GFII
GROUPEMENT FRANÇAIS DE L'INDUSTRIE DE L'INFORMATION

Données culturelles et Linked Open Data

Valoriser le patrimoine public dans le web de données

Synthèse de la journée d'étude du GFII organisée le 26 mars 2013 à la Maison de l'Europe

Journée animée par Jean Delahousse, Consultant Senior, JDC Consult et animateur du groupe de travail « Web Sémantique » du GFII

Sommaire

1. Etat des lieux de l'Open Data culturel, en France et en Europe

Lionel Maurel est co-auteur du rapport Open Glam sur l'ouverture des données et des contenus culturels, publié par Wikimedia France, Creative Commons France, Open Knowledge Foundation, et Veni Vidi Libri. Lionel Maurel est également le rédacteur du blog SI-Lex, à la croisée du droit et des Sciences de l'Information et de la Communication.

Présentation disponible en ligne : http://www.gfii.fr/uploads/docs/LionelMaurel_LOD2013.pdf

L'exception culturelle est inscrite dans le cadre juridique de l'Open Data

La Loi CADA (11 juillet 1978) fixe le cadre juridique de la réutilisation des informations publiques en France. L'article 10 énonce le principe du droit de réutilisation des informations produites ou détenues par les administrations publiques par toute personne qui le souhaite. Mais les *« données culturelles »*, au même titre que les données de recherche, bénéficient d'un statut dérogatoire. L'article 11 énonce que les établissements d'enseignement supérieur et de recherche et les établissements culturels sont libres de fixer les conditions dans lesquelles les informations qu'ils produisent ou détiennent peuvent être réutilisées. La formulation est équivoque et entretient l'incertitude. On peut s'interroger sur la finalité de l'exception culturelle pour les établissements concernés. Fixer leurs propres conditions tarifaires ? Poser des conditions supplémentaires à la réutilisation ? Protéger des données personnelles ? Protéger les droits de propriété intellectuelle de tiers ? Faire obstacle au principe même de la réutilisation ?

L'Etat fixe la gratuité comme principe et la tarification comme exception

En France, EtaLab impulse et cordonne la politique Open Data de l'Etat[1]. Ce moteur politique s'appuie sur plusieurs dispositifs :

- La création d'un portail « data.gouv.fr » dans lequel les administrations d'Etat doivent référencer leurs jeux de données (les administrations territoriales sont invitées à le faire, mais sans obligation).
- La création de la licence EtaLab (Licence ouverte de l'Etat), assise sur le droit des données publiques, mais fonctionnant comme une licence libre, autorisant notamment la réutilisation commerciale à condition de mentionner la source.
- La circulaire du 26 mai 2011 qui fait de la gratuité des données le principe, et de la tarification l'exception, soumise à justification auprès du COEPIA.

A noter : le Ministère de la Culture n'est pas à proprement parler un « établissement culturel ». Il est donc soumis à l'obligation d'ouvrir ses données gratuitement ou d'en justifier la tarification.

[1] La politique d'ouverture en ligne des données publiques (« Open Data »), pilotée par la mission Etalab sous l'autorité du Premier ministre depuis février 2011, a été rattachée directement au Secrétaire général pour la modernisation de l'action publique. Le Premier ministre Jean-Marc Ayrault a créé par décret du 31 octobre 2012 le Secrétariat général pour la modernisation de l'action publique (SGMAP), placé sous son autorité, rattaché au Secrétaire général du gouvernement, et dont la direction a été confiée à M. Jérôme Filippini, conseiller maître à la Cour des comptes. Plus d'informations à l'adresse suivante : http://www.etalab.gouv.fr/

Le Linked Data n'est pas l'Open Data

La circulaire du 26 mai 2011 prolonge l'exception dont bénéficient les « données culturelles » au sein de l'article 11 de la loi CADA : les établissements d'enseignement supérieur et de recherche et les établissements culturels *« peuvent s'ils le souhaitent »* verser leurs données dans le portail pour en autoriser la réutilisation. L'ensemble des établissements culturels (archives, musées, bibliothèques) sont toujours libres d'ouvrir leur données ou non, et d'en autoriser la réutilisation ou non. L'ambiguïté du texte est maintenue. A l'arrivée, les établissements concernés peuvent se lancer dans des démarches « Linked Data » (lier les données) sans nécessairement faire de « l'Open Data » (ouvrir les données).[2]

Au niveau territorial : les données d'archives sont le point sensible

Au niveau des administrations territoriales, le cadre est différent. Le principe de *« libre administration des collectivités locales »* leur accorde une grande autonomie dans les politiques d'ouverture. De nombreuses collectivités ont joué un rôle moteur dans le développement de l'Open Data (Rennes, Nantes …) mais la culture est généralement le parent pauvre de ces initiatives. La plupart du temps, il s'agit de statistiques sur le fonctionnement d'établissements culturels (Région PACA, Bordeaux …). Le point le plus sensible concerne la réutilisation des données des archives départementales. Un grand réutilisateur de données d'archives comme NotreFamille.com s'est plusieurs fois vu refuser le droit de réutiliser les données d'Etat Civil au motif du droit à la protection de la vie privée (données personnelles) ou du droit des bases de données.

Vers une balkanisation des données d'archives ?

La décision du Tribunal de Poitiers de rejeter le recours de NotreFamille.com dans le contentieux qui l'oppose aux Archives Départementales de la Vienne témoigne d'un inquiétant retour du droit des bases de données, et de la possibilité qu'il offre de neutraliser la loi de 1978. Le tribunal s'est appuyé sur le droit des bases de données pour motiver son refus d'accéder à la demande de NotreFamille.com, en considérant le Département de la Vienne comme un producteur de base de données bénéficiant à ce titre de la protection de son contenu, celui-ci attestant d'un *« investissement financier, matériel ou humain substantiel »* [3]. On peut aujourd'hui se demander si on ne se dirige pas vers une balkanisation des données d'archives. Le cas des données d'archives montre que l'exception réservée aux données culturelles encourage les inégalités d'accès et de réutilisation selon les territoires.

L'Open Data culturel : pas une priorité pour le gouvernement

Le gouvernement a publié sa feuille de route concernant l'ouverture des données publiques en février dernier[4]. Les priorités ont été arrêtées pour 6 secteurs stratégiques : *données de santé, données de logement, données de transport, données d'éducation, dépenses publiques,*

[2] http://scinfolex.wordpress.com/2012/10/11/le-centre-pompidou-virtuel-ouvert-ou-sous-verre/

[3] http://scinfolex.wordpress.com/2013/02/01/open-data-rip-la-reutilisation-des-informations-publiques-bientot-dissoute-dans-le-droit-des-bases-de-donnees/

[4] http://www.etalab.gouv.fr/article-la-feuille-de-route-du-gouvernement-en-matiere-d-ouverture-et-de-partage-des-donnees-publiques-115767801.html

environnement. La culture n'est pas identifiée comme une priorité pour le gouvernement alors même que le potentiel de pour la réutilisation est énorme[5].

En France : quelques initiatives institutionnelles fortes

On constate malgré tout plusieurs initiatives intéressantes au niveau national :

- Le Conseil National du Numérique a publié un avis en juin 2012, recommandant la suppression de l'exception culturelle, considérée comme un frein au développement de l'Open Data.[6]
- Deux signaux forts ont été émis cette année par le Ministère de la Culture :
    - La publication du Guide Data Culture dans lequel le Ministère exprime une orientation favorable à l'ouverture et la réutilisation des données culturelles[7].
    - La signature de la Convention SemanticPedia
- Une initiative phare : data.BnF.fr

En Europe : nombreuses initiatives « bottom-up »

En Europe, on trouve de nombreuses initiatives associatives et militantes. En 2012, Wikimedia France, Creative Commons France, Open Knowledge Foundation et Veni Vidi Libri ont publié le rapport Open Glam dans lequel le collectif soulignait la nécessité de réintégrer les données culturelles dans le régime de droit commun de la réutilisation des informations publiques.[8] Cette dynamique « bottom-up » cohabite avec les initiatives de grandes institutions culturelles qui jouent un rôle moteur dans l'Open Data européen (BNF, British Library, Rijksmuseum Amsterdam, Deutsche National Bibliothek …)

Le rôle moteur d'Europeana

Il faut souligner le rôle moteur d'Europeana dans le développement de « l'Open Linked Data ». Europeana a mis en place un nouveau ***Data Exchange Agreement*** en 2011, protocole d'échange obligeant les bibliothèques, archives et musées partenaires à fournir les métadonnées de leurs catalogues sous licence C0.0 et dans des formats compatibles RDF[9].

Questions / Echanges avec la salle

Quelles contradictions dans les politiques open data culturelles en France ?

D'un côté, le maintien de l'exception culturelle dans la circulaire EtaLab permet aux établissements qui le souhaitent d'interdire la réutilisation des données. De l'autre, l'Etat signe un acte fort comme la convention SemanticPédia qui va dans le sens d'une ouverture maximale des données et autorise leur réutilisation commerciale. On peut interroger ces contradictions à l'aune de la problématique des politiques nationales de numérisation. L'Etat accorde, à travers les investissements d'avenir, des

[5] http://scinfolex.wordpress.com/2013/03/03/les-donnees-culturelles-absentes-de-la-feuille-de-route-du-gouvernement-sur-lopen-data/

[6] http://www.cnnumerique.fr/wp-content/uploads/2012/06/2012-06-05_AvisCNNum_12_OpenData.pdf

[7] http://www.etalab.gouv.fr/article-de-nouveaux-jeux-de-donnees-du-ministere-de-la-culture-et-de-la-communication-116421284.html

[8] http://www.donneeslibres.info/

[9] http://scinfolex.wordpress.com/2011/10/10/larchitecture-juridique-ouverte-deuropeana/

budgets importants aux bibliothèques, aux musées et aux archives pour numériser leurs contenus. C'est un travail coûteux et complexe. Une fois numérisé et mis en ligne, ce patrimoine est une vitrine pour les institutions et les élus qui peuvent s'opposer à ce qu'il soit libéré et réutilisable par tous dans un cadre Open Data.

Quelle est la situation à l'étranger ?

Aucun pays européen ne bénéficie d'une politique complète d'open data culturel. La révision de la Directive PSI devrait harmoniser les choses, mais il n'est pas garanti que la Commission fasse de l'Open Data le principe de la diffusion et de la réutilisation des données publiques. L'Europe souffre d'un important décalage avec les Etats-Unis sur ce sujet.

N'y-a-il pas contradiction dans la politique de l'Etat qui demande aux administrations d'un côté d'ouvrir leurs données gratuitement, et de l'autre, de dégager des ressources propres ?

Dans un contexte de réduction générale des dépenses de fonctionnement de la puissance publique, l'Etat demande aux administrations de rentrer dans des logiques d'autofinancement. La gratuité de la mise à disposition a un coût que les administrations doivent absorber pour pouvoir continuer à faire de l'Open Data. Des modèles de partage de la valeur doivent être explorés. Les réutilisateurs commerciaux pourraient par exemple reverser une partie des bénéfices générés à partir des informations publiques aux administrations, à proportion de l'effort de collecte et de diffusion que cela implique.

Mise à jour (août 2013)

La situation a évolué depuis Mars 2013. On constate, au niveau du Ministère de la Culture, une évolution favorable à l'ouverture des données culturelles. La publication du guide Dataculture et la signature de la Convention Semanticpédia constituaient des signaux positifs d'inflexion de sa politique. Ceux-ci se sont vus confirmés par la publication en juillet 2013 d'une feuille de route du MCC pour l'ouverture et le partage des données culturelles. Le document fixe les objectifs à atteindre et dresse la liste des 10 actions à réaliser pour y parvenir[10]. Les données culturelles sont considérées comme un « jeu stratégique » pour le gouvernement. Le manque de la circulaire de février 2013 est en partie comblé. De surcroit, la voie préconisée est clairement celle de l'*Open Linked Data,* comme en témoignent les points 1-2 et 5 :

«1. Inscrire l'action du ministère de la culture et de la communication dans une politique numérique de mise à disposition de ses données publiques sur data.gouv.fr.

2. Ouvrir des jeux de données publiques stratégiques issues du secteur culturel en prenant appui sur les prescriptions du rapport Data Culture.

5. Investir les technologies du web sémantique et amorcer une dynamique de linked open data dans le secteur culturel en contribuant au rayonnement de la culture française et de la francophonie sur Internet »

[10] http://cblog.culture.fr/wp-content/uploads/2013/07/Feuille-de-route-open-data-MCC.pdf

Toutefois, il reste à savoir si cette inflexion sera suivie au niveau des établissements culturels et des collectivités territoriales, dans la mesure où ceux-ci conservent l'autonomie que leur confère la loi dans la définition de leur politique Open Data. L'« exception culturelle » demeure toujours.

Retrouvez le commentaire complet de Lionel Maurel sur l'évolution de la politique Open Data du MCC sur son blog : http://scinfolex.wordpress.com/2013/07/31/ouverture-des-donnees-culturelles-the-times-they-are-changing/ 31/07/2013

2. La stratégie d'exposition des catalogues de la BnF sur le web de données : data.bnf.fr (1)

Gildas Illien est Directeur du Département de l'Information Bibliographique et Numérique à la Bibliothèque Nationale de France.
Sa présentation est disponible en ligne :
http://www.gfii.fr/uploads/docs/GildasIllien_RomainWenz_LOD2013.pdf

Le Linked Data : un changement de métier pour la BnF

L'initiative data.bnf.fr est une petite révolution pour la BnF. Son cœur de métier est le catalogage, qui s'inscrit dans sa mission de service public : constituer la bibliographie nationale. Or, cataloguer revient à modéliser le monde, à le décrire selon des représentations et classifications normalisées propres aux bibliothécaires. A l'inverse, faire du Linked Open Data revient à reverser les données des catalogues dans le monde en les exposant dans le web de données. La démarche Linked Open Data de la BnF vise à révéler les trésors enfouis dans ses catalogues. Pour les bibliothécaires, c'est un changement de paradigme et de métier : on passe d'une logique de catalogue à une logique de métadonnées, d'une logique d'applicatifs, propre au web documentaire, à une logique d'usages, d'une logique de documents à une logique de données. En déconstruisant les catalogues, les notices et les applicatifs qui les supportent, la valeur se déporte du document aux usages.

