Introduction aux technologies sémantiques

Sommaire

Introduction

Dans les textes « Introduction à la gestion des connaissances dans les organisations » et « La modélisation des connaissances » nous avons souligné l'importance de la modélisation des connaissances et des compétences pour la gestion des connaissances dans les organisations. Dans celui-ci, nous présentons une introduction aux technologies sémantiques, une technologie incontournable pour la gestion des connaissances sur le web.

Le capital de connaissances d'une organisation est constitué de connaissances explicites et tacites, au niveau des individus qui composent l'organisation comme au niveau collectif de l'organisation elle-même. Particulièrement importants pour la gestion des connaissances, les processus d'*explicitation des connaissances tacites*, et *la combinaison des connaissances* une fois explicitées, requièrent l'adoption d'un langage formel partagé pour décrire les connaissances. C'est à cela précisément que visent à répondre les technologies du web sémantique ou, en bref, les *technologies sémantiques*.

Nous avons également mis en évidence un langage, des outils et des méthodes de représentation des connaissances et une variété de modèles (voir le texte « La modélisation des connaissances ») réalisés en un langage graphique (MOT), notamment les modèles de processus multi-acteurs (ou scénarios de travail ou d'activités), et les ontologies de domaine. Ces deux types de modèles sont formels, c'est-à-dire qu'ils sont dénués des ambiguïtés inhérentes, par exemple, aux langages naturels ou à des cartes conceptuelles semi-formelles. Les modèles formels servent à la communication entre les personnes, mais leur non-ambiguïté permet surtout le traitement des connaissances par des agents informatiques.

Dans ce texte, nous visons une première synthèse des technologies sémantiques, laquelle sera approfondie par la suite par des lectures actives dans le livre recommandé et par les activités du cours. Ce texte résume en français les premiers chapitres du livre recommandé, et aussi des éléments d'autres ouvrages. Il vous fournira une terminologie en français des principaux concepts qui sont jusqu'à maintenant surtout présentés dans des ouvrages de langue anglaise.

Dans une première section, nous caractérisons d'abord cette troisième génération du web qu'est le *web sémantique*, aussi appelé « le web de données » ou « web 3.0 ». Le web sémantique est fondé sur la représentation des connaissances par graphe RDF (Ressource Description Framework), ce qui permet des constituer des bases de triplets (*triple store*). Ce langage de représentation et ses sérialisations (traduction dans des fichiers de codes) est l'objet de la deuxième section. Une troisième section est consacrée à la topologie du web de données et à l'architecture des applications sémantiques. Une quatrième section présente quelques-unes des applications logicielles.

1. Du web de documents au web de données liées

Le web est devenu un immense réservoir de ressources d'informations et de connaissances avec ses billions de pages, en relation avec plus de deux milliards d'utilisateurs d'internet sur tous les continents. Ce réservoir d'informations subit une croissance extrêmement rapide. En décembre 2011, on recensait 555 millions de sites dont 300 millions nouveaux

sites au cours de la seule année 2011. On comptait plus de 800 millions d'utilisateurs de Facebook dont 200 millions de nouveaux utilisateurs au cours 2011, ainsi que 100 millions d'utilisateurs de Twitter. On note, toujours en 2011, plus de 200 milliards de vidéos visionnées sur YouTube par mois! Flickr héberge plus de six milliards de photos disponibles sur le web[1].

1.1 Les limites du web de documents

Cet immense réservoir de ressources pose un défi énorme aux utilisateurs qui cherchent des ressources adaptées à leurs besoins. Malgré l'efficacité des moteurs de recherche modernes, le web de documents présente un certain nombre de problèmes que le web sémantique vise à résoudre.

