Le web de données, *de « l'information en réseau »*

Le web de données (on préférera ce terme à celui plus ambigu de web sémantique), ce n'est pas compliqué ; ça marche et c'est utile, en particulier pour les bibliothèques.

RETOUR AUX RACINES DU WEB

Le web n'a pas été conçu pour n'être qu'un paquet de documents mis en lien. Il intègre, dès sa conception en 1989 par Tim Berners Lee[1], plus de sémantique que l'utilisation qui en sera faite ensuite. En particulier par la dualité Identifiant/Représentation :

• Identifiant : ce qui commence par « http://... » et que l'on voit dans la barre d'adresse de notre navigateur est une URL, où le « L » est mis pour « Locator ». C'est donc **l'adresse d'un document sur le web** ; mais ce n'est qu'un cas particulier des URI, où le « I » est mis pour « *Identifier* », qui sont des **identifiants, dans le contexte du web, de choses du monde réel.** On comprend donc qu'on peut identifier sur le web n'importe quoi à l'aide d'une URI : Victor Hugo, les pizzas *margherita*, le terme de thésaurus « gouvernance », la Loire, la caractéristique « se situe à », etc. On parle d'une façon générale de **ressources.**

• Représentation : si une URI est l'identifiant d'une « ressource », alors quel « document » obtiendra-t-on en naviguant vers cette URI ? On a l'habitude d'obtenir pour une même adresse toujours le même document, mais d'une façon générale un identifiant peut être associé à **plusieurs représentations** qui varient – de façon transparente – en fonction de préférences de langue, de format, de lieu, etc. C'est ce qu'on appelle la **négociation de contenu.**

Cette capacité des URI à identifier absolument n'importe quoi, indépendamment d'une représentation particulière, est la clé de voûte de l'universalité du web (de données).

Une fois les « choses » identifiées et rendues indépendantes des documents qui les représentent, il devient possible de parler de celles-ci : on peut publier sur le web l'assertion que « La Tour Eiffel se situe à Paris », en utilisant 3 URI pour identifier les 3 composantes de cette assertion : La Tour Eiffel, la notion de « se situer à », et Paris. C'est le standard RDF (*Resource Description Framework*) qui permet d'employer ces assertions en triplets. Notons au passage que, le web étant par nature décentralisé, n'importe qui est libre :

• de créer une nouvelle URI pour identifier Paris ;

• ou de créer une assertion en se référant à une URI déjà existante pour Paris (par exemple celle définie par l'Insee[2]) ;

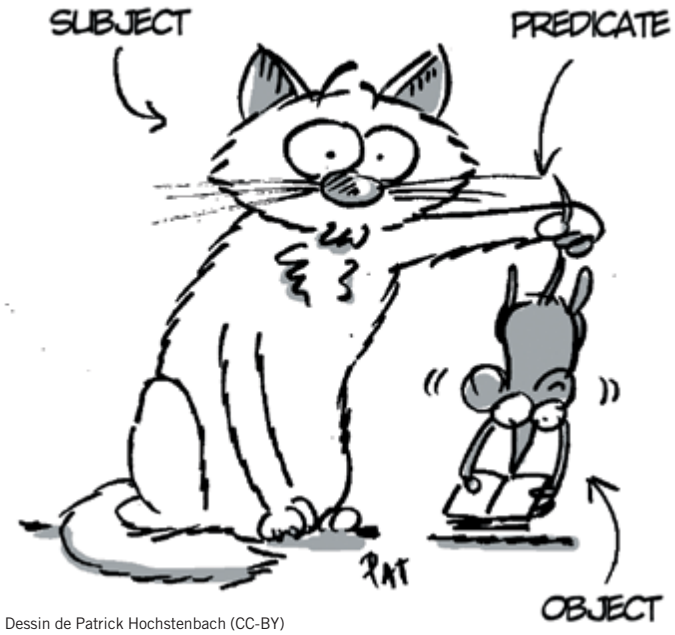
• ou encore d'exprimer des liens d'équivalence entre identifiants : l'URI que l'on définit pour Paris représente la même « chose » que celle définie par l'Insee.

On voit donc se dessiner ce qui nous occupe : un réseau décentralisé de données liées par des triplets.

Mais il faut aller plus loin pour que l'interopérabilité soit complète – puisque le web de données n'est qu'une solution à la problématique de l'interopérabilité. En effet, pour qu'une autre application puisse décoder mon assertion, il faut que je donne une définition précise des identifiants que j'ai utilisés, qui sont sans doute différents de ceux que comprend cette application. En particulier, il faut que je donne une définition précise de mes « verbes » (« est situé à »)

[1] Voir l'article de référence sur le sujet : Tim Berners-Lee, James Hendler and Ora Lassila, « The Semantic Web », *Scientific American,* Mai 2001.

[2] URI de Paris définie par l'Insee : **http://id.insee.fr/geo/commune/75056**, voir **http://rdf.insee.fr/geo**



Dessin de Patrick Hochstenbach (CC-BY)
Source : http://librecat.org/

et mes « types » (Lieu, personne, etc.). C'est ce que permettent les **ontologies**, dont l'objectif est de donner un sens univoque à ce dont je parle, à l'aide de la logique formelle (on parle également de **vocabulaire** ou de **modèle de données**, un peu par abus de langage). Les ontologies permettent également de déclarer des équivalences entre verbes ou entre types, rendant ainsi interopérables des données hétérogènes. Par exemple, je peux dire que, dans mon contexte, « est situé à » relie quelque chose à un « Lieu » et que cela représente la même notion que l'identifiant « *basedNear* » défini dans une autre ontologie bien connue, FOAF[3].
Les ontologies font donc émerger de cet océan de liens des structures interopérables, rendant ainsi les données liées plus « sémantiques », c'est-à-dire plus facilement réutilisables.

