Web de données liées et Web sémantique

Fabien Gandon

Web de données liées et Web sémantique

Fabien Gandon

L'évolution du Web documentaire vers le « Web des données » repose sur des principes et standards, permettant de tout identifier, de tout décrire sur la toile, et de tisser ainsi un graphe de données mondial. Le « Web sémantique », dans un second temps, permet la formalisation, la publication et le liage des vocabulaires utilisés dans ces descriptions. Ces évolutions permettent aux applications d'utiliser plus efficacement les données du Web en reconnaissant les différents types de ressources et de liens qu'elles rencontrent, et en exploitant le sens et les raisonnements qui leur sont attachés. Une application peut ainsi faire la différence entre des ressources nommées « Charles de Gaulle » mais de types différents (le général, une rue, une résidence, le poète, l'aéroport, le porte-avions…).

Identifier des ressources sur le Web

Le Web de données met en relation des sources de données plus ou moins grandes en reposant sur l'architecture classique du Web : *i)* des identifiants universels (URI pour *Universal Resource Identifier*) pour nommer sur le Web n'importe qu'elle ressource; *ii)* un protocole (HTTP pour *Hypertext Transfer Protocole*) pour, à partir d'une adresse (URL pour *Universal Resource Locator*), interroger une ressource et d'obtenir une représentation de celle-ci ; *iii)* un langage (HTML pour *Hypertext Markup Language*) pour représenter et communiquer ces représentations.

Le Web de données n'échange plus nécessairement des documents HTML, mais des données en général et dans différents formats. Ce ne sont donc plus uniquement des pages qui sont liées sur le Web, mais des identifiants de ressources arbitraires. Lorsqu'un identifiant est consulté, les serveurs répondent en fournissant des données décrivant la ressource sans que celle-ci soit nécessairement sur le Web (e.g. une voiture, une espèce animale, une protéine, un auteur….). L'appellation « Web de données » insiste donc sur la possibilité d'ouvrir nos silos de données de toutes tailles, depuis notre carnet d'adresses jusqu'aux immenses bases de génomique, et de les échanger, de les relier, de les composer selon nos besoins. En particulier, l'initiative des Données Ouvertes Liées (LOD pour *Linked Open Data*) promeut la mise en ligne de données en respectant des règles simples constituant cinq niveaux de qualité de publication des données (figure 1).

Décrire les ressources dans un graphe de données mondial

Représenter ces données requiert des modèles, des structures, des formats et des langages. RDF (pour *Resource Description Framework*) est au Web de données ce que HTML est au Web documentaire : le langage qui permet de représenter et de relier des données à propos de

ressources. RDF respecte l'architecture du Web, et notamment les URI pour identifier les ressources et relations décrites (figure 2). De telles descriptions peuvent provenir de n'importe quelle source sur le Web et être fusionnées avec d'autres. Le terme de « gigantesque graphe global » (*Global Giant Graph*) désigne parfois cette toile de données d'envergure mondiale tissée par des milliers de descriptions distribuées sur le Web déclarant des liens entre des nœuds identifiés par des URI.

Par ailleurs, RDF fournit également un modèle de données servant de fondation à d'autres standards. Ainsi, au-dessus de RDF, la recommandation SPARQL fournit principalement trois outils pour l'échange des données : *i)* un langage d'interrogation et de modification des graphes RDF ; *ii)* des formats pour les résultats d'une requête ou d'une modification ; *iii)* un protocole pour soumettre une requête à un serveur distant et recevoir les résultats, notamment au-dessus d'HTTP. Par exemple, sur le site DBpedia (cf. V.16), on peut demander en SPARQL tous les URI des ressources nommées « Paris » en français. A partir des identifiants reçus, on peut à nouveau interroger le site pour avoir des données supplémentaires et ainsi passer de données liées en données liées comme on passerait de page en page.

Vocabulaires et connaissances formelles

Différents types de modèles sont conçus pour fournir des vocabulaires permettant de décrire notre monde sur le Web (comme les ontologies ou les thésaurus). En interrogeant et en raisonnant sur ces modèles informatiques, il est possible d'améliorer des fonctionnalités existantes et d'en proposer de nouvelles. Au-dessus de RDF se dresse ainsi la pile des langages de schémas, ayant une expressivité et un coût de calcul croissants : plus l'on monte dans la pile et plus les définitions logiques du vocabulaire permettent de capturer précisément les structures et le sens des données, mais aussi plus les raisonnements sur ces schémas sont coûteux en termes de complexité et temps de calcul. Le premier niveau dit « des schémas légers » est celui de RDFS (*RDF Schema*) permettant de déclarer et de nommer les classes de ressources (comme les livres, les films, les personnes…) et leurs propriétés (comme l'auteur, l'acteur, le titre…) et d'organiser ces types dans des hiérarchies. Au-dessus de RDFS, la recommandation OWL (*Ontology Web Language*) permet de représenter formellement les définitions et s'organise en plusieurs fragments d'expressivité plus ou moins étendue, qui permettent des déductions supplémentaires en contrepartie de temps de calculs plus longs (figure 3).