- Les documents sont décrits essentiellement par des mots clefs d'une langue naturelle, qui énoncent leurs propriétés (auteurs, sujet, média, etc.) ou se retrouvent dans le texte d'une page web. On les recherche par appariement de ces mots clefs avec les mots d'une requête. Or, un mot comme « Java » peut désigner aussi bien un langage informatique, une danse ou une île d'Indonésie, sans compter d'autres significations que ce mot peut prendre dans des langues étrangères. Cette ambiguïté inhérente aux langues naturelles fait en sorte qu'il faut multiplier et raffiner les recherches pour obtenir quelques centaines, voire des milliers de ressources qui approchent l'information que l'on cherche et qu'il faudrait consulter une à une.
- La réponse à une requête requiert une combinaison et une intégration de plusieurs sources de documents. Par exemple, une requête comme « Où pourrais-je aller en vacances, pour une semaine, le mois prochain, avec deux enfants pour moins de 2000 $ » est impossible avec les moteurs de recherche actuels. La combinaison des sources de données est laissée à l'utilisateur.
- Lorsqu'on recherche des informations sur le web de documents, l'ordinateur sert essentiellement à présenter des textes, des graphiques ou des fichiers audio ou vidéo. On utilise une très faible partie des capacités d'un ordinateur. Devant la masse des informations en croissance exponentielle, la seule façon de vraiment exploiter les ressources du web exige de déléguer aux ordinateurs de nombreuses tâches si on veut faire des rechercher plus intelligente, analyser les données et intégrer les informations.
- Les pages web sont désormais alimentées par des bases de données, en général des bases de données relationnelles, mais dont la structure n'est pas visible sur le web. De plus, ces bases de données ne sont pas interreliées, chacune ayant sa propre structure, ses propres contenus séparés des autres.

[1] Royal Pingdom (2013, 16 janvier) Internet 2012 in numbers. [billet de blogue]. Récupéré le 15 octobre 2013 : http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/

1.2 La gestion des connaissances sur le web

Au début des années 2000, le fondateur du web et actuel directeur du W3C, Tim Berners-Lee, et ses collègues[2] ont proposé d'intégrer au web des informations sur les connaissances traitées dans les ressources du web. En décrivant leur sémantique, au-delà de leur syntaxe, on pourrait traiter les informations du web par des agents informatiques capables de faire des recherches plus intelligentes qu'auparavant parce que fondées sur le sens, sur les connaissances derrière les mots. Cette évolution du web a été qualifiée de web sémantique.

Par ailleurs, les systèmes d'information des organisations transigent de plus en plus sur le web. Or, ces systèmes se sont construits de façon incrémentale, selon les applications et les processus qui apparaissaient critiques à un moment donné. Chaque nouvelle application entraîne la création d'une base de données qui alimente les portails et les pages web. Chaque base de données utilise une structure et des applications logicielles souvent différentes des autres, ce qui rend de plus en plus difficiles la recherche et l'exposition d'informations pertinentes à une tâche donnée.

S'ajoutent à ces bases de données hétérogènes des masses d'informations non structurées qui peuvent représenter jusqu'à 80 % de l'ensemble des données d'entreprise, enfouies dans des textes, des courriels, des échanges sur le web social (web 2.0). Ces connaissances tacites doivent être extraites, coordonnées et exploitées de façon intégrée avec les bases de données structurées. Les organisations doivent réaliser l'intégration de leurs sources de données, d'information et de connaissance en un ensemble cohérent, intégré et évolutif. Elles doivent le faire en laissant les sources d'information existantes en place, sans ajouter à la complexité des systèmes d'information.

Plusieurs solutions techniques ont été proposées pour répondre en partie à ce besoin, tels que les portails « d'intelligence d'affaires », le MDM (Master Data Management), le *big data* et les architectures orientées services. Ces dernières en particulier ont permis de comprendre que l'intégration des sources d'information doit se faire au niveau des données et non des applications et de l'échange de services entre elles. Pour vraiment décloisonner les données, il faut une nouvelle façon de les traiter par les technologies sémantiques (TS), notamment par la construction d'ontologies et de réseaux de données liées qui sont à la base du web sémantique.

Un exemple de technologie sémantique, utilisée avec succès dans un projet récent du Centre de recherche LICEF (projet PRIOWS), a consisté à construire une ontologie regroupant les principales connaissances d'une organisation (barrages, appareils, documentation, personnes, scénarios de travail) et à les mettre en relation avec une ou plusieurs des banques d'information de l'entreprise sans modifier ces dernières. On peut alors interroger les sources d'information par l'ontologie comme s'il s'agissait d'une seule source d'information intégrée. Un autre avantage est la flexibilité inhérente à l'approche, grâce à la facilité de modification d'une ontologie et des liens avec les sources de données.

---

[2] Berners-Lee, T., Hendler, J. et Lassila, O. (2001). The semantic web. *Scientific American Magazine, 284*(5), 29-37.

