Convergence et interopérabilité : l'apport du Web de données

(Convergence and Interoperability: a Linked Data perspective)

Emmanuelle Bermes
Bibliothèque nationale de France (BnF)
Paris, France

Meeting: **149 —B*ridging domains, communities and systems* — Classification and Indexing Section**

Abstract:

The web provides a seamless environment where the user can navigate through resources regardless of their provenance: search engines, public and governmental institutions, commercial websites, social networks, etc. There is a strong contrast between this reality of a seamless web, and the cultural heritage approach of data dissemination. Metadata modelling based on domain-model requirements has led to incompatible standards that make it really difficult to share data between libraries, museums and archives. Cultural heritage institutions maintain catalogues that exist like silos, isolated from one another, and isolated from the wider ecosystem of the web.

In the cultural heritage domain, providing relevant services for Web users requires to seek convergence, and to bridge the gap between libraries, museums, archives and other cultural institutions. Despite the efforts that have been conducted during the past several years, data interoperability is still an open issue. Federated search is permitted by protocols like Z39.50 and Web services & APIs, but the kind of service that can be provided by this technology doesn't seem to fit the needs of Web users. Other interoperability strategies, like exchanging simple Dublin Core records through OAI-PMH repositories, have shown their limits: by seeking the smallest common denominator between disparate data, they create a low-quality data environment with limited perspectives.

The Linked Data provides a perspective for a different kind of interoperability, based on Web architecture principles. Linked Data interoperability is designed to support heterogeneous description models, which is necessary to handle the very different data from libraries, museums and archives. The Linked Data cloud is built following a bottom-up approach, allowing each institution visibility and ownership on their own data.

Yet, in order to create this new interoperability framework for linked data, we need to build strong links between datasets. These links rely on the description of "real things": persons, objects, concepts, places, things that have been described as "authority data" or "subject data" by libraries. The new standards for future catalogues, FRBR, FRAD, RDA, tend to dramatically increase the importance of authority data in the library data landscape. Beyond libraries, these data will provide the hub where cross-domain links can be anchored.

There are still a number of open issues, regarding the creation of a semantic network of data that would be able to connect information from libraries, museums and archives, in a seamless way. Some of these issues are related with the need to be able to convert existing metadata into semantic web enabled data. Some other issues are specific to the Linked Data environment: how to create alignments between datasets, what kind of properties should be used to create those links, etc.

This paper aims at exposing the challenge of interoperability and convergence, and the perspectives offered by Linked Data to address this issue. It will be illustrated by implementation examples such as VIAF, Rameau & LCSH, Europeana, etc. We want to propose a way to renew our perspective on interoperability, based on a user-oriented approach, and an emphasis on the need for cross-domain links for subject data. Our goal is to demonstrate that services built on Linked Data will be more efficient and relevant to the end user than services built on traditional library interoperability frameworks.

Ce qui a fait la valeur ajoutée du Web, ce qui a motivé son adoption quasi universelle et est en train d'en faire le principal média de publication et d'échange d'information, c'est sa globalité et son interopérabilité. Le Web, c'est avant tout un ensemble de standards, qui permettent la dissémination de technologies partagées par tous, et indépendantes des environnements matériels et logiciels. Pour aller plus loin, on pourrait dire qu'aujourd'hui le Web est l'environnement le plus interopérable qui soit.

Le principe de la navigation hypertexte et la généralisation de l'usage des moteurs de recherche a provoqué un changement de paradigme qui devrait encourager les institutions culturelles, et au premier chef les bibliothèques, à prendre plus hauteur dans la démarche orientée utilisateur. Il ne s'agit plus seulement de faciliter l'expérience de l'usager une fois qu'il a pu accéder au service proposé en ligne, mais de considérer son objectif de manière plus globale. Est-il dans une démarche d'apprentissage ? De recherche ? De loisir ? Cherche-t-il une réponse pratique à une question sur sa santé, son emploi, sa maison, son quotidien d'une façon générale ? Tous ces usages existent en bibliothèque, mais pour aucun d'eux la bibliothèque ne devrait plus se considérer comme un passage obligé. C'est maintenant à elle de se positionner sur le parcours de l'usager dans sa démarche quelle qu'elle soit, et non à l'usager de penser que la bibliothèque pourrait avoir des ressources pertinentes à lui offrir.

Cette réflexion doit nous encourager à considérer la convergence entre institutions culturelles comme quelque chose de vital, car on ne pourra pas continuer à attendre de l'usager qu'il comprenne les barrières institutionnelles et les accepte. Le touriste qui prépare sa visite au musée devrait pouvoir trouver aussi bien des livres sur Picasso que les reproductions de ses œuvres ; le généalogiste qui trace l'histoire de ses arrière-grands-parents devrait pouvoir accéder aussi bien aux ressources des bibliothèques qu'à celles des archives.

Pour aller encore plus loin, on peut souhaiter que dans l'écosystème actuel du Web, les ressources que l'usager n'a pas cherchées soient poussées vers lui naturellement, au cours de sa recherche, à travers des résultats fournis par son moteur de recherche favori, des liens depuis une page Wikipédia, des références entre sites Web.

Si tout ceci semble du domaine de l'évidence quand on parle des sites Web, pourquoi n'en est-il pas de même pour les données quoi sont cachées dans nos catalogues ? En effet, l'hypertexte et l'interconnexion des pages Web fonctionne de manière optimale pour les ressources de nature documentaire, mais pour aller plus loin, ce sont les données elles-mêmes qu'il faudrait sémantiser et relier pour les rendre interopérables. C'est justement le principe du Web sémantique et du Web de données.

