Créer des relations, « sortir » de l'information des collections… Les technologies du web de données et du *linked data* jouent un rôle clé dans l'évolution des collections et de leurs utilisations.

Transformer les collections en information grâce aux technologies du web sémantique

Nos catalogues sont conçus, prévus, pour nous permettre de décrire nos collections, de les gérer et d'y donner accès. Cela a des conséquences sur les choix techniques et normatifs que nous faisons. Par exemple, nous mutualisons autant que possible la description bibliographique, identique ; mais nous n'avons pas de raison (sauf pour le Peb) de partager nos données d'exemplaires et encore moins nos données de gestion (liées aux étapes de la commande, par exemple). Les bases de données se sont donc juxtaposées côte à côte, parce qu'aucun usage prévu ne justifiait une autre approche. Hors de la recherche et de la gestion de documents, nos données en tant que telles ne servent (quasiment) à rien d'autre.

Nos données ont une valeur d'usage : c'est-à-dire qu'elles ont d'autant plus de valeur qu'elles « servent », qu'elles sont un apport de connaissance ou une aide à la décision. Si nous améliorons les conditions et facilitons les possibilités de réutilisation, leur valeur grandit et nous nous inscrivons plus solidement dans le *LOD Cloud diagram*[1]. Car le nouvel environnement du web, et les technologies du web de données et du *linked data*, nous invitent à porter sur elles un autre regard : les concevoir non plus comme des métadonnées (c'est-à-dire des données invitant aussitôt à détourner d'elles le regard, pour le porter sur l'objet qu'elles décrivent), mais comme des données, des objets exploitables en tant que tels, ayant un intérêt et une valeur intrinsèques.

LES DONNÉES, AU-DELÀ DES CATALOGUES

Nos bases offrent une masse d'informations considérable. Elles livrent tout d'abord des informations sur la collection et sur sa description. Mais ces deux notions s'élargissent rapidement :

- à la bibliothèque qui a classé, indexé, constitué les collections (histoire et disciplines de l'établissement, idéologie sous-jacente des bibliothécaires qui ont choisi ou non d'acheter « précocement » des livres sur le développement durable ou les *gender studies*, etc.) ;
- aux auteurs qui ont écrit les ouvrages ;
- au corps social qui a produit les objets éditoriaux ;
- aux usages qui en sont faits (statistiques de prêts, citations dans les publications scientifiques, etc).

Bref, il est possible de concevoir la collection comme un artefact produit par un ensemble d'acteurs, et à ce titre fournissant des informations sur ces différents acteurs.

La description de la collection renvoie donc à un ensemble d'objets, individus, structures, qui existent « dans le monde réel ».

Dans le contexte des catalogues, on constate que les données liées à la collection sont, en quelque sorte, « renfermées » sur elles-mêmes. Les PPN d'autorités par exemple ne servent qu'à identifier les notices d'autorités du Sudoc (et n'ont de sens qu'au sein du contexte « Sudoc »), et non les personnes elles-mêmes. On perçoit alors l'importance d'aligner les référentiels : la nature des informations que peut être amenée à diffuser une bibliothèque n'est plus biblio-centrée lorsqu'elle se lie à des concepts qui lui sont extérieurs. Les données bibliographiques sont par exemple liées à des données encyclopédiques (DPpedia[2]), statistiques (Insee), géographiques (Institut Géographique National – IGN) permettant ainsi de les associer à des objets, des lieux, des personnes.

La production bibliographique peut donc enrichir automatiquement l'ensemble des connaissances produites sur une aire géographique (et l'on découvrira alors que les auteurs nés dans telles ou telles régions réagissent plutôt positivement ou négativement aux questions d'immigration, par exemple).

DES TECHNOLOGIES AU SERVICE DE LA CONSTRUCTION D'UNE VÉRITABLE «CONNAISSANCE »

Les technologies du *linked data*, les données structurées en RDF[3], sont précisément conçues pour permettre l'élaboration d'une connaissance qui se construit par associations de concepts, entre bases de données multiples et diverses.

C'est ce processus de liage qui fait qu'une base bibliographique se transforme en un ensemble d'informations. Il s'agit d'associer des données d'une collection à des concepts « extérieurs » à la collection elle-même. Et ces liens ne sont pas seulement des renvois à des bases d'autorités (celles qui servent à indexer, ou à désigner une personne de manière univoque) ; ils s'ef-

[1] *Le Linking Open Data cloud diagram* (http://lod-cloud.net/) est une représentation cartographique de l'ensemble des sets de données publiés sur le web, exposés selon les technologies du web de données (ressources identifiées par des URI et formalisées en RDF) et connectés les uns aux autres.

[2] *DBpedia* est un projet universitaire et communautaire d'exploration et d'extraction automatiques de données dérivées de Wikipédia.

[3] *Resource Description Framework* : modèle de graphe destiné à décrire de façon formelle les ressources web et leurs métadonnées.

fectuent également vers les personnes ou les objets eux-mêmes, à savoir, entre autres, vers DBpedia, LinkedIn, International Standard Name Identifier (ISNI). En prenant le temps « d'aligner les référentiels », on facilite le parcours, les rebonds, les *mash-up*. Nos données sont exploitables, au même titre que beaucoup d'autres données, en tant que masse d'informations sur l'organisme qui les a produites, ou qui en fait usage. Aligner par exemple les identifiants issus de l'Enquête Statistique Générale auprès des Bibliothèques Universitaires (ESGBU) avec les identifiants du Portail d'Aide au Pilotage de l'Enseignement Supérieur (PAPESR) et ceux de DBPedia, peut, pour commencer, se faire dans un fichier CSV simple. La visualisation cartographique des universités françaises en fonction du nombre de leurs étudiants ou de leurs données budgétaires, ou encore l'évaluation du coefficient de corrélation entre taux de réussite aux examens et statistiques de prêts/téléchargements pourront ensuite être obtenues en croisant les informations issues de ces différentes bases.