De l'importance des métadonnées et leur normalisation

Dans ce mouvement de déconstruction, les métadonnées (descriptives, techniques, administratives, sociales, …) deviennent centrales pour assurer tous les processus et maintenir les liens entre ressources. Elles seules permettent de lier l'objet (la ressource), le concept et l'identifiant de l'objet. Leur normalisation est essentielle pour garantir la qualité, la confiance, et permettre l'interopérabilité et les enrichissements mutuels entre catalogues. Le RDF est aujourd'hui le pivot principal du Linked Open Data. Mais il faut garder à l'esprit que l'approche Linked Open Data, aussi innovante soit-elle, s'inscrit dans la continuité de l'effort bibliographique entrepris au niveau international par la communauté des bibliothèques depuis 40 ans (le "contrôle bibliographique universel"). Le RDF est une nouvelle manière d'envisager la description bibliographique, plaçant la notion de triplets au centre, mais il y a derrière un savoir-faire et des acquis de longue date de la profession, qui s'exprime également dans son appropriation du modèle FRBR.

S'adapter aux évolutions des usages sur le web

Auparavant, la stratégie qui dominait consistait à cantonner les catalogues dans les applications internes des bibliothèques, accessibles en ligne mais seulement via leurs interfaces propres, les données elles-mêmes restant enfouies dans les profondeurs du réseau, faute de structuration selon les standards du web. Mais les usages ont changé : les moteurs de recherche sur le web sont devenus le premier guichet pour accéder aux ressources. Pour que les catalogues vivent, leur publication sur le web est aujourd'hui nécessaire. Pour assurer la survie des richesses des catalogues, ce sont leurs notices, décomposées en éléments de données plus fins, aussi riches par leur contenu que par leur

structuration interne, qui deviennent l'enjeu du référencement. La sémantisation des notices dans les standards du Linked Open Data devient le principal levier de leur valorisation et de leur visibilité par tous les internautes, et pas seulement les initiés.

Innover en réalisant des économies : une nécessité

Le passage au Linked Open Data répond aussi un impératif économique. Les budgets des institutions culturelles sont en baisse, alors que le volume de données à traiter ne cesse d'augmenter. Il devient indispensable de réutiliser ce qui a été fait ailleurs, d'innover et de mutualiser les initiatives « sémantiques ». L'ingénierie sociale, qui est dans l'ADN du Linked Open Data, facilite ce travail. Les référentiels sont développés et enrichis par des communautés, des institutions, des entreprises et placés dans des standards ouverts, à la fois techniquement et juridiquement. Cela permet de réduire les coûts de maintenance et de développement des projets « sémantiques ».

Lier les données et les ouvrir : deux principes très différents

Le projet data.bnf.fr s'inscrit dans le mouvement d'ouverture des données publiques et répond à une incitation gouvernementale forte. EtaLab a mis à disposition des outils pour faciliter les initiatives Open Data, notamment la Licence Ouverte de l'Etat et le portail data.gouv.fr. Faire du « Linked Data » (lier les données), et faire de « l'Open Data » (ouvrir les données) sont deux choses très différentes. On peut faire l'un sans l'autre. La BnF a souhaité faire les deux et figure de fait aujourd'hui dans le peloton de tête des institutions culturelles engagées dans le linked open data. Lier les données implique de convertir des notices structurées dans un format pensé par et pour les bibliothèques (MARC) dans les standards du web sémantique (RDF et ses différents vocabulaires). Ouvrir les données implique d'assurer leur interopérabilité technique (compatibilité aux normes et standards existant) et juridique. La Licence Ouverte de l'Etat recommandée par Étalab, qui autorise la réutilisation, y compris commerciale, à condition de mentionner la source, constituait un cadre juridique pertinent pour la BnF. Les données en RDF moissonnables depuis data.bnf.fr (40% du catalogue de la BnF à l'été 2013) sont placées sous licence ouverte. Les données au format MARC restent soumises à redevances en cas de réutilisation commerciale. Les données au format RDF sont sous licence EtaLab.

Data.bnf.fr : les grands objectifs

« Etre plus visible » - L'enjeu de référencement était un objectif prioritaire (cf. concurrence des moteurs de recherche web).

« Etre cohérent et uni » - L'unification et la rationalisation du système d'information interne constituaient le second objectif. La BnF maintient des bases très hétérogènes, avec des catalogues utilisant historiquement des référentiels et des applicatifs différents. Lier les données permet de créer des passerelles d'un silo à l'autre et de gagner en cohérence sans passer par un « Big Bang informatique » : les applications de production bibliographiques sont pour le moment maintenues peu ou prou en l'état, c'est la stratégie de diffusion et d'exposition sur le web qui change fondamentalement. Par ailleurs, il y a une articulation étroite entre le projet data.bnf.fr, tourné vers les publics, et le projet d'archivage interne Spar, qui correspond à la mission patrimoniale de la BnF. En parallèle à l'exposition des catalogues en RDF, les identifiants utilisés à l'interne pour l'archivage des documents ont été normalisés (ARK). Aujourd'hui, les équipes de documentalistes peuvent gérer toutes les étapes du cycle de vie d'une ressource (indexation dans les catalogues, consultation en

interne, consultation depuis le web, archivages en bases profondes). Le système assure la continuité. Les gains en termes d'efficience sont très importants.

« Etre économe et généreux à la fois » - Donner de la visibilité aux ressources les moins sollicitées, se lier à d'autres données de confiance pour enrichir les catalogues de la BnF, se concentrer sur sa valeur ajoutée et la faire rayonner.

« Etre plus utile » - La population des catalogueurs de la BnF représente 300 sur près de 2200 agents. Ce sont de grosses équipes, dotées d'un savoir-faire certain pour produire des données de qualité. Il fallait il fallait valoriser davantage l'investissement public que ce travail représente pour la collectivité, en encourageant une plus grande réutilisation par des tiers à la fois du capital historique des métadonnées BnF (plus de 13 millions de notices) et sa force de description pour toute l'édition contemporaine.

2. Data.bnf.fr : la fabrique (2)

Romain Wenz est conservateur au Département de l'Information Bibliographique et Numérique de la BNF. Sa présentation est disponible en ligne :
http://www.gfii.fr/uploads/docs/GildasIllien_RomainWenz_LOD2013.pdf

Pour de plus amples informations sur les aspects techniques du projet, se référer à :

- La fiche de présentation sur le site data.bnf.fr : http://data.bnf.fr/about
- L'article de Romain Wenz et Agnès Simon : *Des outils automatiques pour le signalement en bibliothèque. Expérimentations autour du projet data.bnf.fr,* bbf 2012 - t. 57, n° 5 : http://bbf.enssib.fr/consulter/bbf-2012-05-0039-008

Le projet

L'instruction a été lancée en 2009. La BnF a pu capitaliser sur l'expérience acquise lors du projet TELplus, soutenu par la Commission européenne, qui s'est achevé en 2008 (conversion partielle de Rameau en RDF). Une preuve de concept a été créée en 2011. En 2012, 10% du catalogue de la BnF a été exposé dans data.bnf.fr, et en 2013, les objectifs sont fixés à 20%, avec l'ambition de faire de la plateforme un service au centre d'un écosystème de données. En termes de méthodologie de projet, un marché public a été conclu avec LogiLab pour le développement et l'intégration du portail, en utilisant le logiciel libre CubicWeb. Une méthode de développement agile a été privilégiée : tester et expérimenter sur des échantillons, puis élargir progressivement. Une attention a été portée aux dimensions métier du projet, à la conduite de l'innovation et du changement.

De nouveaux usages de recherche d'information

Data.bnf.fr agrège les notices de trois bases : Gallica (environ 2 millions d'unités), Archives & Manuscrits (environ 10 millions d'unités), le catalogue général de la BnF (environ 12 millions d'unités). Les volumes concernés sont énormes. Data.bnf.fr doit être considéré comme un Hub de données entre ces trois silos, auparavant déconnectés. A l'arrivée, data.bnf.fr propose une nouvelle expérience de recherche. Il fallait répondre à un besoin de transversalité et d'unification dans la recherche d'information. Les utilisateurs souhaitent aujourd'hui pouvoir explorer l'ensemble du corpus BnF, circuler d'une base à une autre depuis la même interface.

Qualité des données : riches et obscures VS pauvres et accessibles

Une des problématiques à résoudre dans un projet comme data.bnf.fr est, quels que soient les volumes concernés, la complexité et l'hétérogénéité des bases, de produire des états de sortie à la fois compréhensibles par les humains et lisibles par les machines. La qualité des données et de la structuration des contenus est ici fondamentale. La matière bibliographique se prête en natif à faire du Linked Open Data car les données des notices sont déjà structurées finement. La question se pose alors de savoir où placer le curseur : ouvrir des données riches et obscures ou des données pauvres et accessibles ? La tendance actuelle de l'Open Data consiste à baisser le niveau de qualité au plus petit dénominateur commun pour rendre les données solubles dans le web et « hackables » par toutes les communautés. Le choix de la BnF est l'inverse : ouvrir avec un maximum de qualité les données pour ne pas appauvrir les usages. Chaque communauté peut ensuite retravailler les données pour les adapter à ses besoins. De la qualité des données dépendra les possibilités en termes de recherche d'information sur le portail, mais aussi leur référencement par les moteurs et leur réutilisation par des tiers. Cela suppose un important travail en amont, de modélisation et de catalogage rétrospectif, ce qui montre que le Linked Open Data est tout sauf la fin de catalogage. Un projet Linked Open Data doit porter la même attention à ces deux dimensions : structurer et ouvrir. Pour ouvrir les données sans erreur, il faut installer des contrôles techniques, scientifiques et juridiques tout au long du processus.

Le modèle FRBR au centre

Le modèle FRBR *(Functional Requirements for Bibliographic Records)* joue un rôle central. Il s'agit d'une modélisation orientée usages (pour les chercheurs, pour les indexeurs, pour les éditeurs ….) permettant de relier « œuvres » et «personnes » en prenant en compte les différents niveaux d'instanciation d'une ressource (œuvre, expression, manifestation, document …) et les différents niveaux de contributions des « personnes » (auteur, traducteur, préfacier, imprimeur). Ceci permet de faire des regroupements par entité pour proposer une nouvelle expérience de recherche, affranchie des contraintes de la recherche plein texte. Avec data.bnf.fr, il devient possible d'extraire toute l'information disponible dans le corpus de la BnF sur une œuvre, un auteur, une date (ex : « l'année 1472 »), ou de sortir tous les rôles d'un auteur (ex : « Balzac comme écrivain », « Balzac comme destinataire de correspondance », « Balzac comme critique d'art », …). Les catalogues sont déconstruits, les notices sont granularisées, ce qui permet d'humaniser la recherche, de proposer de nouvelles circulations.

L'importance des URI

La normalisation des identifiants est également essentielle. La norme URI (Uniform Ressource Identifier) permet d'établir des liens de navigation stables et pérennes entre les notices, les ressources et les concepts en leur attribuant un identifiant unique (URI ressources, URI http, URI actionnables…). L'utilisation d'URIs permet notamment d'enrichir le parcours de recherche des usagers en le pré-coordonnant (par exemple, la relation « philosophie » et « 20e siècle » renvoie notamment à la « phénoménologie »). Les URIs facilitent également la mise en relation avec un grand nombre de thesaurus externes (Agrovoc, thesaurus du Bureau international du travail, Géonames, etc.) et avec les catalogues d'autres Bibliothèques Nationales (Bibliothèque nationale allemande, Bibliothèque du Congrès, …). La recherche peut ainsi être prolongée dans une logique de

découverte d'information « Discovery ») au sein de catalogues connectés. Les possibilités de rebonds sont démultipliées.

Schema.org

Les notices agrégées dans data.bnf.fr ont été annotées par des métadonnées embarquées compatibles au référentiel Schema.org, ce qui facilite leur indexation et leur trouvabilité par les moteurs de recherche sur le web. Ces métadonnées permettent de faire le lien entre les moteurs de recherche et les ressources des catalogues internes via le Hub data.bnf.fr.

Réflexion métier et accompagnement au changement

Les catalogueurs sont une population en doute sur leur identité professionnelle. On parle souvent de leur disparition avec la concurrence des moteurs de recherche. Il faut souligner l'enjeu métier très fort dans le projet data.bnf.fr, qui montre l'inverse. Exploser les catalogues pour les rendre accessibles aux moteurs de recherche implique une refonte complète des process. C'est un formidable levier pour moderniser et revaloriser le savoir-faire des documentalistes. Mais il a fallu faire de la pédagogie à l'interne et faire comprendre que data.bnf.fr n'allait pas tuer les catalogueurs mais au contraire révéler leur travail et mettre en avant des morceaux « oubliés » du catalogue.

Questions / Echanges avec la salle

Quels exemples de réutilisation ?