Un concept central du web sémantique est le « web de données liées » où ce sont les données elles-mêmes qui sont mises en réseau au moyen de la technologie RDF (Resource Description Framework). À titre d'exemple, l'entreprise française Antidot a enrichi une base de données de 43 720 monuments historiques en France au moyen de six autre sources de données préexistantes : la liste des 3065 gares ferroviaires, celle des 301 stations du métro parisien, les données du code officiel géographique de l'INSEE, une banque de 122 828 photos de monuments, les descriptions des monuments dans Wikipedia (intégrées dans DBpedia dont nous parlerons plus loin) et le service de géolocalisation Yahoo! PlaceFinder qui permet d'afficher les lieux sur un carte géographique. Toutes ces données ont été reliées entre elles par des liens RDF formant un énorme graphe de données liées. Elles peuvent être consultées de façon intégrée dans un portail extrêmement utile pour les touristes et les intervenants de l'industrie touristique.

1.3 Qu'est-ce que le web sémantique?

Prenons un autre exemple. Vous apprenez que votre mère doit suivre une série de traitements de physiothérapie recommandés par son médecin et vous devez l'accompagner. Sur le web sémantique, vous demandez à l'agent informatique qui gère votre emploi du temps de vous organiser un horaire qui tient compte de vos autres engagements. L'agent examine une liste de physiothérapeutes disponibles sur le web. Il sélectionne ceux qui correspondent aux critères de la police d'assurance de votre mère et qui sont situés à moins de 20 kilomètres de sa maison. Puis il recherche les heures de visites compatibles avec votre emploi du temps. En peu de temps, il propose un horaire de traitements pour votre mère.

Pour réaliser cela, les termes des pages ou des documents consultés doivent être reliés à un réseau de relations entre les concepts. De la sorte, quand l'agent consulte la page d'une clinique de physiothérapie, il est capable d'identifier, non seulement les mots « traitement », « physiothérapeute » ou « horaire » par leur syntaxe, mais aussi de préciser que tel physiothérapeute travaille à telle clinique et est disponible les lundis, mercredis et vendredis à certaines périodes de l'année. Autrement dit, il faut que les relations entre les concepts soient disponibles.

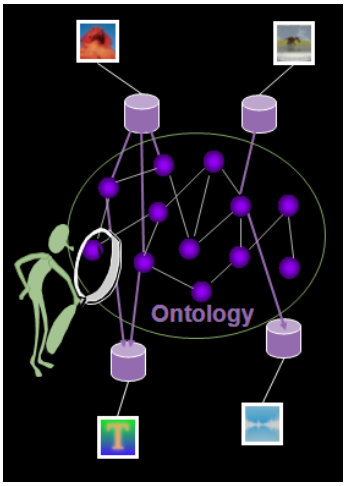

Figure 1 Le web sémantique.

La figure 1 illustre le concept général du web sémantique. Les ressources disponibles sur le web (personnes, organisations, images, textes, pages HTML, etc.) sont décrites par des métadonnées enregistrées dans des bases de données. Ces métadonnées énoncent des propriétés de la ressource.

Par exemple, un document est signé par un auteur, contient un certain nombre de pages, aborde la physiothérapie et ses traitements, possède une table de matière qui pointe sur divers sujets, sur des connaissances traitées dans d'autres livres. Les valeurs de ces propriétés sont choisies dans un réseau de concepts et de relations qui forment un modèle des connaissances, très souvent, une ontologie. En utilisant la structure intégrée dans ce modèle, divers agents informatiques pourront effectuer des inférences et obtenir ainsi des ressources adaptées à une requête.

Le principe essentiel du web sémantique est de soutenir un *web distribué au niveau des données* plutôt qu'au niveau de la présentation des informations, notamment par des liens entre les ressources, entre les pages, entre les éléments de données. Sa réalisation implique un changement d'approche : passer du web des documents (portails, page web, ressources) au web des données liées qui alimentent les pages web, les données étant reliées entre elles au moyen d'un graphe RDF.

