à partir de ces URIs, une information utile et compréhensible par un être humain.

Les pages web sont structurées avec des données embarquées.
Les types et formats des données en RDF sont identiques à ceux que l'on trouve dans le nuage du « linked data ».
Ces deux façons (RDF et HTML) d'accéder aux données sont tout à fait complémentaires.

4. Conclusion

data.bnf.fr ne remplace pas les catalogues existants : il s'agit simplement d'un outil destiné à aider les internautes à trouver dans nos ressources ce dont ils ont besoin, de diffuser des contenus, des liens et des services, et de les partager plus facilement. Cette approche de l'interface web, centrée sur les données, nous permet de combiner nos missions traditionnelles avec l'état de l'art de recherche et ses évolutions.

data.bnf.fr est toujours en développement. Cette première version du site comprend les principaux auteurs français de la littérature, ainsi que des auteurs de l'antiquité, et des juristes. Le nombre de pages est en passe d'être augmenté avec notamment les pages d'œuvres musicales et la création de pages sujets.

A long terme nous souhaiterions lier nos données avec celles d'autres institutions, telles que les universités, les archives et les musées. En France plusieurs projets de web sémantique sous en cours de développement. Néanmoins, il reste de nombreuses difficultés liées aux différences entre les types de ressources et de données.
Dans la mesure où de nombreuses institutions culturelles ont les mêmes enjeux concernant leur présence sur le web et le travail sur les données descriptives et les métadonnées, nous espérons que ce sera l'occasion de partager nos expériences de manière plus collaborative.

Contact : data@bnf.fr
15/11/2011