Dans la continuité du Web de données, le « Web sémantique » met donc l'accent sur la possibilité d'échanger les schémas de nos données et la sémantique associée. Formalisés et publiés selon des standards, ces modèles permettent d'enrichir la gamme des traitements automatiques qui peuvent

être appliqués aux données. En ouvrant les données et leurs modèles, le Web de données et le Web sémantique ouvrent l'ensemble des utilisations qu'il est possible d'en faire.

Références

Fabien Gandon, Catherine Faron-Zucker, Olivier Corby, *Le Web Sémantique Comment lier les données et les schémas sur le web ?* InfoPro, Dunod, Mai 2012, EAN13 : 9782100572946

Tom Heath and Christian Bizer (2011) *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. Version électronique gratuite en ligne: http://linkeddatabook.com/

Affiliation

Fabien GANDON, Docteur et habilité en informatique, Directeur de Recherche chez Inria, responsable scientifique de l'équipe projet commune Wimmics (UCA, Inria, I3S, CNRS) et représentant d'Inria au W3C.

Figures

★ les données sont sur le Web sous licence libre

★ ★ idem + les données sont explicites et structurées

★ ★ ★ idem + les données sont dans un format non propriétaire (RDF)

★ ★ ★ ★ idem + des URI HTTP sont utilisés pour identifier sujets, objets et relations

★ ★ ★ ★ ★ idem + les données sont liées à d'autres données

Figure 1. Les cinq étapes et critères incrémentaux de qualité pour la publication de données ouvertes liées sur le Web.

```
<http://www.uniprot.org/uniprot/P43121.rdf#_5034333132310030> a rdf:Statement ;
    rdf:object <http://purl.uniprot.org/tissues/614> ;
    rdf:predicate :isolatedFrom ;
    rdf:subject <http://purl.uniprot.org/uniprot/P43121> .
```

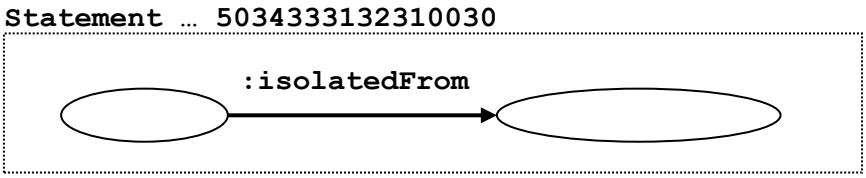
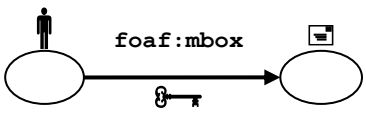


Figure 2. Petit sous-graphe extrait de la description d'une protéine (P43121) identifiée sur le Web par l'URI http://purl.uniprot.org/uniprot/P43121 ; ces données indiquent dans un format standard directement sur le Web qu'une protéine a été isolée à partir d'un tissu et chaque élément et relation est identifié par un URI permettant de découvrir des données supplémentaires.

```
foaf:mbox a owl:InverseFunctionalProperty;
    rdfs:domain foaf:Agent.
```

Avec schéma :



Ou avec image:

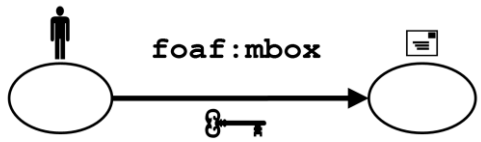


Figure 3. Extrait du vocabulaire FOAF permettant de décrire des personnes et utilisant pour cela les standards RDFS et OWL. Cette définition porte sur la propriété *mbox* qui permet d'indiquer une adresse mail pour une personne ou un groupe. Bien que simple, cette définition implique déjà que la ressource décrite peut être automatiquement classée dans la catégorie des agents (*domain*) et que deux agents ayant la même adresse sont les mêmes agents (*Inverse Functional Property*) l'adresse mail fournissant une clef d'identification de l'agent.