Web sémantique, données libres et liées, UNT

Yolaine Bourda

September 20, 2012

Web sémantique

De nombreux documents sont présents sur le Web. Pourtant il est parfois difficile d'avoir des réponses à des questions simples comme : quelles sont les villes de plus de 15 millions d'habitants ou quelles sont les institutions d'enseignement offrant des formations sur les smart-grids. Pourtant les réponses à ces questions peuvent être trouvées sur le Web mais il faut parcourir plusieurs documents pour y répondre. La solution à ce problème est le passage d'un Web de documents fait pour être consulté par des humains au Web sémantique, extension du Web actuel, fait pour être consulté, non seulement, par des humains mais aussi «traité» par des logiciels. Par traitement, il faut ici comprendre non seulement des calculs mais aussi des raisonnements c'est-à-dire la déduction de nouvelles connaissances.

Cette transformation du Web est actuellement en train de se produire. Elle nécessite la description des ressources[1] ainsi que la spécification des données et des connaissances manipulées (propriétés des données, relations entre elles...). On parle souvent de *métadonnées*[2] pour désigner ces descriptions[3] de ressources. Bien évidemment, si on veut utiliser des outils, il faut s'accorder sur les noms de ces métadonnées, leur signification, les contraintes qu'elles subissent, les valeurs qu'elles peuvent prendre, les ressources auxquelles elles s'appliquent... Cela revient à définir des *vocabulaires* ou *ontologies* en utilisant pour cela des langages formels, ceux du Web sémantique, définis par le W3C[4] et à utiliser ces définitions pour décrire les ressources.

RDF[5] (Resource Description Framework) est le formalisme fondamental du Web sémantique, sa structure de base est le *triplet RDF* : `<Res, Prop, Val>` exprimant que, pour la ressource d'identifiant `Res`, la propriété d'identifiant `Prop` prend la valeur `Val` qui peut-être soit l'identifiant d'une ressource (le créateur de telle ressource pédagogique est telle personne) soit une valeur d'un type de données (la date de création d'une ressource pédagogique est une date, le nom d'une personne est une chaîne de caractères). RDFS[6] (Resource Description Framework Vocabulary) et OWL[7] (Ontology Web Language) permettent de définir des ontologies, c'est-à-dire des classes de ressources (ex: la classe de toutes les personnes), des propriétés (ex: le nom de famille d'une personne), des contraintes (ex: une personne n'a qu'un seul nom de famille)... SPARQL[8] (Query Language for RDF) permet d'interroger des ensembles de triplets RDF.

Un point souvent négligé, est l'importance de la désignation des ressources d'une façon non ambigüe et pérenne en utilisant des IRI[9] (Internationalized Resource Identifiers), extension des URLs permettant de désigner tout type de ressource (y compris non numérique).

[1] Désigne, dans ce cadre, une entité (page web, personne, ville...) possédant un identifiant.
[2] Données décrivant des données.
[3] Le titre d'une page, la date de naissance d'une personne, l'audience à laquelle s'adresse une ressource pédagogique, le nombre d'habitants d'une ville sont des exemples de métadonnées.
[4] http://www.w3.org/standards/semanticweb/
[5] http://www.w3.org/TR/rdf-primer/
[6] http://www.w3.org/TR/rdf-schema/
[7] http://www.w3.org/TR/owl2-overview/
[8] http://www.w3.org/TR/rdf-sparql-query/
[9] http://tools.ietf.org/html/rfc3987

Données ouvertes (Open Data), données liées (Linked Data), données ouvertes liées (Open Linked Data)

La confusion est souvent faite entre données ouvertes et données liées. Elles ont un point commun : elles sont publiées par leur fournisseur et correspondent à une démarche ascendante (bottom-up) et non pas descendante (top-down).