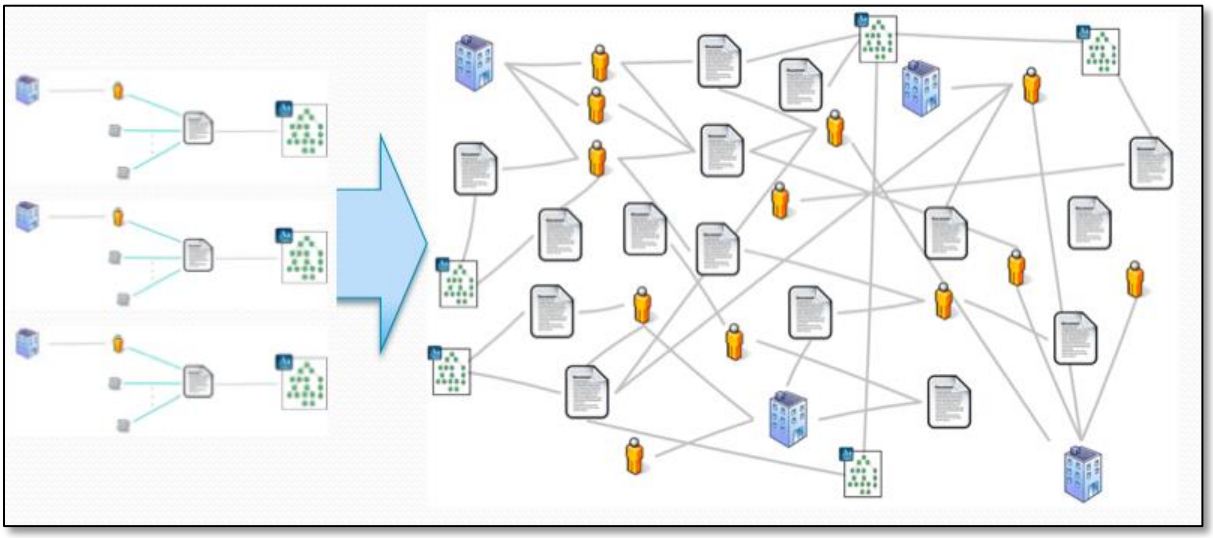

Figure 2 Des données séparées aux données liées.

La figure 2 illustre le concept des données liées. À gauche, on présente plusieurs données sans liens entre elles. Chaque donnée a un ou plusieurs auteurs qui appartiennent à une organisation et un ou plusieurs sujets dans une hiérarchie de termes. Ces données ne communiquent pas entre elles. À droite, on présente un graphe RDF qui les relie. Le graphe relie non seulement les documents, les sujets, mais aussi les personnes et les organisations. Chaque lien du graphe représente une propriété reliant un de ces objets et un autre objet ou une valeur (un nombre, un sujet, une date, etc.) qui appartiennent possiblement à une autre base de données. Cela permet de répondre à des requêtes qui intègrent des termes et des objets qui appartiennent à plusieurs sources de données, par exemple « je cherche des ressources en modélisation, produites par la TÉLUQ et utilisées dans au moins un cours ».

Le web sémantique opère dans un « monde ouvert ». Contrairement aux premiers systèmes d'intelligence artificielle (systèmes experts, analyse de textes, reconnaissance de formes), les systèmes de web sémantique ne sont pas spécialisés utilisant un vocabulaire fixe. Tout le monde peut dire n'importe quoi sur n'importe quel sujet. Le web sémantique englobe une variété de concepts et de terminologies qui évoluent régulièrement, alimentés par un nombre croissant d'intervenants sur le web.

2. Principes du web de données liées[3]

L'idée centrale du web de données liées est d'utiliser la structure du web de documents classique comme base pour structurer les données globalement.

Le web de données liées repose sur un petit nombre de principes additionnels à ceux du web de documents classique :

1. Utiliser des URI pour nommer non seulement les documents, mais les lieux, les personnes, les organisations et d'autres objets qui ne se matérialisent pas comme des documents, pages ou site web, et aussi des concepts abstraits qui les décrivent.
2. Rendre toutes les URI « consultables » via le protocole HTTP pour assurer l'accès sur le web à de l'information sur les entités décrites par ces URI.
3. Lors de l'accès à une URI, donner de l'information sur les objets dénotés par cette URI à l'aide des standards RDF et SPARQL.
4. Lors de l'accès à une URI, prévoir des liens à d'autres URI pour découvrir d'autres objets dans la même base ou dans une autre base de données.

2.1 Les bases du web de données liées

Premier principe : le web de données liées peut être vu comme une extension du web de documents à tout objet ou à tout concept qui inclut des entités concrètes comme les personnes, les organisations ou les concepts abstraits comme celui de véhicules de transport ou des relations entre les entités telles que « connaître quelqu'un » ou « administrer un système ». Les URI fournissent des noms adéquats à ces entités parce qu'elles permettent de créer des identifiants uniques pour les objets à la façon décentralisée propre au web. Ainsi, tout détenteur d'un nom de domaine ou son délégué peut créer de nouvelles références URI.