QUELS ENJEUX ET QUELLES CONSÉQUENCES ?

Souvenons-nous des fausses promesses entendues au milieu des années 2000 à propos du web de données : les machines allaient bientôt comprendre le sens des textes, on nous parlait de web 3.0, de « Twine » (un site qui n'existe plus maintenant mais qui promettait la révolution des réseaux sociaux), on cherchait quelle serait la « *killer-app* » – une application si attrayante qu'elle aurait justifié la technologie à elle seule, etc. Rien de tout cela n'est arrivé, mais d'autres conséquences ont eu lieu.

D'abord une quantité grandissante de « données ouvertes et liées » publiées par une variété importante de producteurs de données : c'est le fameux « *Linked Open Data* »[4]. Citons-en quelques points notables : DBPedia francophone (une extraction en RDF des données de Wikipedia), data.bnf.fr (notices FRBRisées – voir plus bas –, autorités et thématiques de la Bibliothèque nationale de France), ou encore VIAF (Virtual International Authority File, une mise en commun des fichiers d'autorité d'une quarantaine de bibliothèques et de musées).
Dans cet ensemble de données, il faut en mentionner certaines ayant un statut particulier : les thésaurus. Ceux-ci peuvent être représentés et publiés dans le modèle SKOS. Ce modèle permet d'aligner les thésaurus pour permettre l'interopérabilité de catalogues documentaires utilisant des vocabulaires d'indexation différents (« Désobéissance civile » dans Rameau est ainsi rapprochée de « *Civil disobedience* » dans les sujets de la Bibliothèque du Congrès américain[5]). Quant aux ontologies, on se référera au projet LOV - *Linked Open Vocabularies*[6].
Ensuite, une appropriation de cet enjeu des données structurées et liées par les grands moteurs de recherche : c'est l'initiative schema.org[7], qui propose un modèle de description de « plein de choses dont on parle sur le web » (blogs, livres, films, produits, etc.), compréhensible par Google, Yahoo, Bing et consorts. On peut reprocher à schema.org son biais vers le e-commerce, sa vision occidentalisée et son manque de transparence dans la gouvernance, mais si les bibliothèques souhaitent rendre leurs données plus visibles par les moteurs, cela passe par la publication de données compatibles avec schema.org.
D'une façon plus profonde, ces technologies induisent une représentation générale de l'information en **graphe décentralisé**, en réseau. Ce mode de structuration, de pensée, fait suite à celui plutôt tabulaire des bases relationnelles, et plutôt hiérarchique de XML. La conséquence est flagrante sur les notices bibliographiques avec le modèle FRBR. Les *Functionnal Requirement for Bibliographic Records*, successeurs de l'ISBD (*International Standard for Bibliographic Record*) proposent en effet un éclatement de la notice en 4 niveaux conceptuels, eux-mêmes reliés aux personnes ou aux organisations impliquées dans la vie du document (auteur, contributeur, éditeur, possesseur), lesquelles sont elles-mêmes reliées entre elles ou à d'autres données du web.
Cette tendance est à rapprocher du constat que de plus en plus de systèmes informatiques de diffusion des catalogues utilisent une base de graphe RDF (« *triplestore* ») pour centraliser les métadonnées des notices FRBRisées, les fiches d'autorité et les thésaurus. Cette base devient le pivot central des canaux de diffusion (sites web, flux RSS, formats d'échange métier, etc.). Les lois européennes sont notamment diffusées sur ce mode, via la base Cellar et le portail Eur-Lex[8].

PROCHAINES PROMESSES ?

Sans retomber dans les promesses hasardeuses évoquées plus haut, on peut néanmoins esquisser les lignes de force du web de données pour les prochaines années : une utilisation grandissante de schema.org par les moteurs de recherche et les projets de diffusion de données ; l'intégration native des fonctions de publication et de récupération des données du web dans les *Content management system* (CMS) et les SIGB ; la publication et l'alignement de plus en plus de données - dont des thésaurus ou des données de la recherche ; la généralisation de FRBR et de ses dérivés pour la description des notices bibliographiques, etc.
Au-delà des aspects technologiques, ce sont des logiques de partage, de réutilisation, de mise en réseau, de collaboration, ou d'insertion dans un écosystème d'acteurs, qui sont favorisés par cet artefact unique qu'est le web de données.

Thomas Francart
Consultant chez Sparna
thomas.francart@sparna.fr

[3] FOAF : **http://xmlns.com/foaf/spec/**

[4] *Linked Open Data* : **http://linkeddata.org/**

[5] En triplet RDF : **http://data.bnf.fr/ark:/12148/cb12049451f**

skos:closeMatch **http://id.loc.gov/authorities/subjects/sh90000103**

[6] LOV : **http://lov.okfn.org**

[7] **http://schema.org**

[8] EurLex : **http://eurlex.europa.eu**