Une fois les données produites, leur dépôt dans un endroit choisi (serveur local ou plate-forme régionale, portail d'*open data* du ministère, data.gouv.fr, etc) est déjà une forme de recontextualisation, qui incite à certaines réutilisations.

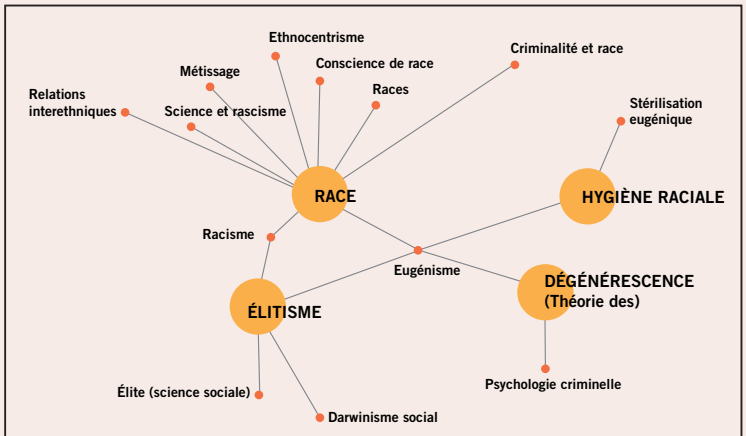
S'il est vain de vouloir anticiper ce qui pourra être fait avec ces données, on peut cependant chercher à être le premier bénéficiaire de cette exposition. Ainsi les bibliothèques de la North Caroline State University (NCSU) publient en Skos[4] une base de fournisseurs de contenus numériques qui alimente leur système de gestion des ressources électroniques développé localement. Plutôt que d'enfermer dans leur logiciel des fiches décrivant les fournisseurs, ils ont choisi d'alimenter directement une base ouverte où toute application (à commencer par les leurs) peut venir puiser à tout moment.

IMAGINER DES UTILISATIONS : L'EXEMPLE DES DONNÉES DE PRÊT

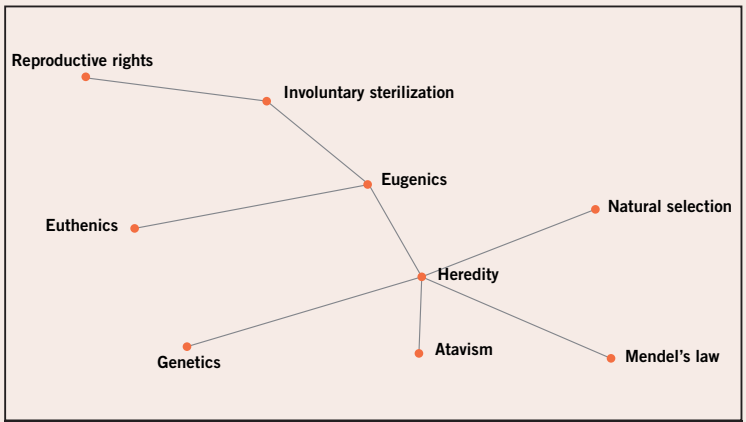
Les statistiques de prêt des bibliothèques permettent de mesurer la proximité entre deux documents ; proposer en *open data* ces données de « proximité » permettrait d'offrir un service de recommandation comparable à celui de certains sites web (« ceux qui ont acheté… ont aussi acheté… »). Une telle base pourrait ainsi alimenter une application d'« activités culturelles » qui suggèrerait à l'usager, en fonction de ses emprunts, une visite au musée, une séance de cinéma ou une lecture particulières. Ce pourrait aussi être des suggestions faites à l'issue d'une commande de billet de train ou d'avion : pour occuper votre voyage, téléchargez tel podcast, lisez tel livre. Certes, le livre en question ne sera peut-être pas l'exemplaire de la bibliothèque, cependant c'est bien grâce à ses données de prêt qu'il aura pu être recommandé…

L'INSCRIPTION DES LANGAGES D'INDEXATION DANS UNE CULTURE DONNÉE

Le langage d'indexation Rameau et son évolution au fil des décennies sont un reflet des modalités de catégorisation de thématiques contemporaines (le féminisme, la laïcité, etc.) qui changent au cours du temps. Comparé à d'autres langages d'indexation produits par d'autres « cultures », il illustre aussi la manière dont nous, bibliothécaires français, organisons la connaissance et structurons les termes de description du savoir comparativement à ces autres cultures.



Représentation en graphes de l'« éco-système » de la vedette Eugénisme produit dans Rameau



Représentation en graphes de l' « éco-système » de la vedette Eugenics produit dans le référentiel de la Bibliothèque du Congrès (Library of Congress, LOC)

La plus grande difficulté réside peut-être là : admettre que nos données, nos précieuses données, soient mêlées à d'autres, et exploitées en dehors des usages de nos collections.

Pourquoi ne pas considérer tout simplement que nos données soient aussi des collections ?

ETIENNE CAVALIÉ
GÉRALDINE GEOFFROY

SCD Université de Nice Sophia Antipolis
etienne.cavalie@unice.fr
geraldine.geoffroy@unice.fr

[4] *Simple Knowledge Organization System* (Système simple d'organisation des connaissances), recommandation du W3C qui permet de publier des référentiels et vocabulaires contrôlés sur le web de données.