Il n'y a pas vraiment de définition reconnue de ce qu'on entend par *données ouvertes*. Il s'agit d'une vision stratégique, politique, sociale, philosophique des données. Les données ouvertes peuvent être reproduites, copiées, publiées, utilisées dans des traitements... Il existe un certain nombre de licences comme celle d'Etalab[10], l'Open Knowledge Definition[11] en recense un certain nombre. L'Open Knowledge Foundation[12] promeut l'utilisation de données ouvertes. Au niveau technique, aucun format n'est imposé ou interdit tant qu'il est lisible, l'utilisation de formats ouverts (par opposition à propriétaires) est recommandée. Comme exemples de données ouvertes, on trouve non seulement des portails gouvernementaux, Royaume Uni[13], France[14], mais aussi des informations sur des villes comme Paris[15] ou Rennes[16].

La notion de *données liées* correspond à une vision plus technologique des données en relation avec le Web sémantique. Il s'agit d'une collection d'ensembles de données publiés en utilisant les langages du Web sémantique (RDF(S) et OWL), ensembles de données reliés les uns aux autres et interrogeables au moyen du langage SPARQL. Les deux points importants sont l'utilisation des langages du Web sémantique et les liaisons entre ensembles de données ce qui permet dans une même requête d'interroger plusieurs ensembles. De nombreux outils commencent à apparaître pour stocker les données, les visualiser de façons diverses et variées (camembert, ligne de temps...), proposer des interfaces de visualisation avec des cartes...

Des données liées peuvent être fermées (utilisées par une entreprise) ou ouvertes (disponibles sur le Web), on parle alors de *Linked Open Data*. Le *LOD cloud diagram*[17] recense des ensembles de Linked Open Data ainsi que leurs relations. De nombreux ensembles existent dans des domaines aussi variés que la musique, la géographie, le e-gouvernement, la chimie... Comme exemples d'applications basées sur les Linked Open Data, on peut citer la partie musique du site de la BBC[18] ou le site de e-commerce BestBuy.

Données liées et Enseignement

Le premier atelier[19] portant sur les applications des données ouvertes liées pour l'enseignement et l'apprentissage, *Linked Learning 2011: 1st International Workshop on eLearning Approaches for the Linked Data Age*, a eu lieu en 2011 lors de la conférence ESWC (Extended Semantic Web Conference) et est publié par CEUR[20]. Le deuxième atelier, *Linked Learning 2012 (LiLe2012)*[21] *2nd International Workshop on Learning and Education with the Web of Data*, a eu lieu au printemps 2012 lors de la conference WWW.

Peu d'applications réelles existent actuellement, citons le projet anglais lucero[22] qui a pour but de favoriser la publication de données issues de l'enseignement supérieur anglais, le projet mEducator[23] qui s'intéresse aux ressources pédagogiques médicales ou l'Open University[24].

[10] http://www.data.gouv.fr/Licence-Ouverte-Open-Licence
[11] http://opendefinition.org/
[12] http://okfn.org/
[13] http://data.gov.uk
[14] http://data.gouv.fr
[15] http://opendata.paris.fr
[16] http://www.data.rennes-metropole.fr/
[17] http://richard.cyganiak.de/2007/10/lod/
[18] http://www.bbc.co.uk/music
[19] http://projects.kmi.open.ac.uk/meducator/linkedlearning/
[20] http://ceur-ws.org/Vol-717/
[21] http://lile2012.linkededucation.org/
[22] http://lucero-project.info/lb/
[23] http://www.meducator.net/
[24] http://data.open.ac.uk/

Projet SemUNIT

Ce projet a pour ambition d'intégrer les métadonnées (et donc les ressources décrites) dans le Web sémantique et plus particulièrement dans les données liées ouvertes.

Dans un premier temps, il a été réalisé, en OWL, une ontologie novatrice du SupLOMFR basée sur les principes du MLR (Metadata for Learning Resources, norme ISO 19788), réutilisant FOAF[25] (Friend Of A Friend) pour décrire les personnes et dont les vocabulaires sont implémentés en SKOS[26] (SKOS Simple Knowledge Organization System).