Pour illustrer ceci par un exemple concret, une bibliothèque dispose en général d'un site Web qui est accessible et relié, via un certain nombre de liens hypertextes, à l'ensemble plus global du Web. Cependant, les données de la bibliothèque comme le catalogue, font généralement partie du Web dit profond, ou caché : c'est-à-dire que ces données sont stockées dans une base de données, accessible à travers un formulaire de recherche, et donc ne peuvent pas être parcourues de lien en lien notamment par des agents logiciels tels que les moissonneurs (*crawlers*) des moteurs de recherche. Ainsi, un usager qui souhaite prendre connaissance de ces données doit obligatoirement accéder à ce formulaire et saisir une recherche. Si les ressources qui l'intéressent sont disséminées dans les bases de plusieurs bibliothèques, il devra recommencer autant de fois cette opération.

Bien entendu, les bibliothèques et les institutions culturelles ont pris conscience depuis plusieurs années de cette problématique, et elles ont mis en place différents moyens pour permettre d'y pallier.

1. Interopérabilité et données culturelles : une situation complexe

Dans les bibliothèques, les modèles d'échange de données et d'interopérabilité se sont particulièrement construits sur le fait que les objets décrits sont des objets multiples. Il existait un enjeu important à éviter la duplication de l'effort de catalogage, et donc à favoriser la récupération des notices d'un catalogue à un autre. Cette forme d'interopérabilité s'est matérialisée le plus souvent par l'adoption d'un format commun, ou la construction de passerelles complexes permettant de convertir complètement un format vers un autre (Marc21 vers Unimarc par exemple.)

Le protocole Z 39.50, développé dès les années 1980, suivi par une nouvelle génération (SRU/SRW) reposant sur des standards appartenant davantage à l'ère du Web, en particulier XML, permet l'interrogation synchrone de plusieurs bases. Toutefois, cette façon de faire présente plusieurs inconvénients : avant tout, le protocole reste très spécifique au monde des bibliothèques, et ne permet pas d'interagir avec des ressources d'autres domaines. D'autre part, l'interrogation synchrone de bases différentes présente des restrictions quant à la précision des requêtes, au temps de réponse, au dédoublonnage des résultats, etc.
Le protocole Z 39.50, très utilisé sur le plan professionnel pour les échanges de notices entre catalogueurs, débouche sur un mode d'interopérabilité que nous appellerons « l'interopérabilité par conversion et copie » (*map and duplicate interoperability*) : si les ressources sont dans un format différent, on effectue une conversion complexe qui permet de récupérer les données avec un minimum de perte (le niveau de perte n'étant toutefois jamais

nul) pour les verser dans une seule base qui sera chargée de fournir le service d'interrogation à l'utilisateur.

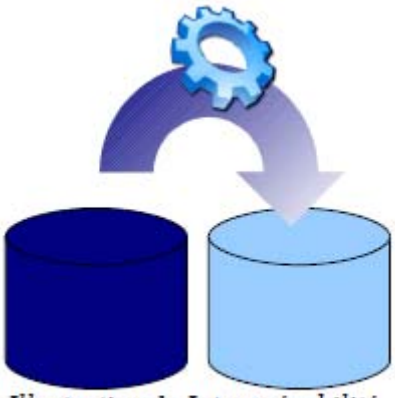
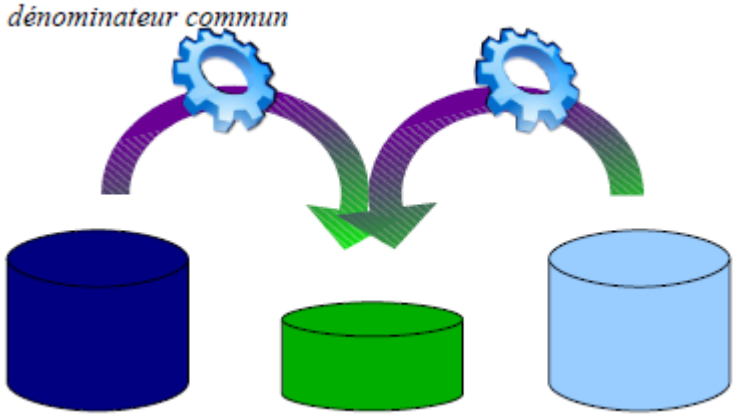


Illustration 1: Interopérabilité par conversion et copie

Le protocole OAI-PMH, mis en place dans les années 1990, propose une approche différente. Issu du mouvement de l'*open access*, il implique dès sa conception une préoccupation de convergence entre les données de la recherche (publications entreposées dans des archives ouvertes) et d'autres données issues notamment du domaine culturel et des bibliothèques.
Pour assurer l'interopérabilité entre ces différentes sources de données, il exige l'utilisation d'un format de données minimal commun, le Dublin Core dit « simple ». Les données ainsi formatées sont moissonnées, c'est-à-dire récupérées dans les bases réparties, pour être versées dans une base commune qui servira de support à la création de nouveaux services. On parlera alors d'interopérabilité basée sur le plus petit dénominateur commun (*smallest common denominator interoperability*).