La première réutilisation est celle des moteurs de recherche (Google, Yahoo !, Bing, Yandex) qui utilisent les métadonnées embarquées compatibles avec schema.org pour améliorer le référencement des pages. Mais data.BnF.fr se positionne dans l'Open Data et a clairement vocation à alimenter un écosystème de réutilisateurs. Il y a un manque encore de visibilité sur les cas de réutilisation. C'est une des problématiques de l'Open Data : comment tracer les usages de données diffusées hors des applicatifs, auprès du plus grand nombre ? La Licence Ouverte de l'Etat limite le suivi sur ce point : le producteur ne connaît le réutilisateur que si celui-ci se manifeste. En l'état, il faut se contenter des statistiques de téléchargement des fichiers DUMP (globaux ou segmentés) rassemblant toutes les données disponibles ou des statistiques renvoyés par le système lors de la négociation de contenus. On peut néanmoins citer la base IF Verso, la plateforme de livres traduits de l'Institut Français[11], qui réutilise les notices des œuvres françaises au modèle FRBR de data.bnf.fr pour regrouper les différentes traductions autour d'une même œuvre. (http://ifverso.com/fr). L'application CatBnF, développée par un particulier, permet la consultation de data.bnf.fr en mobilité. Il y a d'autres projets d'applications, notamment pour les chercheurs, permettant de produire des frises chronologiques en histoire littéraire, des sciences et des idées. On peut imaginer aussi des applications de réalité augmentée permettant de mettre en relation des lieux avec des auteurs et des œuvres en contexte de tourisme culturel.

[11] L'Institut français est l'opérateur du ministère des Affaires étrangères pour l'action culturelle extérieure de la France. Il a notamment pour mission de promouvoir et de diffuser la création intellectuelle à l'étranger auprès de publics concernés par les échanges intellectuels : les professionnels du livre, les traducteurs, les universitaires, les réseaux d'intellectuels, les laboratoires d'idées et les relais culturels.
http://www.institutfrancais.com/

Quel impact sur le référencement des pages « autorités » ?

Tout dépend de l'environnement concurrentiel dans lequel sont exposés les fonds. Lorsqu'il y a peu de concurrence (par exemple en histoire des sciences ou des corpus juridiques, généralement peu visibles), le référencement est dopé. Pour les corpus déjà fortement disponibles et visibles ailleurs (ex : littérature), l'impact est positif, mais moins manifeste. Globalement, la sémantisation des pages favorise largement la longue traîne des œuvres peu disponibles et peu consultées.

3. Linked Data, a radical change?

Richard Wallis est Technology Evangelist chez OCLC. Sa présentation est disponible en ligne : http://fr.slideshare.net/rjw/linked-data-radical-change

Présentation d'OCLC

OCLC est un consortium fédérant plus de 72 000 bibliothèques de par le monde. Les institutions cotisant à OCLC accèdent à l'ensemble des services proposés par l'organisation, notamment l'indexation de leurs catalogues dans WorldCat. WorldCat est aujourd'hui le plus important catalogue collectif mondial (plus de 2 milliards de notices en août 2012). Du fait de sa position, OCLC est devenu un acteur essentiel de la mise en convergence des catalogues et de l'alignement des référentiels entre bibliothèques.

Un passage au Linked Data progressif et expérimental

OCLC produit du Linked Data depuis 2010. Le passage s'est fait progressivement, par expérimentations et ajouts successifs. Aucun projet n'est aujourd'hui terminé. La démarche est incrémentale.

1. Dewey.info : un projet d'adaptation de la Dewey au web de données. Chaque entrée, classe se voit notamment attribuer un URI. L'objectif à terme est de proposer une plateforme de terminologie pour les bibliothécaires compatible aux standards des linked data. Les données de la classification sont disponibles en licence ODC-BY.
2. Passage de VIAF et FAST aux Linked Data[12]. VIAF (*Virtual International Authority File*) et FAST (*Faceted Application of Subject Terminology*) sont deux projets portés par OCLC et d'autres institutions influentes dans le monde des vocabulaires contrôlés (notamment la Library of Congress) dont l'objectif vise à faire converger au sein de thésaurus uniques à vocation universelle les listes de vedettes matières des bibliothèques partenaires. Les données des thésaurus FAST et VIAF sont placées sous licence ODC-BY.
3. Passage de WorldCat au Linked data :
    - Incrustations des métadonnées compatibles Schema.org (RDFa) pour optimiser le référencement et la réutilisation des contenus par les moteurs.
    - Interconnexion des référentiels internes (Dewey.info, FAST, VIAF) et alignements avec des référentiels externes (LCSH[13], LCNAF[14])

[12] http://www.oclc.org/news/releases/2011/201171.en.html

[13] Library Of Congress Subject Heading : la liste de vedettes matières de la Bibliothèque du Congrès

[14] Library of Congress Name Authority : la liste d'autorité de la Bibliothèque du Congrès

- Attribution d'un DOI aux objets (Digital Object Identifier)[15]
- Toutes les données et métadonnées sont sous licence ODC BY (soit 1.2 millions de ressources, 80 millions de triplets).

A noter : La licence ODC BY (*Open Data Commons Attribution License*) autorise la réutilisation des notices WorldCat, y compris pour des usages commerciaux, sans autorisation préalable, avec pour seule contrainte la mention de la source WorldCat et le respect d'un code de bonnes pratiques de réutilisation : le "WorldCat Rights and Responsibilities".

Le passage de WorldCat à Schema.org

Depuis juin 2012, chaque page du catalogue WorldCat est progressivement balisée par des métadonnées embarquées RDFa conformes au modèle Schema.org (7 à 10% des pages balisées au moment du lancement en juin 2012). Schema.org est une initiative conjointe des moteurs de recherche Google, Bing, Yahoo, et Yandex, dont l'objectif est de rendre les contenus des pages web "compréhensibles" par les moteurs de recherche grand public (SEO sémantique). Schema.org n'est pas un vocabulaire pensé pour les bibliothèques. Pour une institution culturelle ou une bibliothèque, ouvrir les pages de son catalogue aux moteurs de recherche était encore impensable il y a 5 ans, mais c'est un choix tactique. D'une part, il s'agit de suivre les usages : les moteurs de recherche web sont de loin le premier outil de consultation numérique, les catalogues des bibliothèques doivent y être référencés. D'autre part, Schema.org a une portée très générique. Le modèle ne concerne pas uniquement les bibliothèques mais toutes les institutions culturelles (archives, radios, TV…). Les possibilités d'interconnexion avec d'autres univers sont facilitées. Par ailleurs, il faut aussi souligner la rareté d'une initiative comme Schema.org dans le monde de l'information numérique et de la recherche. Des acteurs industriels en concurrence frontale ont collaboré pour mettre en place des standards communs afin de développer le SEO Sémantique. Une telle collaboration était nécessaire car il y avait un vrai besoin d'améliorer la pertinence des résultats et la qualité du filtrage dans la recherche sur le web. Mais Schema.org ne répond pas à tous les besoins. Des améliorations sont attendues pour faciliter l'échange et le partage de données entre bibliothèques, communautés. Dans cette optique, OCLC intervient au sein du « Schema Bib Extended Community Group » du W3C[16], dont la mission est d'étendre la portée du modèle pour affiner le balisage des pages web et faciliter le partage des données.

"User-centered is data-centered ! »

Pour l'utilisateur, l'expérience est transformée. Le passage au Linked Data permet de gagner en richesse sur les résultats, en fluidité et en temps sur la recherche. WorldCat devient une source exhaustive sur des concepts, des sujets, et plus seulement des mots-clés. Le système renvoie toutes les ressources disponibles sur un concept dans les catalogues de toutes les bibliothèques membres.

[15] Un DOI (Digital Object Identifier) est un identifiant unique et universel d'un objet numérique, quel qu'il soit (article, ebook, fichier sonore, vidéo, etc.). Il est attribué par le producteur ou le diffuseur de l'objet par l'intermédiaire d'une agence DOI agrée par la DOI Fondation. L'identifiant DOI est amené à jouer un rôle prépondérant dans la gestion des métadonnées sur les réseaux numériques.

[16] http://www.w3.org/community/schemabibex/

Des enrichissements avec des ressources tiers sont possibles (DBpédia). La puissance de l'agrégation est démultipliée et les pratiques de recherche s'affinent. Avant, la logique de rebonds vers la source prévalait : l'utilisateur sortait de WorldCat pour consulter le catalogue source et y revenait si nécessaire, en réitérant la démarche. Avec les données liées, toute l'information essentielle est poussée vers l'utilisateur qui reste dans l'environnement WorldCat le temps de la recherche : *« user-centered is data-centered !»*.

BIBFRAME : future standard de description bibliographique dans les Linked Data

OCLC travaille aussi au développement de BIBFRAME en partenariat avec la Library of Congress. Complémentaire à Schema.org, BIBFRAME a vocation à remplacer MARC pour la description bibliographique dans l'univers du web des données. Historiquement, MARC a été porté par la Bibliothèque du Congrès depuis les années 60 pour faciliter l'indexation des notices dans l'optique de l'informatisation des catalogues (*MAchine Readable Catalogue*). MARC correspond à une époque où les catalogues n'avaient pas vocation à sortir des applicatifs et n'est pas optimisé pour l'interconnexion des informations. BIBFRAME repose sur le principe selon lequel la bibliothèque de demain sera principalement numérique, et projetée en réseau. Les possibilités pour l'interconnexion et le portage des données sont beaucoup plus grandes.

« Bibliographic Description As Part of the Web »[17]

Les bibliothèques se sont tenues à distance du pilotage du web et sa normalisation, jusqu'à présent tenue par les acteurs industriels et les instances politiques. Avec le développement des Linked Data, la description bibliographique peut devenir un puissant levier de transformation du web vers le web des données. Les bibliothèques peuvent jouer un rôle nouveau dans la gouvernance des réseaux, en s'impliquant dans l'élaboration de nouveaux standards comme BIBFRAME, en diffusant des bonnes pratiques et en facilitant les passerelles avec d'autres univers (médias, musées, archives …). Le Linked Open Data est un des rares lieux d'innovation technologique où l'on peut faire coopérer les acteurs et les référentiels. L'exemple de Schema.org montre qu'il y a une complémentarité à trouver avec les acteurs du *search* et qu'on ne peut plus penser la relation bibliothèques – moteurs de manière uniquement concurrentielle.

Les enjeux : *« Stop Copying, Start Linking ! »*

Aujourd'hui, il faut apporter de la cohérence et de la coordination au web de données. Les acteurs ne doivent pas faire du Linked Data seuls. L'interconnexion entre les fonds et les référentiels est au cœur des principes du web de données. Les institutions doivent coopérer. Il n'est pas nécessaire de s'interconnecter à tout le monde. Les Hubs les plus centraux, auxquels tout le monde se connecte, suffisent car ils redistribueront. DBpédia peut constituer un point de départ privilégié. C'est une initiative du groupe « *Semantic Web Education* » du W3C qui remonte à 2006. Ce groupe devait à l'origine réfléchir à faciliter l'adoption du web sémantique, en perte de vitesse. DBpédia a été conçu pour démontrer l'intérêt des données liées et le projet est petit à petit devenu le point central dans l'édification du web de données. Autour de DBpédia, il y a d'autres acteurs qui agissent en coordination (les moteurs de recherche avec Schema.org, les Bibliothèques Nationales, les grands

[17] Devise de l'initiative BIBFRAME

médias comme la BBC[18], …). C'est à ces Hubs qu'il faut se connecter pour accroître le rayonnement de ses propres contenus et les enrichir.

"From cataloguing to catalinking"[19]
Le métier des documentalistes est lui-même amené à évoluer avec la déconstruction des catalogues : du *« catalogage »* au *« cataliage »*. La notion d'exemplaire (manifestation individuelle d'une œuvre) ne sera plus aussi centrale. Les usagers devront pouvoir circuler entre plusieurs niveaux de représentation, du concept à l'œuvre et ses différentes manifestations (FRBR). Les bibliothécaires devront apprendre à parler et à contribuer au développement des langages du web. Le web de données n'est pas une rupture pour le métier des bibliothécaires mais une évolution.

Questions – Echanges avec la salle

Aujourd'hui, les Bibliothèques Nationales éditent les fichiers d'autorités dans leur langue. « Victor Hugo » et ses attributs dans Rameau n'est pas « Victor Hugo » dans le LCSH. Comment internationaliser ?
VIAF doit apporter une réponse à ce problème en facilitant l'alignement et la convergence des listes entre les institutions. VIAF synthétise aujourd'hui partiellement de nombreux registres nationaux, dont RAMEAU. Il faut encore progresser, mais la convergence doit être pensée sur le temps long. Avec le web des données, il y a un petit nombre de ressources très importantes. L'interopérabilité n'est pas un rêve. Mais une chose est certaine, il n'y aura jamais un standard unique, seulement des mashups de référentiels et des transpositions. Il est aussi intéressant de suivre le développement des standards interprofessionnels comme l'ISNI (*International Standard Name Identifier*) et de les adapter aux Linked Data. L'ISNI doit servir d'identifiant unique pour toute la chaîne de l'information (éditeurs bibliothèques, agrégateurs, diffuseurs, …). Une fois lié à VIAF, ce sera un levier d'unification très puissant pour le monde de l'information.

Quel sera le facteur décisif dans l'adoption des standards sémantiques chez les producteurs d'information ?
Il y a le moteur politique mais il penche aujourd'hui nettement plus en faveur de l'Open Data que du Linked Open Data. Dès lors que les fournisseurs de solutions de gestion de l'information (SGIB, ERMS, agrégateurs …) concevront des plateformes intégrant nativement les standards du web des données, les acteurs du contenu devront suivre.