Puis, une application basée sur les descriptions des ressources issues de plusieurs UNT (UNIT, UNISCIEL, UVED, UOH) a été créée avec les fonctionnalités suivantes :

- génération automatique d'URI,
- transformation automatique des métadonnées codées en XML en triplets RDF,
- import dans un entrepôt de données RDF,
- proposition de quelques services
  - interrogation des métadonnées pour la recherche d'une ressource pédagogique et présentation des résultats avec des «facettes»,
  - recherche d'un «expert» (dans ce cadre d'un enseignant) sur un sujet,
  - présentation des éléments de métadonnées en utilisant des outils existant et en fonction de leur type. Pour ceux prenant leurs valeurs dans un vocabulaire contrôlé (comme la nature de la ressource, son type pédagogique...), présentation sous forme de camemberts ou d'histogrammes, pour ceux ayant des dates comme valeur, présentation sous forme d'une ligne de temps. Comme nous ne disposions pas de données géographiques, nous n'avons pas pu montrer d'affichage sous forme de carte.
  - point d'entrée SPARQL (indexé par google) qui a servi, non seulement pour faire des interrogations, mais aussi pour valider les données, citons par exemple
    * ressources pour lesquelles tel élément de métadonnées n'est pas renseigné (intéressant pour les éléments recommandés),
    * vérification conjointe de la structure et du type d'agrégation.

En plus des services proposés, on peut en imaginer d'autres, comme

- la possibilité d'ajouter simplement des annotations (par exemple des étoiles ou des commentaires...)
- la mise en évidence simple des «trous» dans les ressources disponibles (avons nous des exercices et des cours dans tous les domaines ?)
- la prise en compte du profil utilisateur pour lui fournir les ressources appropriées (si besoin ou envie)
- des requêtes prenant en compte une ontologie de domaine
- des liaisons avec d'autres entrepôts de données ouvertes liées (dblp, dbpedia...)
- l'utilisation de l'élément de métadonnées relation pour récupérer une ressource et toutes celles qui en font partie, uniquement la dernière version d'une ressource, toutes les ressources référencées par une ressource donnée ou référençant une ressource donnée, toutes les ressources ayant comme pré-requis une ressource donnée...
- ajouter des informations de nature géographique pour des affichages sous forme de cartes
- ...

[25] http://www.foaf-project.org/
[26] http://www.w3.org/TR/skos-reference/

Données ouvertes liées et UNT

Actuellement, les ressources pédagogiques recensées par les UNT sont accessibles soit par le portail de chacune d'elles (UNIT[27], UNISCIEL[28]...), soit par un portail commun[29]. Ces ressources sont indexées en utilisant le schéma de métadonnées SupLOMFR[30]. Des logiciels, basés sur le protocole OAI-PMH (par exemple : ORI/OAI), prennent en charge cette indexation ainsi que le partage et le moissonnage des métadonnées. Celles-ci sont stockées dans des entrepôts XML et peuvent être interrogées via des formulaires d'interrogation. Bien que nécessaires, les services proposés sont limités et faits pour être utilisés par des être humains via des formulaires. Ils ne tirent pas complètement partie du travail d'indexation qui a été réalisé, travail couteux en temps et qui doit être rentabilisé au maximum. Les métadonnées ne sont pas aussi visibles qu'elles pourraient l'être, elles ne peuvent pas être traitées par des logiciels extérieurs. En conclusion, les ressources décrites ne sont pas aussi visibles qu'elles pourraient l'être.

Le projet SemUNIT a montré la faisabilité de l'intégration des métadonnées décrivant les ressources de l'enseignement supérieur dans le monde des données ouvertes liées. Cette approche commence à être de plus en plus largement répandue. Il est possible de faire le parallèle avec le début du Web.