Illustration 2: Interopérabilité basée sur le plus petit dénominateur commun



A nouveau, cette approche présente des limites. Les différentes sources sont contraintes d'appauvrir leurs données pour les faire entrer dans ce format commun, ce qui débouche soit sur la suppression de nombreuses informations, soit sur leur concaténation dans des champs de métadonnées généralistes, difficiles à exploiter. Ce n'est pas le Dublin Core lui-même qui est ici mis en cause, mais la façon dont il est employé dans le protocole OAI-PMH. Si on considère les données de bibliothèque, le Dublin Core simple fait par exemple perdre tout le bénéfice du système des notices d'autorité, qui fait le lien entre les notices bibliographiques. Toute l'information se retrouve à plat.

Pour ajouter une couche de complexité à ce panorama, notons que si l'on ne s'intéresse pas seulement aux bibliothèques, mais aussi à d'autres institutions culturelles telles que les archives et les musées, il faut prendre en compte la diversité des modèles de données définis par ces trois communautés.

Le modèle de base des bibliothèques repose sur deux concepts complémentaires, les notices bibliographiques qui décrivent les documents, et les notices d'autorité qui décrivent des entités (personnes, collectivités, concepts, etc.) que plusieurs notices bibliographiques peuvent avoir en commun.

Le modèle des archives met en avant la notion de contexte et de hiérarchie. Le format EAD, qui s'appuie sur le modèle de description de l'ISAD-G, permet de représenter les inventaires sous la forme d'une arborescence de composants qui favorise le respect des fonds. Des notions comme le titre ou l'auteur sont moins pertinentes dans ce contexte, alors qu'elles sont basiques dans celui de l'information bibliographique.

Enfin, l'information des musées est profondément déterminée par le fait qu'elle porte essentiellement sur des objets uniques. Ainsi, le contexte de ces objets est décrit, pas seulement comme dans le cas des archives en fonction de l'organisation des ressources, mais en fonction des différents événements auxquels l'objet est confronté, de sa création à sa conservation en passant par les différentes opérations de restauration et d'exposition qui ont pu l'affecter. Ce concept d'événement devient central dans le modèle, et c'est à travers lui que l'on relie les œuvres aux personnes. Ainsi le modèle CRM du CIDOC accorde une place structurante à l'événement.

Ces profondes différences de modèle au sein même des métiers du patrimoine culturel font de la convergence des données un véritable challenge. Réduire des données de bibliothèques, d'archives et de musées à un modèle commun implique de renoncer aux particularités de traitement et de conception de chacun de ces domaines, et réduit la construction de services communs à son plus simple élément. De plus, ces méthodes d'interopérabilité ne prennent pas vraiment acte de l'évolution des usages dans le contexte du Web, tel que nous l'avons évoqué en introduction. En effet, elles impliquent toujours un postulat de départ qui est que l'usager connaît l'existence de ces services, et fait la démarche de se rendre sur la page d'accueil de la bibliothèque ou du portail pour se positionner dans une démarche de recherche.

2. Le Web de données et l'interopérabilité basée sur les liens

Le Web de données propose une forme d'interopérabilité qui ne repose ni sur l'interrogation synchrone de bases réparties, ni sur la réduction de bases diverses à un format commun, mais sur la création d'un espace global d'information, utilisant les liens pour permettre de naviguer de manière transparente d'une ressource à l'autre.
Le Web de données est une extension du Web qui doit permettre de créer un espace global d'information, au-delà des documents, pour les données. Les règles de bonne pratiques du Web de données, énoncées par Tim Berners Lee puis adaptées par le groupe SWEO (Semantic Web Education and Outreach), sont au nombre de quatre :

- utiliser des URI (uniform ressource identifier) pour identifier les ressources : chaque ressource sur laquelle on veut pouvoir faire des assertions doit se voir affecter un identifiant Web, une URI ;

- ces URI doivent être formulées suivant le protocole HTTP afin qu'on puisse les actionner pour accéder à la ressource identifiée, ou à des informations sur cette ressource ;
- lorsqu'on accède à une ressource via son URI, celle-ci doit renvoyer des informations utiles et pertinentes en utilisant les standards (RDF, SPARQL) ;
- enfin, les ressources doivent être reliées, c'est-à-dire qu'il ne suffit pas de publier des informations, mais il faut les relier à des informations publiées par d'autres, afin de créer un écosystème basé sur les liens.

L'objectif est de créer un espace global d'information où les données sont décrites suivant un modèle commun, le modèle RDF, et reliées par des liens actifs, exploitables par des machines.

Grâce aux principes du modèle RDF, les liens entre les données sont typés, c'est-à-dire qu'ils qualifient le type de relation qui relie deux ressources : similarité, relation de sujet (« *aboutness* »), ou autre. Dans cette approche, il est possible de créer des liens entre des ressources décrites en utilisant divers modèles, à partir du moment où la grammaire de base, commune à tous ces modèles, est le RDF.

Deux modèles d'interopérabilité permettent de représenter cette nouvelle façon de travailler les données : le modèle de la roue et de l'essieu (*« hub and spoke »*) et le modèle de la navigation intuitive (*« follow your nose »*).