4. Les enjeux de la réutilisation : le cas des Archives

Emmanuel Condamine est Directeur Général Généalogie de NotreFamille.com. Sa présentation est disponible en ligne[20] :

[18] http://fr.slideshare.net/metade/linked-data-on-the-bbc
[19] Eric Miller – Zepheira.

http://www.gfii.fr/uploads/docs/EmmanuelCondamine_LOD2013.pdf

NotreFamille.com

NotreFamille.com est un grand réutilisateur de données d'archives sur le marché de la généalogie. Au titre du développement de l'activité généalogique de son site web généalogie.com, le groupe NotreFamille.com travaille pour l'accessibilité des données "culturelles", notamment les archives de l'état-civil et des recensements français. NotreFamille.com souhaite procéder à la numérisation et la transcription systématique de ces documents, afin de proposer à ses utilisateurs un moteur de recherche patronymique centralisé permettant l'exploration profonde des différents silos rassemblés.

Le paradoxe généalogique français

Il y a un paradoxe généalogique en France. La richesse du fonds archivistique est unique au monde. La généalogie passionne les Français : 61% des Français ont déjà fait une recherche sur l'histoire de leur famille[21], et le nombre d'associations de généalogistes amateurs est très important. Le contenu et l'intérêt sont là, mais le marché de la généalogie ne se développe pas et les besoins des utilisateurs ne sont pas satisfaits, ceci alors même que 90% des archives départementales ont déployé des plans de numérisation ambitieux.

Pas d'outil pour la recherche fédérée par patronyme au niveau national

Les raisons sont multiples : méconnaissance par le grand public des sources et des modalités de recherche[22], fracture numérique et inégalité d'accès selon les territoires, activités chronophages… Surtout, il n'existe aucun outil centralisé permettant la recherche par patronyme sur l'ensemble du territoire. Les usagers sont contraints d'interroger chaque portail départemental où les données sont le plus souvent indexées de façon sommaire. Depuis 5 ans, NotreFamille.com travaille à combler ce manque, mais s'est affronté à plusieurs reprises au refus d'archives départementales.

La problématique des archives

Les archives départementales bénéficient d'une grande autonomie dans la définition des politiques de diffusion de leurs données. Elles sont en situation de monopole de fait sur les données qu'elles détiennent, notamment pour les registres d'Etat Civil. Par ailleurs, la généalogie n'entre pas directement dans le périmètre de leur mission de service public. Or, les données qui intéressent les acteurs de la généalogie ne représentent que 2 à 3% des fonds archivistiques. Dans un contexte de baisse des budgets, les investissements à consentir pour mettre à disposition ces données à des réutilisateurs sont très importants (coûts de transcription, numérisation, mise en ligne, indexation,

---

[20] Pour de plus amples précisions sur le cadre juridique et la spécificité de la réutilisation des données d'archives, se référer à la présentation de Lionel Maurel : ***« Etat des lieux de l'Open Data culturel, en France et en Europe »,*** en tête de cette synthèse

[21] 61% des Français ont déjà fait une recherche sur l'histoire de leur famille (Ipsos – Avril 2010)

[22] La recherche généalogique nécessite une expertise et du temps. Il y a de nombreux paramètres : l'époque à laquelle le registre d'Etat Civil a été constitué, l'histoire des migrations familiales, les variantes patronymiques, …

etc.). Le *crowdsourcing* est insuffisant : la mise en place d'outils de contrôle qualité et de validation a un coût, et les résultats sont toujours parcellaires.

Les vertus de la réutilisation

NotreFamille.com se positionne comme un grand réutilisateur de données. Il y a encore un important travail de sensibilisation à mener dans le champ culturel sur la réutilisation commerciale. Celle-ci est trop souvent perçue comme une forme d'exploitation ou de privatisation d'un bien commun (*« payant = méchant »*). La valeur ajoutée apportée par les réutilisateurs est rarement perçue. En tant qu'intermédiaire entre les producteurs de données et les utilisateurs finaux, NotreFamille.com se considère comme un « facilitateur d'usages ».

Valeur ajoutée du projet de NotreFamille.com

La composante technologique est très forte dans le projet de NotreFamille.com. Les cycles d'investissement et de R&D de la société se concentrent sur trois étapes :

1. **Accès :**
   - Nécessité de constituer un fonds généalogique, historiographique et iconographique suffisamment riche pour permettre l'exploration profonde des différents silos.
   - Les données d'Etat Civil ne suffisent pas, il faut des informations contextuelles (ex : collections sur les prisonniers de la seconde Guerre Mondiale, bulletins des Lois, Tableau d'Honneur de la Grande Guerre, etc.)
   - Achats de licences pour les images déjà numérisées
   - Numérisation des fonds originaux pour lesquels NotreFamille.com a réussi à négocier les droits de transcription - reproduction
2. **Indexation :**
   - Un poste d'investissement et de R&D très important car c'est sur cette partie que se concentre la valeur d'usage du service proposé par NotreFamille.com.
   - Transcription systématique des informations : l'automatisation est impossible car l'OCR n'est pas performant sur des registres d'Etat-Civil (hétérogénéité de la structure selon les départements, problème de la graphie, vocabulaires, etc.)
   - Constitution d'une base de données rassemblant plusieurs silos : archives départementales pour lesquelles les droits ont été négociés, et fonds privés (rachat des éditions SWIC en 2007, intégration du fonds de généalogie successorale Coutot en 2009).
   - Création de référentiels (patronymiques, géographiques, etc.) : champs de R&D important pour permettre la recherche par noms propres croisée à des critères de provenance géographique.
   - Création d'index nominatifs de recherche : on peut penser l'activité généalogique comme la « mère du linked data » dans la mesure où elle consiste essentiellement à lier des personnes entre elles par lien de parenté. Les généalogistes ont développé des modèles, des approches formalisées, bien avant le développement des TIC (généalogie ascendante, descendante, successorale, etc.).
   - Interaction images-sources : les résultats fournis par le service généalogie doivent être contextualisés et enrichis, la recherche doit être multidimensionnelle pour répondre aux besoins d'une enquête généalogique.

3. **Diffusion**
    - Création d'une interface web : moteur de recherche unifié pour l'exploration des silos depuis 2011 (arbres généalogiques, relevés d'Etat Civil, etc.)
    - Hébergement des données
    - Mise en visibilité et animation des fonds (référencement)

La valeur ajoutée de NotreFamille.com tient essentiellement dans ces investissements technologiques et sa connaissance métier. Le travail de modélisation des requêtes et de qualification de la donnée est essentiel. Les problématiques sont nombreuses : comment gérer les variantes patronymiques quand les algorithmes de rapprochement phonétiques sont inopérants ? Comment gérer les homonymes ? *etc*.

Intérêt pour les détenteurs

Le projet de NotreFamille.com est parfois perçu comme une appropriation, mais les avantages sont nombreux. Les fonds détenus font l'objet d'usages qu'ils n'auraient pas connus sans l'intervention de NotreFamille.com sans surcoûts supplémentaires (importance de la transcription systématique). Le travail et l'investissement réalisés par la collectivité sont valorisés grâce à des intermédiaires comme NotreFamille.com qui portent les fonds à la connaissance de nouveaux publics. Des revenus additionnels peuvent être générés pour la collectivité par le biais de redevances qui associent les détenteurs de fonds au succès de l'entreprise. Des collaborations scientifiques sont également envisageables dans la construction de référentiels communs et de formats standardisés d'échanges.

Des obstacles importants demeurent

Le projet genealogie.com a été lancé en 2006 et de nombreux obstacles demeurent. La généalogie cumule toutes les difficultés :

- **Difficultés culturelles :** les collectivités doivent accepter de voir leurs fonds sortir de leurs systèmes. La réutilisation par un opérateur économique peut être également mal perçue par les associations de généalogistes, qui étaient jusqu'à présent les seuls à utiliser des documents d'archives.
- **Difficultés politiques :** les investissements consentis pour numériser les fonds et les diffuser sur les portails départementaux sont importants, et ces initiatives sont considérées comme des vitrines par les élus.
- **Difficultés administratives :** il n'y a pas de guichet unique, NotreFamille.com doit négocier les droits département par département.
- **Difficultés juridiques :** l'exception culturelle entretient le flou mais en réalité, la plupart des données concernées ne rentrent pas dans le champ de l'exception. L'argument des « données personnelles » est souvent invoqué, le problème a été tranché par la CNIL : les données nominatives concernant des personnes mortes depuis plus de 70 ans sont communicables. Pour les données les plus délicates, les réutilisateurs doivent demander une autorisation de la CNIL. NotreFamille.com a obtenu cette autorisation fin 2011.

Quels signaux positifs ?

NotreFamille.com a signé une première licence avec le département du Rhône[23] en novembre 2012 et une seconde avec la Vendée en mars 2013[24]. Après les contentieux qui ont opposé la société aux archives départementales du Cantal et de la Vienne notamment, ces avancées attestent d'une stabilisation en cours du cadre légal. Il faut y ajouter l'arrêt de la Cour Administrative du Tribunal de Lyon en juillet 2012 qui a permis de clarifier la question de l'exception culturelle[25]. Mais la récente décision de la Cour d'Appel du Tribunal de Poitiers montre que le combat n'est pas encore fini. Il faut encore faire accepter la légitimité de la réutilisation commerciale des données généalogiques[26].

Tendre vers le modèle britannique ?

Le modèle britannique montre qu'il y un ROI direct et indirect à trouver pour les archives nationales autorisant la réutilisation de leurs données. Le marché anglais se partage entre 3 grands réutilisateurs. Des centaines d'emplois ont été créés et des partenariats mis en place entre ceux-ci et de grands opérateurs publics. BrightSolid, un prestataire de numérisation de masse spécialisé dans le domaine culturel, a noué plusieurs partenariats avec The National Archives, le site des Archives Nationales du gouvernement britannique sous la tutelle du Ministère de la Justice. Les fonds détenus par The National Archives sont progressivement transcrits et numérisés (census, crime courts & convicts collection, etc.) moyennant un système de redevances. Le trafic sur le site The National Archive a considérablement augmenté, les contenus sont davantage sollicités. 2 millions de £ de revenus additionnels ont été générés par la mise en place des redevances. Les fonds numérisés sont réutilisés par Brightsolid pour alimenter FindmyPast, son principal service commercial[27].

Questions - échanges avec la salle

Privé / Public : concurrence ou complémentarité ?

La réutilisation des données d'archives rappelle par certain point le débat sur la concurrence public / privé au début du projet Légifrance. Lorsque le projet est sorti, les éditeurs juridiques avaient peur que le dispositif, financé sur fonds publics, ne fasse concurrence à leurs activités. L'arbitrage s'est fait sur la notion de valeur ajoutée. Légifrance répond à une mission de service public de la DILA (diffuser la loi – que nul n'est censé ignorer - au plus grand nombre) et repose sur une logique quantitative (diffuser le corpus légal de manière exhaustive mais brute). Les éditeurs développent des services complémentaires à valeur ajouté nécessaires aux juristes dans l'exercice de leur métier

[23] http://www.gfii.fr/fr/document/notrefamille-com-premiere-signature-d-une-licence-de-reutilisation-de-donnees-publiques-avec-le-departement-du-rhone

[24] http://www.rfgenealogie.com/s-informer/infos/nouveautes/notrefamille-signe-une-nouvelle-licence-avec-la-vendee

[25] http://www.gfii.fr/fr/document/notrefamille-com-se-felicite-de-l-arret-du-4-juillet-2012-de-la-cour-administrative-d-appel-de-lyon

[26] http://scinfolex.wordpress.com/2013/02/01/open-data-rip-la-reutilisation-des-informations-publiques-bientot-dissoute-dans-le-droit-des-bases-de-donnees/

[27] Findmypast.co.uk, .com et .ie permettent à tous les utilisateurs des marchés britanniques, irlandais et néozélandais d'effectuer des recherches généalogiques profondes sur une base de données unifiée. Le service permet notamment d'agréger et corréler l'intégralité des dates de naissance, de mariage et d'enterrement pour réaliser des arbres de familles, ce qui lui a valu de recevoir la Queen's Award for Innovation en 2007.

(commentaires, versioning ...). La situation est un peu différente pour les archives car la généalogie n'entre pas dans leur mission de service public. On peut toutefois envisager un modèle similaire, avec la création d'un guichet public unique pour centraliser les données au format brut. A charge ensuite aux réutilisateurs d'y apporter de la valeur ajoutée à travers leurs services.

Partage de la valeur privé / public : quelles pistes ?
De nouvelles formes d'échanges et de partages de la valeur entre réutilisateurs et détenteurs peuvent aussi être explorées. La licence ODBL pose notamment les bases d'une rétribution du détenteur proportionnelle à la valeur générée par le réutilisateur sur ces données. Quelle est la position de NotreFamille.com ?

Des systèmes de redevances existent déjà pour permettre aux archives de générer des revenus additionnels. La licence signée avec les départements du Rhône et de la Vendée montre qu'il y a un intérêt pour les archives départementales à instaurer ces systèmes. Sur les autres formes de coopération, NotreFamille.com est ouvert à l'échange scientifique de données pour compléter les fonds archivistiques, mais il n'est pas concevable de livrer son index. Celui-ci est le cœur de métier et la clé de l'avantage concurrentiel de la société. C'est le fruit de 10 années de R&D et il faudra plusieurs années pour le rentabiliser.

5. The Digital Museum As A Platform

Jacob Wang est le responsable des activités digitales du Muséum National du Danemark (NATMUS). Sa présentation est disponible en ligne sur le site du GFII :
http://www.gfii.fr/uploads/docs/JacobWang_LOD2013.pdf

Présentation
Le Muséum National du Danemark (NATMUS) a été fondé en 1857. Il emploie aujourd'hui 550 salariés et est doté d'un budget annuel de 35 millions d'euros. Les collections numériques du musée représentent 10 millions d'objets et 2 millions d'images.