L'intégration dans le monde des données ouvertes liées permet

- une meilleure exposition des métadonnées et donc des ressources décrites facilitant ainsi la réutilisation de ces dernières
- la possibilité, pour des utilisateurs extérieurs, de lier leurs données avec les données publiées augmentant encore la visibilité de celles-ci
- la possibilité de lier simplement les ressources pédagogiques aux formations quand celles-ci seront décrites elles-aussi sous forme de données ouvertes liées
- la possibilité de construire des parcours de formation ou d'apprentissage (en utilisant les relations)
- l'accès à un ensemble de données facilement extensible (ajout d'une ontologie de domaine pour faire des requêtes la prenant en compte)...
- la possibilité d'utiliser des outils ouverts prenant en compte ce genre de formats
- la transparence, l'inscription dans la mouvance data.gouv.fr
- la possibilité de faire des liaisons avec d'autres entrepôts : données issues de la recherche, données pédagogiques issues d'autres pays francophones ou non...

Conclusion

Parmi les valeurs ajoutées de l'approche données ouvertes liées citons, entre autres, des métadonnées «bien formées», l'interopérabilité des métadonnées entre des entrepôts différents, la possibilité de faire de l'analyse de données et des raisonnements ainsi qu'une meilleure visibilité des ressources décrites. Cette visibilité accrue entrainera une utilisation accrue des ressources décrites.

En ce qui concerne l'insertion des UNT dans le monde des données ouvertes liées, il ne faut surtout pas se satisfaire de l'application qui a été réalisée (maquette). Il faut, non seulement la pérenniser mais aussi l'étendre à la description des offres de formation (CDM-FR/MLO).

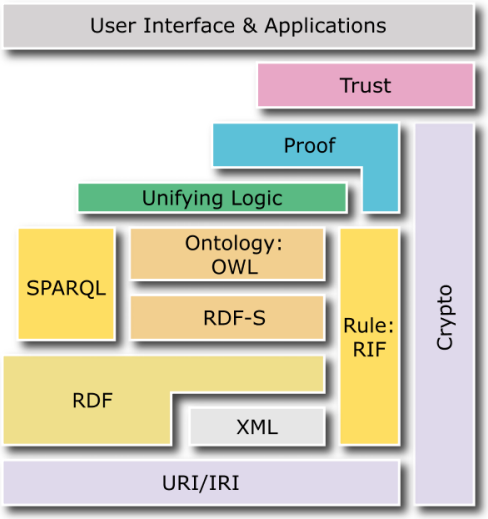
[27] http://www.unit.eu/fr
[28] http://www.unisciel.fr/
[29] http://www.universites-numeriques.fr/
[30] http://www.sup.lomfr.fr/index.php?title=Accueil

L'insertion des données produites par les UNT et, plus largement des données produites dans l'enseignement supérieur est stratégique et inscrit celui-ci dans la mouvance de plus en plus largement répandue de mise-à-disposition des données publiques (data.gouv.fr) dans un format exploitable.

Annexes

Le Layer Cake

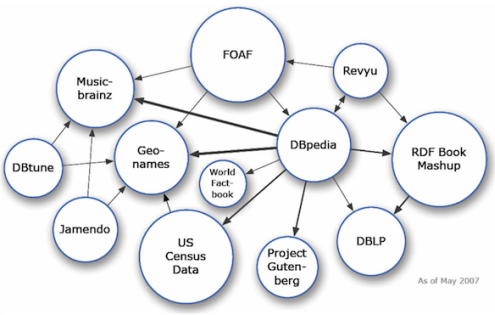
Les différentes couches du Web sémantique représentée par le fameux "gateau".



Linked Open Data

Une représentation graphique de la vision de Tim Berners-Lee des Open Linked Data[31].

Le LOD cloud en mai 2007



[31] http://www.w3.org/DesignIssues/LinkedData.html

Le LOD cloud en septembre 2011