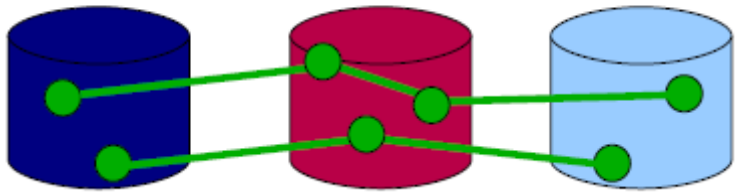
Les référentiels ou vocabulaires sont appelés à jouer un rôle vital dans le Web de données, en particulier lorsqu'il s'agit de construire l'interopérabilité entre des données issues de domaines différents. Sur le Web, un utilisateur a la possibilité de naviguer d'un site Web à un autre sans avoir connaissance des moyens techniques utilisés pour publier les données, sans même qu'il n'existe véritablement de rupture ou de frontière entre ce qu'on appelle les sites Web. De la même manière, sur le Web de données, la navigation de lien en lien doit pouvoir se faire, d'un jeu de données (*dataset*) à un autre, sans nécessité de percevoir les limites des différentes bases de données ni leur format.

Les référentiels sont volontiers associés au modèle « hub and spoke » : ils agissent comme un point nodal ou une colonne vertébrale permettant de créer un point de contact entre des jeux de données différents. Dans le Web de données, ce point de contact est suffisant pour naviguer sans contrainte d'un jeu de données à l'autre, en utilisant les URI, que les données soient ou non exprimées suivant le même modèle.

Illustration 3: Interopérabilité basée sur les liens : modèle "hub and spoke"

Pour aller encore plus loin, dans le Web de données, n'importe quel jeu de données dont on réutilise les données peut jouer ce même rôle de passerelle, quoique pas de manière centralisée : le fait de parcourir ces liens permet alors de découvrir de nouvelles ressources de façon intuitive (*« follow your nose » interoperability*).

Illustration 4: Interopérabilité par les liens : modèle "follow your nose"



3. Le rôle des vocabulaires dans le Web de données

Les vocabulaires sont donc appelés à jouer un rôle vital dans le contexte de l'interopérabilité basée sur les liens. Dans ses travaux, le groupe Library Linked Data du W3C (LLD XG) a défini deux types de vocabulaires : les vocabulaires de métadonnées, et les vocabulaires de valeurs (ou référentiels de valeurs) [LLD XG, 2011].

On englobe sous le terme de vocabulaires de métadonnées, ou jeux de métadonnées (*metadata element sets*), les vocabulaires qui servent à exprimer des propriétés et des classes (des types de ressources) réutilisables pour créer des descriptions. Ces vocabulaires sont exprimés à l'aide de schémas RDF (RDFS) ou d'ontologies en OWL (Web Ontology Language.) Les Dublin Core metadata terms sont un bon exemple d'un tel vocabulaire : ils fournissent un ensemble de classes et de propriétés utiles pour décrire des ressources documentaires (exemples de classes : Agent, Document... Exemples de propriétés : Creator, Format...) D'une certaine façon, les vocabulaires de métadonnées contribuent à créer du lien dans le Web de données, en autorisant le partage de typologies et de relations communes. Le projet Vocabulary Mapping Framework (http://cdlr.strath.ac.uk/VMF/) proposait une intéressante application de ce principe, dans laquelle différents vocabulaires de métadonnées étaient reliés suivant un modèle « hub and spoke » à une matrice commune permettant de gérer les conversions d'un format à un autre.

On parle de vocabulaire de valeurs ou référentiel de valeurs (*value vocabulary*) pour désigner un ensemble de termes organisés en système de connaissance (*Knowledge Organization System* ou KOS) pour être utilisés, notamment, en tant qu'objet dans les triplets. Les vocabulaires de valeur sont généralement utilisés pour lister des valeurs contrôlées dans le cadre de notices bibliographiques. On peut citer par exemple les LCSH (Library of Congress Subject Headings) ou encore le référentiel des codes de langues ISO 639-2, tous deux publiés en RDF sur le site http://id.loc.gov maintenu par la Bibliothèque du Congrès.

Grâce au principe du Web de données, l'utilisation d'un référentiel commun tel que les LCSH permet de faire des liens entre deux jeux de données même si ceux-ci sont exprimés suivant un modèle différent, en faisant appel à des classes et propriétés différentes. Ils partagent alors un même vocabulaire de valeurs. La navigation de lien en lien dans le Web de données doit

rendre possible l'exploitation conjointe de ressources décrites différemment, pourvu qu'elles aient un point de contact.

Pour les bibliothèques, le modèle des notices bibliographiques et d'autorités fonctionne déjà d'une manière similaire dès lors que des liens sont créés entre ces deux types de notices, et que leur cohérence ne repose pas que sur l'emploi de chaînes de caractères (les noms) normalisées. L'évolution vers les nouveaux modèles avec FRBR, FRAD et FRSAD, et ensuite l'évolution des règles de catalogage vers RDA qui s'appuie sur les mêmes concepts, prend également acte de la nécessité de mutualiser davantage les informations par la création de liens, non plus en recopiant les notices d'une base à l'autre, mais à l'intérieur même d'une notice. Dans ces nouveaux modèles, les notions qui peuvent prétendre au rang de référentiel, au sens d'informations qui sont partagées et réutilisées dans différentes descriptions et servent à faire du lien, se multiplient : l'œuvre, l'expression, les personnes, les collectivités, les familles, les sujets. Ces notions qui deviennent centrales dans les nouveaux modèles de l'information bibliographique pourront être mutualisées avec d'autres métiers et ainsi contribuer à porter les données des bibliothèques sur le Web.

Le fait de s'intéresser aux données des archives et des musées implique de prendre en compte une modélisation principalement déterminée par l'existence d'objets ou de documents uniques.