Les musées face au virage numérique et à l'Open Data
Les musées sont aujourd'hui dans la situation des bibliothèques il y a 20 ans, en pleine transition numérique. Mais la diversité des outils disponibles et la baisse des coûts démultiplient les possibilités. Pour le NATMUS, il y a deux évolutions majeures :

- L'usage des réseaux sociaux pour la communication envers les communautés
- En termes de gouvernance de SI, le passage d'une logique de catalogues intégrés, centrés sur la notion de « produit », à une logique de de flux, de plateforme et d'API.

La réutilisation des données d'institutions culturelles est source de création de valeur économique, mais aussi et surtout de valeur d'usages (enseignement, recherche, diversité culturelle …). Pour le NATMUS, promouvoir cette valeur d'usage est l'objectif principal d'une démarche Open Data.

Un Hackaton pour inspirer les équipes et rénover l'image du NATMUS
En octobre 2012, le Muséum a expérimenté son premier Hackaton à partir des données libérées au cours de l'année 2011. Il s'agissait d'une part d'encourager la réutilisation du patrimoine, mais aussi de rentrer en contact avec les communautés de développeurs car ce sont eux qui peuvent inspirer aux équipes du musée une culture de la donnée. Pour un professionnel de la conservation, imaginer des usages innovants à partir des données sur les collections ne va pas de soi. C'est une culture propre aux communautés de Hackers. En termes marketing, jouer sur l'image du « hacking » est aussi un moyen de rénover l'image de l'institution, de tendre la main aux jeunes pousses, ceci pour un coût très modique. En utilisant les réseaux sociaux, notamment Twitter, il est très simple de viraliser la communication et d'intéresser les développeurs. En termes de budget, cela représente 1450 euros de subventions de Microsoft, et 4x650 euros aux frais du Muséum, pour un investissement temps lui aussi relatif (2 semaines de mobilisation intense seulement en amont de l'événement). L'événement a réuni 35 « développeurs » (programmateurs, designers d'interfaces, graphistes) et un pool de journalistes spécialisés pour commenter l'événement sur les réseaux, le documenter. Il était aussi important d'inviter d'autres institutions (Danish Agency For Culture, The State Archives, Royal Library of Danemark), car l'initiative du Muséum était très observée et devant servir de modèle.

Quels résultats ?
A l'arrivée, 10 prototypes ont été réalisés en 2 jours :

- Kulturarev (lauréat) : une application de réalité augmentée pour smartphone android donnant délivrant en temps réel des informations contextuelles sur les monuments
- Datafest
- Adopt a mound
- NEar Something
- Dead neighbours
- Image analysis of artworks
- Image mozaic, smashup
- Danebrosmaend
- SMK Geoscraping
- Mapping Artworks

Quel ROI pour le Muséum ?

D'un point de vue immédiat, l'événement a donné une belle visibilité à la démarche et a permis de toucher le public des développeurs. Des contacts ont été pris pour des projets plus structurants concernant le Muséum mais aussi les autres institutions invitées. La plupart des projets concernent la création d'API officielles pour le Muséum, pour la Royal Library, ou pour les Archives Nationales. Pour un musée qui souhaite se transformer en plateforme, mettre à disposition des communautés une API ouverte et cohérente facilite l'animation d'un écosystème autour de ses données. Pour le NATMUS, il est important de ne pas faire de l'Open Data de manière isolée et de collaborer avec les autres institutions car la mise à disposition des « données culturelles » représente une opportunité très importante pour le secteur culturel de réinvestir l'espace social.

Perspectives

L'expérience du Hackathon a révélé deux perspectives technologiques pour le Muséum : le mashup des données (mélanger les données du musée à celles d'autres institutions, combiner des applications créées par d'autres pour développer des services originaux …), et les technologies sémantiques. Le Muséum a fait le choix d'ouvrir une partie des données, d'expérimenter les réutilisations à travers un Hackaton, avant de structurer. Mais l'ouverture doit être globale, et elle est aujourd'hui partielle. Les données non-structurées sont les plus riches et ne peuvent pas être facilement réutilisées. Les Muséums connaissent aujourd'hui la migration numérique qu'ont connu les bibliothèques il y a 20 ans, mais le spectre des possibles et beaucoup plus large car tout se télescope (web 2.0, sémantique, IA, Big Data, Open data,…).

Un nouveau Hackaton à l'automne 2013

Les budgets sont en baisse. Le Hackaton est un moyen d'apporter progressivement de nouvelles compétences aux équipes, dans un nouveau modèle économique. Le prochain est prévu pour l'automne 2013 avec de nouveaux jeux de données. Il devra réunir une cinquantaine de développeurs. Les profils seront plus diversifiés (storytellers, enseignants …) et l'événement sera thématisé (culture tourisme, éducation, jeu …). L'idée sera de tester l'API officielle du Muséum, et peut-être d'envisager des croisements avec les données d'autres institutions. Aujourd'hui, toutes les applications sont développées par des tiers. A terme, il faut aussi viser la réutilisation interne : chaque service du musée doit pouvoir développer ses propres applications à partir de ses données et celles des autres.

6. Faciliter la publication des données sur le web : présentation du projet DataLift

Alexander Polonsky est Directeur Marketing chez MONDECA. Sa présentation est disponible sur le site du GFII : http://www.gfii.fr/uploads/docs/AlexanderPolonsky_LOD2013.pdf

Présentation de DataLift

Datalift est un projet de recherche co-financé par l'ANR sur les méthodologies et les outils de production de données ouvertes et liées dans le web des données. Le financement court de septembre 2010 à mars 2014. Concrètement, il s'agit de mettre en place une plateforme ouverte pour faciliter la production et la publication de données liées. Le développement du web des données est aujourd'hui freiné par la complexité des projets. Il y a de nombreuses piles technologiques. Le fractionnement des applications permettant de conduire une à une toutes les étapes d'un projet rend difficile l'appropriation des méthodes et des outils. Face à la grande diversité des référentiels et des vocabulaires, il y a un besoin de centraliser et qualifier l'ensemble des ressources disponibles pour orienter les porteurs de projets (quel vocabulaire utiliser dans quel contexte ?). Le développement des Linked Data souffre aussi du manque de lisibilité et de l'isolement des démarches. Trop d'acteurs font de l'Open Linked Data seuls, en publiant leurs données avec leurs propres vocabulaires. A l'arrivée, les données sont bien ouvertes, dans un format partagé, mais prisonnières d'un silo sémantique. A contrario, l'ingénierie sociale est un accélérateur puissant pour le développement du web de données et un facilitateur de projets.

Solidité et crédibilité des partenariats

Datalift rassemble toute la chaîne des acteurs du web de données sous la forme d'un consortium :

- Des fournisseurs de données (IGN, INSEE, …)
- Des instances de pilotage, de coordination et de communication (FING)
- Des intégrateurs de solutions dans le CLOUD (ATOS)
- Des fournisseurs de technologies sémantiques (MONDECA)
- Des instituts de recherche et des grandes écoles (Mines Telecom, Université de Montpellier, INRIA)

Centraliser les applications

La plateforme Datalift centralise une série de modules entièrement intégrés permettant de réaliser toutes les étapes d'un projet Linked Open Data au sein d'une même application :

- Publication des données via SPARQL End-Point (paramétrage privé ou public possible)
- Manipulation des données des données via des modules de visualisation, navigation, requêtage, alignement
- Transformation des données en RDF
- Réconciliation et interconnexion avec des entités partagées (LOD – *Linked Open Data*)
- Réconciliation et interconnexion avec des vocabulaires partagés (LOV – *Linked Open Vocabularies*)

- Conversion de formats sources (CSV, SQL, XML) en RDF pour nettoyage

Quelles données ?

Les fournisseurs partenaires du projet ont déjà publié un certain nombre de données de référence en utilisant la suite d'applications DATALIFT :

- Ontologie topographique et données administratives sur data.ign.fr
- Code officiel géographique, Code et nomenclatures NAF, données de recensement extraites sur data.insee.fr

ZOOM sur LOV *(Linked Open Vocabularies)*

L'identification des vocabulaires pertinents et leur maintenance sont des points durs dans la conduite de projets Open Linked Data. LOV est un sous-projet de Datalift, intégré à la plateforme. Il s'agit d'un mapping des vocabulaires disponibles que l'utilisateur peut explorer à travers une interface de visualisation. Celle-ci permet notamment d'observer les relations de dépendances entre chaque vocabulaire qui sont catégorisées et qualifiées par contexte d'usages

Quelques fonctionnalités de LOV

« LOV Search » permet une recherche par type de vocabulaire ou par propriété (ex : « j'ai besoin d'une propriété pour décrire la « latitude », quels sont les vocabulaires disponibles et pertinents ?). Des fiches documentent les vocabulaires.

« LOV aggregator » réunit l'ensemble des vocabulaires disponibles dans un seul dump RDF dans un endpoint SPARQL. Un outil de mapping des références permet de de visualiser les relations entre vocabulaires (extension, dépendance).

« LOV Stats » est un outil de mesure de la « popularité » des vocabulaires et des propriétés (LOV Distribution, LOV property).

Un cas d'usage sur la publication des données des collèges de Gironde fournies par l'INSEE est disponible sur la présentation d'Alexander Polonsky :
http://www.gfii.fr/uploads/docs/AlexanderPolonsky_LOD2013.pdf

Quels usages ?

Le catalogue répertorie 363 vocabulaires, qui sont catégorisés, qualifiés par contexte d'usages, par relations de dépendance, et mis à jour. DataLift ne recense que les vocabulaires compatibles aux standards du web sémantique, avec un niveau de développement suffisant pour faire l'objet d'une réutilisation. On constate un noyau dur très réutilisé, et une majorité de vocabulaires peu ou pas utilisés. DublinCore est le plus réutilisé en nombre de vocabulaires, SKOS est le plus populaire en nombre de classes et propriétés réutilisées, FOAF et Uniprot sont les plus populaires en nombre d'occurrences de la classe / propriété dans tout le web de données.

7. La stratégie de Radio France dans la production et la gestion de données musicales

Caroline Wiegandt est Directrice de la Documentation à Radio France. Isabelle Canno est Chargée des projets transverses à la Direction de la Documentation à Radio France. Leur présentation est disponible en ligne :
http://www.gfii.fr/uploads/docs/CarolineWiegandt_IsabelleCanno_LOD2013.pdf

Radio France : quelle offre radiophonique ?
Radio France, 1er groupe radiophonique français, regroupe 3 stations nationales (France Inter, France Culture, France Musique), 3 stations multi-villes (FIP, LeMouv', France Info), un réseau de 43 stations locales (France Bleu) et 4 formations musicales (Orchestre National de France, Orchestre Philarmonique de Radio France, Chœur de France, Maîtrise de Radio France).

La production musicale générée est colossale. Il faut distinguer la production traditionnelle et la production nouveaux médias :

- **Production traditionnelle :**
  - Production radio (+ de 1000 titres diffusés par jours sur les antennes, + de 800 concerts enregistrés et diffusés par an, concerts live)
  - Production des formations musicales (50 créations par an)
- **Production nouveaux média :**
  - Offre spécifique pour les nouveaux médias AOD et podcasts sur les différentes plateformes
  - Nouvelles offres spécifiques sur internet (chaînes vidéos comme « Nouvossons », « Concerts en vidéos »

Problématiques de la gestion des données
La gestion des flux de données musicales est centrale à Radio France. Les problématiques sont complexes. Il y a une multiplicité des lieux et des instances de saisie (+ de 3000 salariés répartis sur l'ensemble du territoire), la volumétrie des données musicales est considérable et le suivi des droits d'auteurs pour les musiques vivantes et les disques diffusés doit être assuré en permanence. Il faut distinguer les bases de gestion internes des bases d'écoutes :

- **Bases internes :**
  - Base de gestion des concerts des formations musicales de Radio France et édition des programmes de salles (Direction de la musique)
  - Base éditoriale de production des programmes (Préparation des grilles et conduites d'antennes)
  - Fichiers Word et Excel sur les postes individuels
  - Outils de programmation musicale
  - Outils de diffusion
- **Bases externes :**
  - Discothèques numérique (1 693 650 titres, 129 300 albums) et documentaire (75 616 notices 78 tours, 803 148 notices Vinyls + CD, 31 000 notices œuvres)

- Bibliothèque musicale : livres et revues, partitions et matériels d'orchestre (146 619 notices, 27 600 notices concerts, 62 100 œuvres)
- Documentation sonores : description des émissions et concerts indexés par l'INA (55 600 notices)
- Documentation d'actualité (agrège les articles issus de la PHN et de la PQR)
- Sites internet des chaînes et des émissions

L'offre musicale augmente. Pour s'adapter, Radio France entre progressivement dans la logique des Big Data en important directement les flux des lieux de production (labels, salles de concert …). Harmoniser la publication des données musicales devient indispensable.

Penser « Linked Enterprise Data » avant de penser « Linked Open Data »

Il y a un vrai besoin de mutualiser et normaliser la saisie des données. Avant de lier et de publier les données, une réflexion de fond sur les référentiels doit être conduite à l'interne, ce qui revient à déployer une gouvernance des données. Le champ de la documentation musicale n'est pas aussi bien modélisé et structuré que celui de l'IST ou de la littérature grand public. La démarche est en cours chez Radio France. Les différents services, antennes, doivent s'entendre sur un référentiel commun. Ceci implique a minima de trouver un consensus sur le schéma de description des œuvres et tout ce qui les caractérise (genre, instruments, …) et sur les formats d'export. La réutilisation interne est le premier objectif : les données publiées doivent être réutilisables par tous les documentalistes de Radio France.

