Association
des Bibliothécaires
de France

Le web des données

Date de publication : 20/10/2010

Le web des données est **le projet de rendre le web (et les informations qu'il véhicule) aussi aisément lisible et exploitable par des ordinateurs que par les internautes**, ce qui n'est actuellement pas le cas et est à l'origine de plusieurs "frustrations" vis-à-vis du web "normal".

Frustration 1 : les liens ne sont pas typés

Les pages Wikipedia de Jean-Paul Sartre et Paris sont liées par un lien hypertexte. Un ordinateur voit ce lien, mais ignore que **la relation entre eux est : le premier est né dans le second**.

Le web des données consiste notamment à faire comprendre à un ordinateur :

- une information globale disant : l'objet décrit dans cette page est un être humain ;
- au moment de pointer vers la page Paris : l'indication "Ceci est le lieu de naissance du sujet principal de cette page".

Mais qu'est-ce qu'un être humain ? Dans le web des données, un être humain **est une ressource à laquelle on relie d'autres ressources selon certains types de relations**. L'essence ontologique d'une ressource est d'être **relationnelle**. Le fait qu'un être humain soit lié à un lieu de naissance, à d'autres personnes, etc., voilà ce qui le définit.

Plutôt que de décrire une ressource comme ayant un certain nombre de propriétés le web des données estime que chacune de ces propriétés est en soi une ressource liée à d'autres ressources :

Jean-Paul Sartre

- est né à Paris
- a écrit Huis Clos
  - a vécu avec
    - Simone de Beauvoir
      - qui est l'auteur du Deuxième Sexe
      - A joué un rôle dans l'histoire du féminisme
      - D'autres femmes ont joué un rôle important dans l'histoire du féminisme
      - etc.

Plutôt que de représenter cela sous forme de listes à puces imbriquées, il apparaît rapidement plus lisible de faire un schéma, où le centre est forcément glissant.

Une ontologie vise à définir l'essence de son objet d'étude.
Une ontologie informatique définit, pour une ressource donnée, quelles relations peuvent lier cette ressource à d'autres ressources. Un être humain est un type de ressource, un lieu géographique en est un autre. Chaque type de ressources est liée à d'autres ressources par des relations différentes.
L'ontologie Relationship définit quelles sont les relations définissant un être humain (ami de, fils de, etc.).

L'ensemble Ressource-lien-Ressource (ou Sujet-Verbe-Complément) est un triplet. La deuxième ressource peut n'être qu'une chaîne de caractères : Individu - a pour nom - Sartre.

Frustration 2 : affichées sur le web, les données ne sont plus structurées

Le moteur de recherche indexant les pages d'un catalogue de bibliothèque ne "sait" pas qu'il s'agit de livres, et il ignore ce que signifie "être le titre d'un livre", etc. Une des premières concrétisations du web des données est de restituer la structure de la base à l'intérieur des pages web.

Exemple : pour *Huis clos*, le PPN de la notice Sudoc est : 000563668. Cet identifiant peut s'exprimer sous forme d'URL : http://www.sudoc.abes.fr/DB=2.1/SRCH?IKT=12&TRM=000563668.

Cette ressource

- **est un livre**
  - les types de relations qui caractérises un livre sont listées et définies par l'ontologie BIBO (BIBliographic Ontology).

- **et, à ce titre, a un certain nombre de propriétés** :
  - il a pour titre : *Huis Clos*
  - il a pour auteur : Jean-Paul Sartre, qui est une ressource (associée aussi à d'autres livres) qui a pour **PPN 027123227**.

En Dublin Core, l'auteur est appelé : **creator**. Le triplet reliant l'œuvre à l'auteur peut être représenté et présent dans la page web dans des balises cachées, ou dans une page alternative liée.

L'information structurée initiale n'est plus perdue quand affichée sur le web. Elle devient réexploitable très facilement pour des sites voulant combiner les informations tirées de multiples sources (mashups).

Frustration 3 : quand elles sont structurées, ce n'est pas assez normalisé

Le **Z39.50** comme les **API** récupèrent des informations structurées. Mais le Z39.50 est extérieur au web (ce n'est pas du http), il est spécifique aux bibliothèques, et limité aux métadonnées bibliographiques. Quant aux API, chaque site a les siennes et il faut faire des développements spécifiques pour les exploiter.

Alors qu'en intégrant dans ses pages web de résultats des informations structurées sur le modèle de triplets, toutes les limites mentionnées ci-dessus sautent.

Réalisations

On peut en distinguer trois types :

- **les sites qui exposent leurs données** (le catalogue Libris fournit pour chaque notice l'équivalent en triplets) ;
- **les sites servant de sources à des applications** (DBpedia : Vous pouvez ainsi comparer l'article Jean-Paul Sartre sur Wikipedia et sur DBpedia. Bpedia est pour beaucoup dans l'essor du web des données, en exposant des milliers d'informations tripletisées de la Wikipedia, récupérables par d'autres bases) ;
- **les sites qui exploitent les données des autres** pour produire un contenu nouveau (Geonames, BBC Music). Ces sites réexposent à leur tour leurs données en triplets.

Le web "actuel" est déjà capable d'exploiter les données d'autres sites (cf. la notice de *Huis Clos* sur Calice68). Donc ça ne semble pas nouveau. Mais ces enrichissements déjà existants nécessitent l'analyse du code non normalisé, mouvant des sites sources. Ils peuvent aussi utiliser des API Amazon ou LibraryThing, spécifiques à chaque source -- alors que le web des données universalise l'encodage des métadonnées, donc facilite leur récupération.

Pour trouver d'autres applications, vous pouvez partir du schéma fourni par Linked Data, en cliquant sur un des projets pour voir "ce que ça donne".

RDF ?

Le **RDF** liste des consignes pour décrire des ressources : pour avoir le label "RDF", il faut décrire ses données sous forme de triplets, et que chaque ressource soit désignée par un identifiant unique. Cela permet ensuite à un développeur de se brancher sur plusieurs sites RDF, pour en extraire les informations qui l'intéresse, grâce à cette structure commune.

Le RDFa est du RDF appliqué aux pages web. Comment intégrer ces triplets dans les pages web sans gêner l'internaute ? Le RDFa explique quels balises et attributs HTML utiliser.

Perspectives

L'objectif n'est pas de remplacer le web mais d'interconnecter le plus grand nombre de sites. Il est vain de se demander si tous les sites "migreront" vers du web des données : Un site proposant un fil RSS mais pas de possibilité de commenter est-il 2.0 ? il y a plusieurs niveaux d'intégration, pour le web 2 comme pour le web des données.

Concernant les bibliothèques, on pourra commencer par :

- le remplacement d'OpenURL par RDF ;
- l'intégration facilitée de contenus enrichis (pages de couvertures, bibliographies, biographies, etc.) de manière beaucoup plus fine. Actuellement, si on veut ajouter du contenu Wikipedia à un site, il faut prendre de gros blocs d'informations. Avec le web des données, l'affichage d'informations sera beaucoup plus fin.

- des catalogues fédérés à la carte, loin du Z39.50 et des connecteurs.

Et plein d'autres choses, existantes ou qui viendront plus tard, avec la familiarisation et l'imagination.