C'est sans doute pour cette raison que ces communautés se sont moins tôt intéressées à la problématique des référentiels, et longtemps il n'a pas existé l'équivalent des notices d'autorité (qui toutefois se développent aujourd'hui dans les archives avec l'EAC – encoded archival description). Dans les musées, il existe des référentiel de valeurs de type thésaurus et classifications (ex. les différents thésaurus du Getty pour les sujets, les lieux, les artistes, etc., ou encore le système de classification iconographique IconClass) qui permettent de rendre tangible le contenu des objets graphiques.

Les nouveaux modèles tendent à développer l'idée de mise en relation de ressources entre elles en se basant sur des liens, favorisant ainsi la découverte de nouvelles ressources par rebond. Les liens qui vont permettre de connecter ainsi les ressources sont des entités telles que des personnes, des événements, des lieux, des concepts. Or ce type d'entité, qui correspond aux notices d'autorité des bibliothèques, est également le type même de ressources qui peuvent être partagées au-delà des limites d'un type d'institution culturelle en particulier.

De la même manière, les référentiels particuliers que sont les classifications telles que la Dewey, la CDU, IconClass... utilisent des valeurs chiffrées qui permettent, en plus de jouer le rôle de « hub and spoke » que nous avons déjà souligné, de construire des services comme le multilinguisme [Dunsire, 2010].

L'utilisation des référentiels de valeurs par différents jeux de données va permettre de créer naturellement une interopérabilité de type « hub and spoke » sans développements supplémentaires. Un exemple : les données bibliographiques de la BnF contiennent une référence à un plan de classement Dewey de haut niveau utilisé pour la *Bibliographie nationale française* notamment. La conversion de ce plan de classement en lien vers les URI fournies sur le site http://dewey.info est quasiment instantanée, et facilitée par le fait que les URI de Dewey.info sont construite à partir de l'indice Dewey lui-même (ex. pour la littérature française de fiction : http://dewey.info/class/843/). Ainsi, les ressources de la BnF seront dès

leur publication reliées au Web de données par ce biais, et on pourra faire des liens avec d'autres jeux de données qui seraient reliés à Dewey.info [Wenz, 2010].

Pour aller plus loin, on peut dire que certains jeux de données, qui ne sont pas particulièrement conçus pour jouer le rôle de référentiels de valeur, voient leur usage si largement répandu qu'ils vont finir par se comporter exactement de la même manière. Si les responsables de jeux de données font le choix de réutiliser des URI existantes au lieu de générer leurs propres URI locales, on aboutit au modèle de l'interopérabilité intuitive (« follow your nose ») : on passe directement du nouveau jeu de données ainsi publié à celui dont on réutilise les URI. Un exemple : DBPedia, extraction en RDF des données de Wikipédia réalisée par les chercheurs de l'Université Libre de Berlin et de l'Université de Leipzig en Allemagne, joue actuellement un rôle de « hub » pour le Web de données : en raison de sa dimension encyclopédique, DBPedia est souvent le premier choix pour se relier pour les jeux de données de toute nature. Si une bibliothèque décide, plutôt que de générer des URI pour les auteurs de ses ouvrages, de réutiliser des URI existantes, celles de DBPedia par exemple, il devient possible de naviguer directement non seulement de cette bibliothèque à DBPedia, mais aussi directement vers le jeu de données d'une autre institution, fond d'archives, musée, qui aurait fait le même choix.

Enfin, l'alignement des référentiels entre eux peut aussi créer des passerelles. Par exemple, les Archives nationales de France utilisent un thésaurus généraliste nommé « Thésaurus W », désormais publié dans le Web de données (http://www.archivesdefrance.culture.gouv.fr/thesaurus/). Ce thésaurus est relié à RAMEAU, le vocabulaire des vedettes matières de la Bibliothèque nationale de France. On pourrait ainsi relier entre elles une ressource des archives et une ressource de la bibliothèque en utilisant ces deux thésaurus et leurs liens.

4. Exemples

Nous pouvons présenter quelques exemples d'utilisation de ces principes pour construire des applications favorisant la convergence des données.