Importance de la qualité des données

La qualité des données publiées est essentielle, aussi bien pour la réutilisation interne qu'externe. La stratégie de Radio France est de prendre le temps de produire des données de qualité, de structurer à l'interne pour mieux ouvrir. Ceci s'oppose au discours selon lequel il vaut mieux ouvrir rapidement un maximum des jeux quitte à simplifier l'information pour atteindre la masse critique. La structuration des contenus et des SI prend du temps, mais démultiplie l'impact de l'ouverture. La réflexion de Radio France sur la publication dans le web des données est en cours. L'ouverture se fera progressivement une fois le niveau de qualité de la structuration et de la description atteint.

Mettre en place une « gouvernance des données »

C'est un point dur de la démarche, il faut prendre en charge tout le cycle de vie des flux, tenir compte de leur hétérogénéité et envisager les rétroactions possibles : maîtriser les process d'acquisition, de diffusion et se demander à quel moment un flux pourra enrichir le référentiel. En ce sens, la démarche de Radio France est inverse de celle de la BNF. La BNF produit des référentiels en amont des usages, là où Radio France part des usages pour produire des référentiels car ceux-ci n'existent pas ou sont incomplets.

Modéliser le champ musical : la notion d'œuvre au centre

Le modèle vers lequel Radio France tend sera centré sur la notion d'œuvre musicale. C'est une notion complexe à modéliser dans toutes ses dimensions. Il y a l'œuvre (titre, auteur) et il y a tout ce qu'il y a autour et qui la détermine :

- Ses différentes interprétations (noms des interprètes)
- Ses différentes éditions et supports
- Son genre (le « jazz » chez Fip n'est pas le « jazz » chez France Inter)
- Les noms d'instruments, différents selon l'ère culturelle, géographique et l'époque (médiéval, âge baroque, classique, musique électronique …)
- Les ayants droit (noms de personnes physiques et morales : labels, producteur, distributeur studio d'enregistrement, etc.)

Enjeux pour la fonction documentaire

Avec la création et la maintenance de référentiels, la fonction documentaire acquiert un nouveau rôle :

- Assurer l'homogénéité des données selon les services, antennes
- Apporter de la granularité de l'information pour offrir une meilleure flexibilité selon la publication selon le type de support (ordinateur, Smartphone, tablette, TV connectée …).
- Equiper les différents services avec des outils de publication et de mise à disposition des données.

L'expertise métier est à souligner : il faut des documentalistes musicologues pour faire ce travail. Par exemple, comment normaliser la description des instruments de musique ou des genres ? En tant qu'institution, Radio France a un rôle à jouer sur la création d'index de référence au niveau international, notamment en ce qui concerne la musique du monde, très peu décrite dans une optique de gestion documentaire. C'est un moyen de revaloriser la fonction documentaire et de faire rayonner l'institution. Pour l'instant, la valeur ajoutée est bien perçue par les utilisateurs internes, mais il faut encore convaincre sur le long terme DRH et Direction Financière.

Perspectives pour l'institution

La culture de Radio France est celle du flux, du broadcast, du temps réel. Jusqu'à récemment, la question de la conservation ne se posait qu'à la marge, car c'est d'abord la mission de l'INA. Cette position change progressivement. Radio France doit s'impliquer dans la conservation et la valorisation du patrimoine musical. La gestion sémantique des données permettra de renforcer le rôle de l'établissement sur le web culturel. Cette action s'articule autour de trois axes :

1. Constituer une base de connaissance riche autour des fonds musicaux accumulés par Radio France depuis 1934 (date de création de l'Orchestre National de France). Ex : mettre à disposition toute la documentation disponible sur une symphonie : retranscriptions des coups d'archets utilisés lors d'une représentation de référence)
2. Participer à la création d'ontologies de référence en musicologie
3. Enrichir les référentiels par alignements avec ceux d'autres établissements

Questions – Echanges avec la salle

Quelle coordination interne et quel calendrier ?

La BBC a ouvert ses données en Linked Open Data pour les réutiliser en les mélangeant à des données externes afin de produire de nouveaux contenus. Le cas de Radio France est différent car il y a de nombreux gisements dormants, et la structure décentralisée de Radio France rend difficile la mise en place d'une politique « top down ». La démarche fonctionne par capillarité et par effets d'opportunité entre les services. Le travail est bien avancé sur la mise en place des référentiels œuvres, personnes physiques et morales. Il faut encore progresser sur la mise en place d'identifiants communs. Ce travail d'harmonisation doit être mis en parallèle avec un autre chantier en cours : la fusion de la base documentaire et du magasin numérique au sein de la discothèque. C'est un champ d'expérimentation important pour tester les référentiels (Œuvres, personnes physiques, morales, identifiants). Fusionner est l'occasion d'harmoniser. Enfin, la Direction de la Documentation et la Direction de la Musique collaborent sur le montage d'une nomenclature de référence pour les concerts des orchestres programmés (description des instruments, des applications : comment une œuvre se monte concrètement ?).

8. Perspectives offertes par les "données liées" à la Cité de la Musique

Marie-Hélène Serra est Directrice du département « Pédagogie et Médiathèque » à la Cité de la Musique. Rodolphe Bailly est Responsable du Système d'Information Documentaire et de la Numérisation à la Cité de la Musique

Les missions de la Cité de la Musique

La Cité de la Musique est un EPIC crée par décret du Ministère de la Culture en 1995. L'établissement a pour mission la promotion du spectacle autour de trois axes :

1. L'appropriation par les publics : programmation de concerts « live » ou différé, en VOD ou vidéos d'archives
2. La médiation, pédagogie auprès des publics : réalisation de produits multimédias complémentaires aux concerts (ex : visualisations, animations pour comprendre une œuvre de l'intérieur)
3. La conservation du patrimoine en spectacle vivant (médiathèque).

En plus des salles de spectacles (Pleyel+Amphithéâtre), la Cité de la Musique comprend :

- Le Musée de la musique : collection permanente de + 7000 pièces, expositions temporaires, un laboratoire de recherche et de restauration (Labex)
- Une médiathèque musicale : + 60 000 ouvrages et partitions, portail de plus de 52 000 ressources numérisées (concerts et œuvres, photographies, captations vidéos, fiches pratiques, dossiers pédagogiques)
- Des espaces d'activités éducatives : ateliers de pratique, cours, master class, conférences, visites commentées du musée …

A noter : en tant qu'EPIC, la Cité de la musique n'a pas obligation légale d'ouvrir ses données. Les EPIC sont aujourd'hui hors champs de la Directive 2003 obligeant les établissements publics à ouvrir leurs données.

Vers un nouveau marketing culturel, numérique et sémantique

Aujourd'hui, l'établissement rassemble un public d'utilisateurs habitués de la création contemporaine. L'enjeu est de s'ouvrir à d'autres cibles sans perdre cette population acquise. En 2012, la Cité de la Musique a initié une réflexion sur la mise en place d'un nouveau marketing culturel, centré sur les déploiements numériques et sur l'enrichissement des contenus. La démarche sémantique de l'établissement s'inscrit dans cette nouvelle politique. Il s'agit de décloisonner les silos et de réutiliser les données pour enrichir les produits multimédias (pédagogiques, documentaires, programmatiques), et d'assurer le continuum entre la fonction documentaire et la fonction marketing. L'internaute doit être guidé tout au long du « tunnel d'achat », en lui proposant des contenus exhaustifs et de qualité. Un établissement culturel doit viser l'appropriation des dispositifs et de l'offre par les publics, et pas simplement leur consommation. Le marketing et la fidélisation passent par une politique forte et qualitative de production de contenus.

La situation du SI : fragmentation des contenus et de l'accès

Comme Radio France, la Cité de la Musique a amorcé l'élaboration d'un modèle de données et doit faire face à la quantité et l'hétérogénéité des contenus à traiter. Trois types de données sont concernés : les données liées à l'offre et à la demande (bases communication, relations presse, relations au public…), les données liées aux différentes bases documentaires (médiathèque, musée), les données liées à l'activité pédagogique (ateliers, cours …). Le SI est organisé dans une logique de silos par métiers. Aucune application n'agrège l'ensemble des données disponibles sur un compositeur, une œuvre ou un événement. Ces données sont fragmentées au sein de plusieurs bases, correspondants à plusieurs services (activités, ressources, publics …). Du point de vue des utilisateurs, il y a une grande diversité dans les modes d'accès et de consultation. Plusieurs portails existent pour des usages très différents : billetterie, médiathèque.cite-musique.fr, CitéLive, education.citedelamusique.fr, media.citedelamusique.fr applis mobiles, …

La libération des données d'abord à l'interne

L'interconnexion des différentes bases de données internes est un premier développement pour préparer les bases d'une Cité de la musique dématérialisée. La « libération des données » se fera d'abord à l'interne, ce qui nécessite d'élaborer des référentiels communs pour créer des ponts entre des bases qui ne se parlent pas et faciliter la réutilisation d'un service à un autre. Par exemple, il n'est pas aujourd'hui possible de croiser les données de la base « relations publics » (statistiques de fréquentation des spectacles) avec les données du service communication. Des passerelles sont à l'étude avec d'autres établissements comme la BNF. L'idée sera de contextualiser l'information sur les œuvres dans un patrimoine culturel plus large.

Vers un data.citedelamusique.fr ?

Un laboratoire expérimental est en cours de montage pour tester les technologies « Linked Open Data » à petite échelle. Un entrepôt RDF maison a été créé pour tester les différents modèles et les formules d'enrichissements adoptées sur des données choisies. L'entrepôt agrège des données de la saison musicale en cours, la description des enregistrements audios et vidéos et des données complémentaires issues de la médiathèque. En parallèle, des passerelles sont réalisées avec DBpédia pour enrichir les informations sur les personnes. A ce stade, l'entrepôt repose sur des relations d'équivalence entre des bases de données cloisonnées. Les alignements ne peuvent être réalisés qu'au niveau des noms de personnes (compositeurs, interprètes, …). Une interface en HTML permet la navigation dans l'entrepôt sur le web. Le système permet d'extraire toutes les informations disponibles au sein des bases de la Cité de la Musique sous un format structuré. Ex : obtenir toutes les descriptions des symphonies de Berlioz, l'ensemble des enregistrements disponibles, les différentes interprétations et la date de la prochaine représentation.

Prochains développements

L'extension du périmètre des données concernées et l'établissement de relations d'équivalence au niveau des œuvres sont à prévoir. Un autre développement à moyen terme consistera à interfacer l'entrepôt au site web de la Cité de la Musique afin de lier les contenus sur le programme en cours à la billetterie et de mesurer le ROI. L'interconnexion des ressources avec data.BnF.fr pour contextualiser les ressources est à l'étude. La Cité de la Musique participe également au projet ANR de modélisation du champ musical DOREMUS, dont les résultats alimenteront le développement des référentiels.

Questions - Echanges avec la salle

Quel(s) soutien(s) des différentes directions ?

Pour l'instant, le développement de l'entrepôt n'est pas prioritaire. Toute la difficulté est de faire comprendre aux donneurs d'ordre l'impact positif d'un projet d'infrastructure sémantique. Il y a des ROI à tous les niveaux : visibilité – marketing de l'offre, mise en cohérence et efficience du SI, enrichissements des contenus, etc. Pour une institution culturelle, il est très difficile d'obtenir une vue d'ensemble de ses fonds. Un projet de cette nature permet d'y arriver sans remettre à plat l'existant. Aujourd'hui, Il faut apporter une « couche éditoriale » au projet pour convaincre les directions culturelles d'en faire des investissements prioritaires. Bien souvent, c'est cette couche qui suscite l'adhésion.

Quel(s) arbitrage(s) entre la simplification et la préservation de la qualité des données ?

Il faut trouver un équilibre entre la tentation de simplifier les informations et celle de les sur-enrichir. C'est un arbitrage complexe, surtout dans le secteur culturel où la documentation doit encore souvent prouver sa légitimité par rapport à la communication. Un des points essentiels est la création de tables de correspondances entre les fonctions. Il est important que chacun continue à parler dans sa langue (ex : les genres musicaux de communication ne sont pas ceux de la documentation). Si la documentation force les autres services à utiliser ses vocabulaires, la démarche sera perçue comme prescriptive avec un risque d'échec du projet. Par la multiplicité des alignements possibles, les Linked Data sont un bon instrument pour faire se parler les services, sans attenter à la langue et aux outils de chacun.

Quelle(s) méthode(s) de liaison(s) aux sources externes ?

Pour l'instant, l'entrepôt est seulement connecté à DBpédia car il s'agit de s'approprier les technologies du Linked Data sur un échantillon. Ensuite, il faudra développer des partenariats avec d'autres institutions de références, notamment la BNF. La Cité de la Musique pourra aussi s'appuyer sur les acquis du projet MIMO qu'elle a coordonné jusqu'en juillet 2011. MIMO (Musical Instruments Museums Online) est un agrégateur intégré à Europeana rassemblant plus de 43 234 notices descriptives sur des instruments de musique issus des différents muséums musicaux[28]. La Cité de la Musique a notamment participé au développement du thésaurus implémenté dans l'agrégateur, qui décrit les 6 grandes familles d'instruments sur 3 niveaux. La période d'austérité budgétaire est propice aux collaborations. Les institutions culturelles ont tout intérêt à rentrer dans des démarches de mutualisation.

Quelle position vis-à-vis de l'Open Data ?

La politique de l'établissement est de faciliter au maximum la diffusion de ce qui peut l'être, conformément à sa mission de promotion du spectacle vivant. L'évolution de la Directive 2003 est surveillée, tout comme l'action d'Europeana, afin d'apporter des éclaircissements sur l'exception culturelle et le statut des EPICS. Mais toutes les données de la Cité de la Musique ne peuvent pas être ouvertes. Le périmètre est celui de la création musicale contemporaine. Beaucoup de ressources sont encore protégées par le droit d'auteur. Sur les captations vidéos et les photographies professionnelles, le recours à des sociétés de gestion collectives est obligatoire. Il y a aussi des problématiques de droit à l'image. On est parfois loin du domaine public, l'ouverture ne pourra être que partielle et raisonnée.

9. Enjeux culturels et linguistiques autour des données liées : le projet Semanticpédia

Thibault Grouas est Chef de la mission des langues et du numérique, Délégation Générale à la Langue Française et aux Langues de France.

Sa présentation est disponible en ligne sur le site de l'INRIA :

http://wimmics.inria.fr/projects/dbpedia/fichiers/presentations/Presentation_GFII-2013.pdf

Pour de plus amples informations sur les liens entre Wikipédia et DBpédia, voir la présentation de Fabien Gandon, Senior Research, Wimmics Team Leader, INRIA AC Representative at W3C disponible à l'adresse suivante :

http://upload.wikimedia.org/wikipedia/commons/7/73/Lancement_Semanticpedia_-_Discours_Fabien_Gandon_avec_commentaires.pdf?uselang=fr

Missions de la DGLFLF

Rattachée au Ministère de la Culture, la Délégation générale à la langue française est devenue Délégation générale à la langue française et aux langues de France en 2001. La DGLFLF est chargée de

[28] http://pro.europeana.eu/mimo-edm

coordonner la politique linguistique de l'État au niveau interministériel et national. La DGLFLF participe à l'enrichissement de la langue française à travers les différentes instances de contrôle et d'évolution de la langue (listes terminologiques de l'Académie française, Commission générale de terminologie et de néologie, dictionnaire terminologique FranceTerme). La préservation de la diversité du patrimoine linguistique est au centre de ses préoccupations (« les langues de France »). Il faut distinguer en métropole l'ensemble des langues régionales territoriales (alsacien, basque, breton, catalan, corse …) des langues non territoriales (arabe dialectal, arménien occidental, berbère …). Il y a ensuite l'ensemble des langues parlées sur les territoires français d'Outre-Mer (on compte 28 langues canaques en Nouvelle-Calédonie), et la langue française des signes[29]. La DGLFLF s'est équipée d'une mission des langues et du numérique, reconnaissant que les outils numériques sont devenus un vecteur essentiel à la promotion, l'étude et l'enregistrement de la diversité linguistique nationale. Cette mission coordonne des projets ayant trait au traitement informatique des langues, avec une forte composante traitement automatique du langage (TAL) et sémantique.

Importance historique du TAL en France

La France est très active en TAL. C'est un des pays européens où on trouve le plus de laboratoires sur le sujet. Les liens avec l'industrie sont nombreux et il y a un tissu de PME-PMI très innovantes, notamment en traduction automatique[30]. Du fait de sa transversalité, le TAL est un champ stratégique en R&D car il y a des applications potentielles dans tous les secteurs. De nombreuses technologies utilisées par des grands groupes sont nées dans les laboratoires français.

Semanticpédia : nouveau levier de l'influence linguistique française

Le développement du Web sémantique déterminera à l'avenir les logiques de publication et d'accès aux contenus. La France doit se positionner dans cette transformation et intervenir dans les instances de gouvernance du web de données sous peine d'être écartée du web de demain. Il faut en complément une politique culturelle et linguistique forte, produire et exporter des contenus francophones compatibles avec ces nouveaux standards. L'initiative Semanticpédia traduit une prise de conscience de l'État français de ces nouveaux enjeux. Semanticpédia est une convention signée en novembre 2012 entre le ministère de la Culture et de la Communication (représenté au niveau du secrétariat général par le département des programmes numériques et par la DGLFLF), l'INRIA et Wikimédia France. Cette initiative vise à stimuler l'innovation de services, l'innovation technologique, la production de contenus originaux et l'émergence de nouveaux usages à partir de la réutilisation des données francophones issues de Wikipédia[31]. Les compétences de chaque signataire sont complémentaires : Wikimédia apporte son réservoir de données et son expertise sur l'économie de la contribution, l'INRIA apporte la couche technologique et le ministère de la Culture et de la Communication facilite les synergies institutionnelles et octroie les fonds.

[29] Voir le site de la DGLFLG pour plus d'informations : http://www.dglf.culture.gouv.fr/

[30] Voir par exemple les retours d'expériences proposés lors de la journée d'étude du GFII sur le sujet en 2009 : *« Le multilinguisme : des solutions existent pour réduire les barrières des langues : retours d'expériences d'EADS/MBDA, de l'INA et de TOTAL »* : http://www.gfii.fr/fr/document/le-multilinguisme-des-solutions-existent-pour-reduire-les-barrieres-des-langues-br-i-retours-d-experiences-d-eads-mbda-de-l-ina-et-de-total-i

[31] http://www.inria.fr/actualite/actualites-inria/partenariat-semanticpedia

Synthèse de la journée d'étude du GFII : *« Données culturelles et Linked Open Data : valoriser le patrimoine public dans le web des données »*, 26 mars 2013, Maison de l'Europe.

Wikipédia et le domaine culturel

Wikipédia est devenu la base de connaissances la plus exhaustive pour le domaine culturel francophone. On trouve plus de 600 000 pages de contenus français liées au domaine culturel. Tous les corps de métiers et toutes les disciplines sont représentées (peinture, architecture, archéologie, histoire …). C'est une spécificité nationale : le Wikipédia anglophone est nettement plus orienté « sciences et média ». Wikipédia est aussi la base de connaissances la plus facilement réutilisable d'un point de vue juridique et technique. Les contenus sont considérés comme un bien commun et placés sous licence Creative Commons. Une partie de l'information présentée sur chaque page est semi-structurée (listes, infobox, vignettes ….), et peut-être convertie en format structuré tel qu'RDF, ce qui la rend lisible par les machines.

Semanticpédia : un DBpédia.fr

Toutes ces données sont agrégées dans les standards des données liées, ou *« linked open data »* (LOD) au sein de l'entrepôt DBpédia. Crée dans le sillage de l'initiative LOD en 2006, DBpédia a apporté une nouvelle dynamique au développement du web de données : un espace de connaissances multilingues, interconnectées, interrogeables, partageables, réutilisables et mises à jour en temps réel. Le projet est aujourd'hui le référentiel central du web de données[32]. A l'origine, DBpédia était centré sur les contenus anglophones, mais des versions locales se sont développées, DBpédia en français, projet développé dans le cadre du partenariat Semanticpédia[33]. Cela représente un énorme gisement de données (noms de personnes, dates, monuments, bâtiments …) pour produire des contenus et des services francophones innovants (réalité augmenté, géolocalisation …).

Les enjeux linguistiques : s'affranchir de la traduction pour accéder aux contenus

La DGLFLF s'intéresse particulièrement aux possibilités qu'offre DBpédia pour la création de nouveaux modes d'accès linguistiques aux contenus (interfaces à facettes multilingues). Les infrastructures sémantiques permettent de s'affranchir de la traduction pour exporter ou importer des ressources car les systèmes n'analysent plus leur contenu mais leur description, qui est normalisée dans le langage de la machine. Dans le domaine de l'Histoire des Arts, l'initiative HDA-Lab a montré que des contenus francophones sur l'architecture gothique deviennent identifiables par un chercheur japonais, et qu'inversement, des contenus japonais sur le même sujet sont identifiables par un Français.

Les enjeux linguistiques : exporter et relocaliser les langues de France

Ces possibilités renouvellent les modes d'influence de la francophonie : l'approche n'est plus défensive, il s'agit de décloisonner et faire dialoguer des patrimoines linguistiques auparavant hermétiques. Le mouvement est double : exporter des contenus francophones et faciliter leur signalement dans le reste du monde mais aussi identifier des contenus en langue étrangère depuis la

[32] Pour plus d'informations sur l'importance de DBpédia dans le développement du web de données, voir la présentation d'Alexandre Monnin, IKKM, au Forum du GFII 2013 : *« DBpédia et les Linked Open Data ou la question du public »*. http://forum.gfii.fr/uploads/docs/Alexandre_Monnin_DBPedia_ForumduGFII.pdf

[33] Selon le site de l'initiative, le DBpédia anglophone décrit actuellement 3.77 millions d'objets, dont 764 000 personnes, 573 000 endroits, 333 000 œuvres, 192 000 organisations, 202 000 espèces et 5500 maladies. Au total, l'ensemble des bases DBpédia (base anglophone + 111 versions locales) s'élève à 20.8 millions d'objets. http://dbpedia.org/About

langue française. La traduction reste nécessaire pour l'appropriation des contenus mais pas pour leur signalement. C'est une révolution car la connaissance dans le domaine culturel était jusqu'à présent bornée par les frontières linguistiques. L'anglais n'a pas acquis, dans le champ culturel, le rôle de vulgate qu'il joue dans le champ de la connaissance scientifique où il est la langue internationale. Le désenclavement culturel et linguistique des données sera l'enjeu le plus important du web de données.

Les enjeux linguistiques : documenter le patrimoine oral

Un autre aspect de l'initiative Semanticpédia qui intéresse la DGLFLF est la préservation des langues orales, pour lesquelles il y a peu de supports écrits. Dans les sociétés orales, lorsque les mots disparaissent, les coutumes et les savoir-faire disparaissent aussi. Il faut créer des ressources spécifiques, des dictionnaires qui alimenteront la base. C'est un sujet très politique. Les directions régionales des affaires culturelles (DRAC) sont particulièrement intéressées par cette initiative car cela dessine des possibilités pour relocaliser l'accès et la contextualisation du patrimoine dans les régions (ex : proposer un accès aux collections des musées en français et dans la langue régionale).

Les enjeux linguistiques : alimenter le Wiktionnaire francophone

La DGLFLF soutient aussi l'alimentation du Wiktionnaire francophone : 2 millions de termes viendront s'ajouter au réseau sémantique des articles de Wikipédia. La volonté de sémantiser ce dictionnaire est forte, mais il faut trouver un consensus sur l'approche : syntaxique VS sémantique ? spécialisation VS ambition encyclopédique ? Le périmètre du Wiktionnaire est large : contenus belges, suisses, québécois, africains … Il faut mettre tout le monde d'accord sur le sens des mots. Une première version RDF est en cours de développement. Elle servira à stimuler l'innovation dans la traduction automatique.

Questions / échanges

Quels exemples de réutilisation ?

À un premier niveau, Semanticpédia doit faciliter la réutilisation par les institutions culturelles. L'échange de données entre musées choisissant d'ouvrir leurs fiches sur DBpédia a une vraie valeur ajoutée. Les institutions peuvent coopérer pour enrichir mutuellement leurs fiches là où il y a des manques. Cela permet aux établissements d'être identifié comme une source de référence sur une œuvre, une collection, un artiste. À un second niveau, Semanticpédia doit stimuler la R&D à partir de la réutilisation des données. Dans le cadre de Semanticpédia, le MCC a débloqué des fonds pour financer des initiatives internes de sémantisation. De nombreux projets sont en cours d'expérimentation, comme HDA-Lab ou la sémantisation de la base JOCONDE. JOCONDE est le catalogue collectif des collections des Musées de France, accessible au public sur le web. C'est la source de référence la plus exhaustive sur les illustrations, les œuvres picturales et graphiques détenues par les musées français. La base comprend 500 000 notices descriptives qui sont en cours de sémantisation et d'interconnexion avec DBpédia en français. JOCONDE est l'une des briques principale de l'agrégateur national COLLECTIONS, qui moissonne plus de 40 archives de musées nationaux et régionaux pour les reverser dans Europeana. Aujourd'hui, s'interconnecter à JOCONDE est le moyen le plus simple pour un musée d'être visible dans COLLECTIONS et donc sur Europeana. Des projets de R&D très pointus sont aussi à l'étude dans des domaines émergents, notamment la

navigation au sein de corpus sonores. Aujourd'hui, le sémantique fonctionne très bien sur les textes et progresse sur l'image, mais le son reste un champ à explorer.

10.HDA-Lab : un regard prospectif sur le tagging sémantique

Bertrand Sajus est Chef de projets au département des programmes numériques (DPN) du Ministère de la Culture et de la Communication (MCC).

HDA-Lab est un exemple de réutilisation expérimentale des données du Corpus « Histoire des Arts » constitué par le département des programmes numériques du MCC, réalisé en collaboration avec l'Institut de Recherche et d'Innovation (IRI) Cette collaboration a contribué à la mise en place de la Convention SemanticPédia.

Le portail « Histoire des Arts » : faciliter la production de supports d'enseignement.
Le portail « Histoire des Arts[34] » donne accès à plus de 5000 ressources éducatives alimentées par 350 institutions culturelles, réparties sur l'ensemble du territoire. Le projet, coordonné par le DPN, vise à faciliter la constitution de supports éducatifs par les enseignants et les médiateurs culturels. L'Histoire de l'Art est une discipline obligatoire dans l'enseignement secondaire général depuis 2010. En 2010, l'édition en histoire de l'art ne proposait que très peu d'ouvrages destinés au programme scolaire. Il fallait donner accès aux communautés enseignantes à des ressources fiables et peu onéreuses. Le corpus a été structuré en fonction du programme pédagogique. C'est une structuration simple, adaptée aux usages des enseignants (structuration verticale, par grandes catégories : « primaire », « secondaire », …) Les notices descriptives de chaque ressource renvoient, par des liens profonds, vers autant de ressources en ligne, chaque notice contenant des mots-clés, initialement saisis comme de simples tags.