Europeana

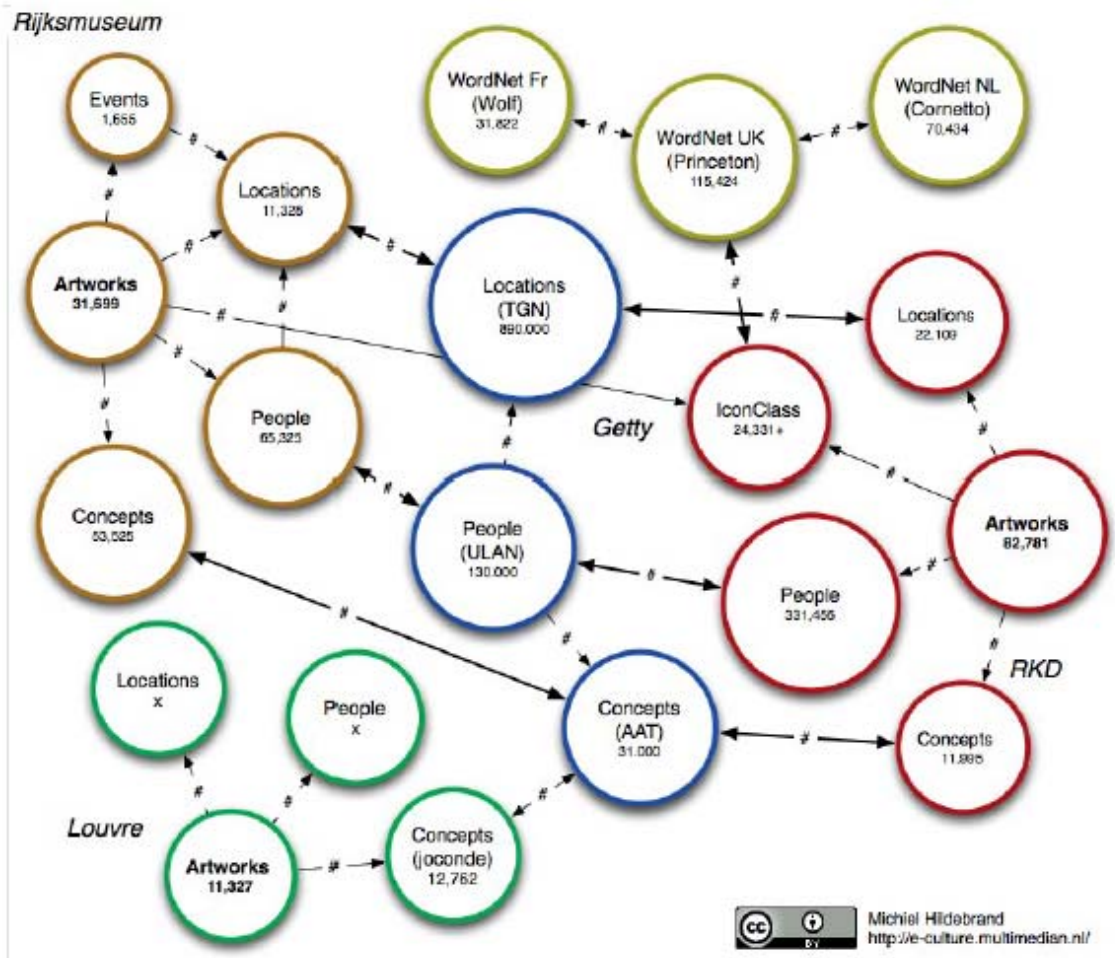
Le projet Europeana (http://www.europeana.eu) a pour vocation de relever le challenge de faire converger des données de bibliothèques, d'archives, de musées et d'archives audiovisuelles.

Europeana a construit son prototype en utilisant un modèle de type « plus petit dénominateur commun », basé sur le format ESE (Europeana Standard Elements) qui est une extension du Dublin Core simple pour lui ajouter principalement des informations de provenance et des éléments qui permettent de construire les liens vers des objets numériques distants stockés sur les sites des partenaires.

En parallèle de cette réalisation, Europeana développe le modèle EDM (Europeana Data Model), un modèle généraliste basé sur les principes du Web sémantique qui doit permettre de construire un réseau d'informations pour relier les ressources numérisées qui sont agrégées dans la bibliothèque numérique [Doerr, 2010]. Le modèle EDM permet d'agréger des ressources qui sont décrites suivant une logique documentaire aussi bien que suivant une logique orientée événement. C'est le réseau sémantique, c'est à dire les référentiels qui

décrivent les personnes, les lieux, les concepts etc. qui doivent faire le lien entre les ressources.

Il existe dans l'espace d'innovation d'Europeana, le Europeana Labs, un prototype qui démontre ces principes (http://eculture.c s.vu.nl/europeana/session/search). Ce prototype contient les données du Rijksmuseum Amsterdam et du Musée du Louvre, de la base Joconde du Ministère de la Culture français, ainsi que du Rijksbureau voor Kunsthistorische Documentatie (Netherlands Institute for Art History) à La Haye. Des thésaurus de lieux (The Getty Thesaurus of Geographic Names), de personnes (ULAN – The Union List of Artists Names), de concepts (WordNet et AAT – Art and Architecture Thesaurus), et une classification iconographique (IconClass) permettent de créer du lien entre les entités de ces différentes bases.



L'exemple du Centre Pompidou

Cet exemple montre que cette approche peut aussi être intéressante dans un cadre institutionnel. Dans le cadre de sa stratégique numérique développée depuis 2007, le Centre Pompidou a créé une nouvelle plateforme de diffusion de contenus numériques culturels sur Internet : le Centre Pompidou Virtuel. Ce nouveau site offre dans un espace unique l'ensemble de la production numérique du Centre Pompidou et de ses établissements associés (Bpi, Ircam) : œuvres numérisées, documents sur l'art, vidéos d'artistes, podcasts, notices des livres de la Bpi, etc. Les contenus artistiques et culturels (œuvres du musée, captations audiovisuelles...) sont reliés avec les événements (expositions, spectacles, conférences) et avec d'autres ressources pertinentes (affiches, photos de vernissages, livres, archives d'artistes...), permettant de parcourir le site de lien en lien pour découvrir ses contenus de façon intuitive.

L'un des principaux enjeux du projet était d'unifier dans un espace commun, permettant de nombreux liens et rebonds, des données issues de différentes bases structurées suivant des formats variés (EAD pour les bases archivistiques, MODS et Dublin Core pour les bases de bibliothèque, et des modèles locaux pour les bases du Musée et des archives audiovisuelles).
Pour cela une ontologie RDF a été créée, et articule autour de concepts majeurs (œuvre – ressource – personne – événement – collection, et quelques autres) toutes les données de ces différentes bases. Le Centre Pompidou Virtuel démontre ainsi la valeur ajoutée de l'utilisation des technologies du Web sémantique pour construire des rebonds entre des ressources et créer une expérience utilisateur innovante.