HDA-Lab : explorer les possibilités du tagging sémantique
La collaboration du DPN avec l'IRI a pour périmètre le projet HDA-Lab qui repose sur deux modules, reprenant les données du portail « Histoire des Arts » : un module de tagging sémantique des ressources pour leur indexation (module « HDA-BO »), et une interface de recherche et de navigation dans la version sémantisée du corpus (site « HDA-Lab »). Il s'agit d'un projet de recherche, une « preuve de concept » qui repose sur l'expérimentation et la recherche des nouvelles possibilités offertes par le tagging sémantique en recherche d'information. Pour l'instant, HDA-Lab est une annexe du portail « Histoire des Arts » : la sémantisation du corpus est partielle et les fonctionnalités proposées ne sont pas définitives. A terme, certaines fonctionnalités expérimentées dans HDA-Lab seront intégrées au portail « Histoire des Arts ». Mais il faut encore pousser plus avant la sémantisation du corpus et analyser les retours des utilisateurs (indexeurs et utilisateurs finaux).

HDA BO : indexer

[34] Accessible à l'adresse suivante : http://www.histoiredesarts.culture.fr/

Le module HDA-BO (Back Office) permet d'apparier les tags librement renseignés par les indexeurs avec les entrées des articles de Wikipédia. Cette opération montre qu'il est important, pour les institutions culturelles francophones, de disposer de la totalité du corpus DBpédia RDF en français car de très nombreuses entrées n'existent pas dans la version anglaise (ex : « Yvette Horner », « Bernard Lavillier », …). Les tags sémantiques sont constitués des entrées de Wikipédia et accompagnés de métadonnées (liens vers l'article Wikipédia, permaliens, URIs DBpédia). L'extraction a permis de rapatrier 17 000 tags dans le back office. Ces tags sont proposés aux utilisateurs pour l'indexation par auto-complétion. L'objectif était de voir si les utilisateurs (indexeurs) supporteraient la contrainte d'un semi-contrôle sur leur vocabulaire. Les retours montrent que la suggestion de termes par auto-complétion est perçue comme une aide.

HDA-Lab : explorer

Le second axe du projet était la création d'une interface de navigation facilitant l'exploration « multidimensionnelle » des ressources. L'utilisation de tags sémantiques, normalisés par un référentiel, permet de créer des facettes à partir desquelles de nouveaux usages de recherche sont possibles. Le tagging sémantique permet aussi de désambiguïser les termes (ex : « roman »). A l'origine, les ressources agrégées dans le portail « Histoire des Arts » n'avaient pas de référentiel propre en raison de l'amplitude des champs couverts (de la préhistoire à la période contemporaine, peinture, sculpture, musique, jardins, littérature, BD, …). L'utilisation d'un référentiel très large (1 300 000 entrées), constitué par les usages, et d'ambition universelle comme Wikipédia / DBpédia, a permis de sémantiser environ 80% des tags.

Valeur d'usage pour la recherche d'information

De nombreuses facettes sont disponibles : recherche par zone géographique (pays), par ligne de temps (année, décennie, siècle), par disciplines, par courants – genres artistiques, par mots-clés rattachés aux œuvres (nuage de tags)… Le moteur est dynamique : ces facettes peuvent être re-combinées dans le temps de la requête pour affiner l'exploration. Les facettes sont à la fois des critères de recherche et des modes de restitution visuelle de l'information. La navigation par concept est également proposée, via les arbres thématiques de Wikipédia et via des extraits de thésaurus utilisés par le MCC[35]. Les métadonnées sont enrichies par les « infobox » extraites des articles de Wikipédia, disponibles en format RDF, grâce à DBpédia. Ce dispositif propose également des liens vers des ressources connexes (recherche sur l'auteur d'une œuvre, sur une date, sur un musée …). Un module de géolocalisation permet de localiser l'institution qui détient une ressource (musée, archive, …). L'interface permet de pousser loin la contextualisation de la recherche sans demander à l'utilisateur de savoir écrire des requêtes expertes.

Alignements et interconnexions linguistiques

L'utilisation du RDF permet de procéder à des équivalences linguistiques sans passer par la traduction. Aujourd'hui des alignements sont possibles vers 5 langues : anglais, italien, espagnol, allemand, japonais. Un Japonais ne connaissant pas le terme « gothique international » en français

[35] Encore en développement, la fonctionnalité est pour l'instant disponible en démonstration sur les branches « secteur urbain » du thésaurus de l'architecture et de l'urbanisme et « Architecture d'habitation » du thésaurus iconographique de Garnier.

pourra l'entrer dans sa langue, les ressources francophones lui seront proposées par simple liaison avec le DBpédia francophone. HDA-Lab montre que le web de données permet de créer de nouveaux accès interculturels et translinguistiques pour des investissements éditoriaux minimes, sans qu'il y ait à modifier le système d'indexation. DBpédia sert de « hub sémantique » entre les langues.

Questions - échanges avec la salle

Quelle déperdition dans les alignements linguistiques ? Comment le MCC se positionne-t-il par rapport aux œuvres qui restent « enclavées » ?

Il faudrait plutôt poser la question à l'envers. Si les Japonais accèdent ne serait-ce qu'à 50% des ressources francophones disponibles sur le patrimoine artistique français (titres d'œuvres, noms d'artistes, courants artistiques, …), c'est 50% qui n'existaient pas avant pour eux. Pour accroître le ratio, il faut encourager les synergies entre institutions. Celles-ci doivent entrer dans le cercle vertueux de l'enrichissement des référentiels. Les musées ont tout intérêt à contribuer à l'enrichissement de Wikipédia dans leurs domaines d'expertise et à s'interconnecter à d'autres institutions via des Hubs sémantiques. Il est à noter qu'Aurélie Filipetti e encouragé cette démarche lors de la signature de la Convention Semanticpédia.

Quel(s) retour(s) de l'Education Nationale ?

Des réticences étaient attendues en raison de l'image parfois encore négative de Wikipédia dans les communautés enseignantes. Mais celle-ci s'améliore d'année en année. D'une part, Wikipédia a mis en place un code de bonnes pratiques suffisamment efficace pour garantir la fiabilité des données. D'autre part, les enseignants comprennent bien la complémentarité entre Wikipédia et des sources plus classiques. De plus en plus d'institutions alimentent Wikipédia et rédigent des contenus. Il y a également des initiatives dans des lycées : les élèves rédigent et / ou réutilisent des articles de Wikipédia dans le cadre de projets scolaires. Wikipédia est de facto la source la plus utilisée par les jeunes pour s'informer sur un sujet. La question n'est donc pas de la rejeter mais bien d'étudier les manières pertinentes d'exploiter cet outil.

11. Comment faire parler le Web des données?

Hicham Tahiri est président de Vocal Apps

Contexte de création de Vocal Apps

La création d'un DBpédia francophone a redynamisé l'écosystème du web sémantique, mais il y a encore assez peu d'exemples de réutilisateurs commerciaux des données « libérées » avec Semanticpédia. Vocal Apps, crée en juillet 2012 et incubée à Paris Incubateurs, est une des premières start-up à se positionner sur ce gisement. A l'origine de Vocal Apps, il y a la volonté de développer des systèmes capables de « faire parler les machines ». La société a conçu un prototype de robot

parlant reposant sur 23 moteurs avec le CRIIF : SAMI.[36] Le robot peut dialoguer avec les visiteurs et répondre à des questions factuelles en allant chercher les réponses dans une base de connaissance alimentée par DBpédia.

Pourquoi faire parler le web ?

La langue orale est notre « interface naturelle ». Aujourd'hui, pour trouver des réponses à des questions factuelles (ex : *« Où se situe le Louvre ?»*), les usagers doivent requêter par écrit dans des moteurs. Les résultats doivent être filtrés dans une liste à plat, ce qui représente une perte de temps. Il y a un vide à combler sur ces usages ciblés. Croiser les technologies de reconnaissance vocale à l'Intelligence Artificielle (IA) et au sémantique permet de concevoir des Systèmes de Questions-Réponses (SQR) automatiques, renvoyant des réponses concises, autonomes des documents qui les contiennent (pages web). Le traitement avancé de la parole a atteint un stade de maturité satisfaisant, mais il faut encore progresser sur l'injection de connaissances dans l'IA des systèmes. Avec les Google Glass, Google expérimente la réalité augmentée, mais la surcharge cognitive et visuelle imposée aux utilisateurs pourrait se révéler un frein à l'adoption. Il vaut peut-être mieux marcher dans la rue avec des écouteurs renvoyant une réponse auditive synthétique, qu'avec une paire de lunettes surimposant une couche d'informations à l'environnement …

De l'importance de DBpédia pour les technologies de la parole

DBpédia offre une immense base de connaissance *« machine readable »* pour développer des SQR automatiques francophones. C'est la matière qui manquait jusqu'à présent pour injecter des connaissances de nature « encyclopédique » dans les systèmes. DBpédia consigne sous forme structurée « le registre des connaissances de la multitude à un instant T » dans un état de langue en phase avec le parler contemporain, ce qui ce qui facilite leur implémentation dans des SQR.

Le démonstrateur IZIPEDIA

En plus de SAMI, Vocal Apps a développé un autre démonstrateur : IZIPEDIA. Il s'agit d'un SQR permettant de répondre à des questions de culture générale (géographie, culture, personnalités, politique, entreprises …) en allant extraire en temps réel des contenus structurés dans DBpédia. Le système construit des réponses en langage naturel à des questions exprimées en langage naturel (ex : *« Quel est le débit du Mississipi ? » => « Le débit volumique du Mississippi est de 18000 m3/s »)*. La réponse est enrichie d'une vignette permettant de contextualiser l'information à partir des « infobox » extraites de Wikipédia. La brique de reconnaissance vocale n'est pas encore implémentée sur toutes les versions. http://izipedia.com/

Quels usages ? Quelle valeur ajoutée ?

Dans une optique professionnelle, on peut imaginer des services d'informations optimisés pour le *fact-checking* au sein de gros corpus de données (Big Data), pour le rapprochement et la comparaison d'entités, d'ordres de grandeur (ex : comparer des populations entre des pays, des villes ...). En termes de hiérarchisation de l'information, la valeur ajoutée est importante : l'utilisateur distingue clairement les données factuelles, objectives, et les informations de contexte, ce qui offre

[36] SAMY a été présenté pour la première fois au public à l'occasion du festival « Futur en Seine » en juin 2013: http://www.atelier.net/blog/2013/06/18/joujoux-de-technologie-exposent-futur-seine_421463
Synthèse de la journée d'étude du GFII : *« Données culturelles et Linked Open Data : valoriser le patrimoine public dans le web des données »*, 26 mars 2013, Maison de l'Europe.

aussi des perspectives pour la production de nouveaux supports d'information. Des synergies avec des sources plus « professionnelles » sont aussi possibles (bases de données scientifiques, juridiques, …).

Le mariage du web sémantique et de l'IA

Le mariage entre les technologies sémantiques et celles de l'IA paraît naturel. Les systèmes doivent disposer de contenus structurés et décrits dans les formalismes du web sémantiques pour « accéder au sens » et développer, via des algorithmes propres à l'IA, des capacités d'auto-apprentissage. Mais la mise en œuvre pose de nombreux défis en termes de modélisation et d'interconnexion. Avec IZIPEDIA, Vocal Apps a produit un travail important sur l'architecture des données pour faciliter leur mise en relation. L'enrichissement et les interconnexions entre les données comptent plus que leur exhaustivité pour permettre au système d'établir des rapprochements, des inférences, entre la requête et les connaissances enregistrées dans la base. Les « infobox » et la liaison des données en RDF dans DBpédia facilitent considérablement le travail. La désambiguïsation des termes de la requête est une autre étape importante (ex : Saint-Louis : la ville ? l'île ? le Roi ?). Vocal Apps a développé ses propres algorithmes pour interpréter les termes de la requête au niveau du concept, et les rapprocher de ceux présents dans DBpédia. Il faut aussi concevoir un modèle de réponse suffisamment contraignant pour fournir des réponses précises et suffisamment souple pour construire des réponses dans le langage des humains.

Questions – Echanges avec la salle

Quelles perspectives pour Vocal Apps et pour le marché français des SQR automatiques ?

Il y a de très nombreuses applications possibles, dans le domaine de la domotique, de la voiture connectée, des applications smartphones, ou sur le marché de la réalité augmentée. Les usages d'un SQR sont nombreux. Vocal Apps ambitionne de devenir le Wolfram Alpha ou le SIRI français. Wolfram est très actif aux Etats-Unis mais a peu pénétré le marché européen, et SIRI, en France, ne permet de répondre à l'ensemble des questions des utilisateurs car il n'est pas adossé à une base de connaissance suffisamment exhaustive. Vocal-Apps cherche aujourd'hui à nouer des partenariats avec des acteurs de la recherche d'information (moteurs, portails) ou des éditeurs d'assistants personnels pour apporter son expertise technologique. Le monde du SEARCH est trusté par la Silicon Valley. Le SQR automatique est un des rares domaines où il y a une carte à jouer pour les acteurs français.

Quels axes d'amélioration ? Quelles interrogations ?

La principale difficulté réside dans la nature même de DBpédia qui est une « source rebelle ». Les infobox ne représentent que 5% des pages. Les contenus les plus riches sont ailleurs. Il faudra un jour parvenir à développer des systèmes capables de traiter les contenus non-structurés et semi-structurés pour exploiter en profondeur la mine d'information de DBpédia. Une autre interrogation consiste à savoir comment intégrer les usagers dans la boucle technologique, et exploiter l'ingénierie sociale. Aujourd'hui, les usagers interviennent pour sourcer et documenter les pages de Wikipédia. On peut imaginer des systèmes permettant d'intégrer leur retour sur les résultats. De manière plus immédiate, il faut améliorer le mapping de DBpédia et les mises à jour. Aujourd'hui, la base est actualisée toute les 6 semaines à partir des DUMP de DBpédia. Vocal Apps souhaiterait disposer

d'un endpoint « live » pour faciliter l'actualisation en temps réel des connaissances. Il faudra aussi suivre l'éventuelle concurrence de Wikidata sur DBpédia et l'intégrer à la base si nécessaire.