5. Les enjeux et les problématiques

Il existe néanmoins un certain nombre de barrières à la création d'un espace d'information global permettant de relier de manière transparente les données des bibliothèques, des archives et des musées.

La première barrière porte sur la conversion des données pour permettre leur publication dans le Web de données. L'adoption du modèle RDF questionne les formats et modèles existants, en partie parce que certaines informations sont complexes à représenter en utilisant le modèle de triplet, en partie parce que ce modèle ouvre des possibilités nouvelles qui invitent à remettre en cause les anciens modèles. Pour les bibliothèques, par exemple, le modèle du Web sémantique basé sur les liens suggère naturellement la compatibilité avec le modèle FRBR (et les autres modèles de la famille FR). Or, les données existantes, stockées en masses énormes dans les catalogues actuels, ne sont pas pleinement compatibles avec ce nouveau modèle, et de nombreux liens sont manquants pour pouvoir procéder à une conversion satisfaisante de ces données en RDF [Koster, 2011]. Le même type de question va se poser pour les données des archives, avec le modèle de description hiérarchique de l'EAD qu'il faut faire évoluer vers un modèle de graphe.

Une autre difficulté majeure est liée à l'attribution des URI. En effet, dans le modèle RDF, chaque ressource sur laquelle on veut pouvoir faire des assertions, et chaque propriété permettant de relier des ressources entre elles ou de les qualifier, doivent se voir attribuer un identifiant Web pérenne, une URI, et pour être utilisable sur le Web de données, une URI compatible avec le protocole HTTP. La question de la maintenance des URI se pose donc à plusieurs niveaux.

En ce qui concerne les jeux de métadonnées, il faut que les instances qui ont autorité pour publier les jeux de métadonnées (l'IFLA par exemple, dans le cas des FR** ou de l'ISBD) effectuent la conversion de leurs standards vers le Web de données et affectent des URI qui pourront ensuite être réutilisées dans d'autres contextes. Il est important de noter que si ce ne sont pas ces instances qui entreprennent ce travail, les URI seront publiées par d'autres, et il existe un risque de voir se multiplier différents vocabulaires RDF pour le même modèle, sans qu'aucun ne soit véritablement validé : c'est ce qui avait commencé à se produire pour les FRBR. Il faut noter qu'à l'IFLA, ce travail a maintenant été entrepris, au sein de la section de catalogage (FRBR review Group, ISBD/XML group) avec la création d'un Namespaces task group qui prend en charge la coordination et la dimension stratégique de cette action. Du côté des jeux de données eux-mêmes, il faut attribuer des URI qui permettent de représenter les documents, les auteurs, mais aussi les sujets, c'est-à-dire les concepts, lieux, etc. qui font partie des référentiels de valeurs. Il est très important que les référentiels de valeurs ne fonctionnent pas uniquement avec des chaînes de caractères, comme c'est encore souvent le cas dans les modèles de données actuels des bibliothèques, mais en attribuant à chaque ressource un identifiant qui permet de la gérer de manière interopérable et reliée dans le Web de données.

Les questions habituelles de gestion d'identifiants vont se poser : comment pérenniser les URI, vaut-il mieux choisir des URI « opaques » dont les chaînes de caractères ne transportent aucune sémantique, ou des URI signifiantes plus faciles à manipuler pour des agents humains... [Bermes, 2009]

La modélisation des référentiels de valeurs n'est pas anodine non plus [Vatant, 2010]. Les travaux du LLD XG ont montré que la dualité des référentiels tels que les autorités des bibliothèques pouvait poser problème : ces référentiels sont à la fois des listes de « noms », de labels, avec des recommandations sur la manière de nommer les entités qu'elles décrivent, et des référentiels qui donnent des informations sur des entités du monde réel comme les personnes en particulier. Il est nécessaire de recourir à des modèles complexes tels que celui de VIAF, qui dissocie le nom ou « label », le concept dans le contexte de la liste d'autorité, et la personne qu'il représentent.

Cette complexité n'est pas seulement un problème de pure modélisation. Elle se justifie dans le Web de données avec la problématique des alignements. Comme nous l'avons vu, il est nécessaire pour être présent dans le Web de données de créer des liens avec des jeux de données existants. La tâche de celui qui veut publier des données est donc triple :

- identifier les points de contact entre son jeu de données et ceux qui existent,
- déterminer précisément la nature du lien entre ses entités et celles du jeu de données avec lequel il veut se relier,
- et mettre au point la meilleure méthode pour créer ces liens, manuellement ou automatiquement.

La modélisation des liens entre les différents jeux de données, et entre les référentiels de valeur, est une question sensible, source potentielle de problèmes logiques dans l'environnement du Web de données [Bergman, 2010]. Le premier réflexe pourrait être de déclarer les ressources qui décrivent une même entité comme équivalentes (par exemple, Victor Hugo dans DBPedia et Victor Hugo dans le fichier d'autorité de la Bibliothèque nationale de France). Or l'utilisation de la propriété owl:sameAs, qui indique que les deux ressources sont totalement équivalentes, implique que toutes les propriétés qui s'appliquent à l'une doivent aussi s'appliquer à l'autre. Ainsi, si on avait une information telle que la date de

création de l'autorité « Victor Hugo » dans le fichier BnF, on créerait une incohérence logique en déclarant l'équivalence avec DBPedia, cette propriété devant s'appliquer également à la ressource équivalente.

Pour cette raison, des relations d'équivalence moins fortes sont mises à disposition par différents vocabulaires de métadonnées : par exemple skos:exactMatch et skos:closeMatch qui permettent de gérer les alignements automatiques dont la certitude n'est pas toujours optimale. Il existe aussi d'autres relations de similarité ou de proximité assez génériques, comme umbel:isLike ou rdf:seeAlso. Enfin, rien n'interdit de relier un jeu de données de bibliothèque à un jeu de données comme VIAF (personnes et collectivités) avec des propriétés de type dc:creator ou dc:subject, ce qui revient au modèle d'interopérabilité « follow your nose » que nous avons évoqué.

La question de la génération des alignements n'est pas non plus anodine quand on se repose sur d'importantes masses de données, comme c'est le cas des bibliothèques. L'idéal serait bien entendu de pouvoir générer ces alignements automatiquement, mais dès qu'on sort d'un contexte thématique ou institutionnel précis, la comparaison des chaînes de caractères n'est plus suffisante. Par exemple, au Centre Pompidou, nous avons constaté que dons un contexte où le corpus d'artistes est relativement restreint, l'alignement sur la chaîne « prénom, nom » est pertinent dans la plupart des cas. En revanche, pour aligner les événements issus de différentes bases, la diversité possible de saisie des titres d'événements est telle (“la subversion des images”, “exposition la subversion des images”, “exposition au Centre Pompidou : la subversion des images”, “subversion des images 2010” etc.) qu'il faudrait combiner de nombreux autres critères pour éviter les doublons et les faux positifs (“conférence dans le cadre de l'exposition la subversion des images”). Même pour les personnes, si la base est plus large, comme dans le cadre de VIAF, on va être confronté à de nombreux homonymes. Le meilleur moyen d'assurer les alignements est de disposer de « clefs » ou d'identifiants, qui ne sont pas forcément des URIs, mais qui sont suffisamment normalisés pour être utilisés dans différents jeux de données. Les identifiants ISO type ISSN, ISBN, ISNI ont un rôle essentiel à jouer dans le Web de données, d'où l'importance de les réinjecter dans les catalogues locaux pour préparer les alignements de demain. L'autre solution pour aligner les référentiels est d'utiliser la logique des liens : ainsi dans VIAF, on utilise les données bibliographiques auxquelles chaque auteur est relié pour augmenter l'indice de confiance des alignements.

6. Conclusion

En conclusion, le développement du Web de données repose sur la capacité des institutions à publier leurs données, mais peut-être encore davantage sur les liens qu'on pourra créer entre ces données, afin de construire l'interopérabilité basée sur les liens du modèle « hub and spoke » ou du modèle « follow your nose ». Pour cela, on a plus besoin que jamais des référentiels, outils partagés de modélisation de la connaissance, qui permettent de créer des points de contact entre des jeux de données qui partagent une similarité de contenus, mais adoptent des modèles de description différents. Cette nouvelle forme d'interopérabilité s'inscrit dans l'écosystème du Web et permettra de développer de nouveaux services, plus adaptés à la découverte intuitive des ressources et à la prise en compte des vrais objectifs des usagers, en synergie avec les ressources du Web.

7. Références

[Bergman, 2010] Mike Bergman, « Bridging the Gaps: Adaptive Approaches to Data Interoperability. » Keynote presentation for DC 2010, Pittsburgh, Octobre 2010. <http://dublincore.org/workshops/dc2010/DC-2010_20101022_Bergman_keynote.pdf>

[Bermes, 2009] Emmanuelle Bermes, « Linking Open Data : a case for releasing library data on the Semantic Web. » Emerging trends in technology, IFLA satellite conference, Florence, August 2009. <http://www.ifla2009satelliteflorence.it/meeting3/program/assets/EmmanuelleBermes.pdf>

[Doerr, 2010] Martin Doerr, Stefan Gradmann, Steffen Hennicke, et. al. « The Europeana Data Model. » IFLA 76th congress, Göteborg, August 2010. <http://www.ifla.org/files/hq/papers/ifla76/149-doerr-en.pdf>

[Dunsire, 2010] Gordon Dunsire, « Not just numbers on shelves: using the DDC for information retrieval ». EDUG meeting, Alexandria, April 2010. <http://www.slainte.org.uk/edug/meetings.htm>

[Koster, 2011] Lukas Koster, « Missing links : the challenge of generating linked data from legacy databases. »Common Place, March 28th 2011. <http://commonplace.net/2011/03/missing-links/>

[LLD XG, 2011] Report from the Library Linked Data W3C incubator group (in progress). <http://www.w3.org/2005/Incubator/lld/wiki/Main_Page>

[Vatant, 2010] Bernard Vatant, « Porting library vocabularies to the Semantic Web, and back. A win-win round trip ». IFLA 76th congress, Göteborg, August 2010. <http://www.ifla.org/files/hq/papers/ifla76/149-vatant-en.pdf>

[Wenz, 2010] Romain Wenz, « data.bnf.fr: describing library resources through an information hub. » Semantic Web In Bibliotheken 2010, Cologne, October 2010.

About the Author

Emmanuelle BERMES, head of Prospective & data services at the National library of France (Bibliothèque nationale de France, BnF), has been working at BnF since 2003, first in digital libraries, then in metadata. These past two years, she's been involved in a semantic Web project, data.bnf.fr. Emmanuelle Bermes is a member of IFLA IT section
where she's contributing to the creation of a Semantic Web special interest group (SIG). She's also a co-chair of the Library Linked Data W3C incubator group.