Deuxième principe : dans le web classique, le protocole HTTP assure le lien entre l'identificateur d'une ressource et son affichage dans une interface utilisateur qui permet sa consultation. L'extension de ce mécanisme au web de données consiste à « déréférencer » toute URI pour rendre visible une description de l'objet ou du concept correspondant.

[3] Cette section est inspirée de Heath, T. et Bizer, C. (2011). *Linked data : Evolving the web into a global data space* (1ère éd.). Morgan & Claypool Publishers, p. 7-27.

Troisième principe : de façon analogue au langage HTML, qui standardise les contenus du web de documents, le web de données utilise un unique langage de publication de contenus sur le web de données : le Resource Description Framework (RDF)[4] qui est un modèle de données sous forme de graphe regroupant des triplets (objet, propriété, valeur).

Quatrième principe : le web de données généralise l'utilisation des hyperliens pour connecter non seulement des documents, mais toutes sortes d'objets, par exemple entre une personne et un lieu, une personne et une compagnie, une personne et une ressource dont il est l'auteur. Contrairement aux liens du web de documents qui sont non typés, dans le web de données les liens ont des types spécifiés par des relations abstraites; par exemple les liens « ami-de » ou « auteur-de ». On les appelle « liens RDF » pour les distinguer des hyperliens classiques (syntaxiques). C'est ainsi que le web devient sémantique, en ce sens que ces concepts et ces relations abstraites constituent un modèle de connaissance d'un domaine donné.

Un exemple de graphe RDF est présenté à la figure 3[5]. Il établit des liens entre trois types de sujets différents : les organisations (p. ex. TÉLUQ), les personnes (p. ex. Gilbert_Paquette, Josianne_Basque) et les cours (p. ex. Cours_Inf6070). Ce graphe contient quatre triplets qui décrivent chacun un lien RDF :

```
(Gilbert_Paquette, travaille-à, TÉLUQ)
(Josianne_Basque, travaille-à, TÉLUQ)
(Gilbert_Paquette, connait, Josianne_Basque)
(Gilbert_Paquette, est-auteur-de, Cours_Inf6070)
```

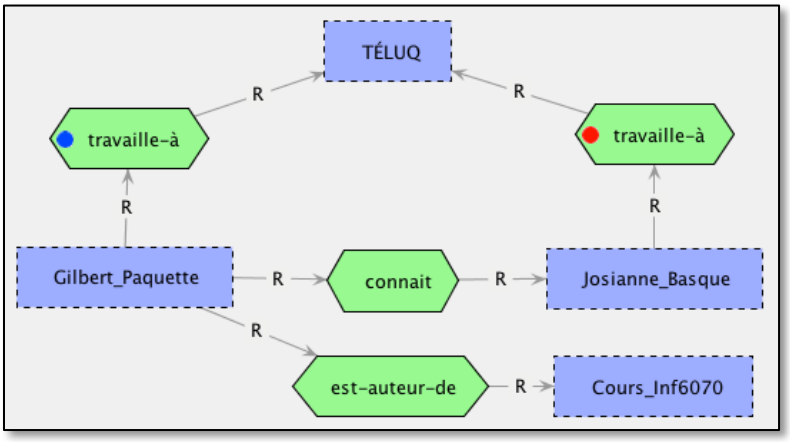

Figure 3 Un graphe RDF contenant quatre triplets.

[4] Graham, K. et Carroll, J. J. (dir.). (2004, février). Resource description framework (RDF) : Concepts and abstract syntax. W3C Recommendation. Récupéré le 10 octobre 2013 de : http://www.w3.org/TR/rdf-concepts/

[5] Ce graphe est réalisé avec le logiciel G-MOT présenté précédemment. Il n'y a ici qu'un type de lien du vocabulaire MOT (le lien R), mais d'autres liens seront nécessires dans d'autres sections du cours. Un lien RDF est représenté par trois objets du vocabulaire MOT, liés ensemble par des liens R. Dans plusieurs autres représentations par graphe, on préfère lier le sujet à l'objet par un lien surmonté d'une URI de la propriété, mais cela rend plus difficile l'interprétation les liens entre les propriétés.

Soulignons que chacun des termes de ces triplets (sujet, relation, objet) sera en fait une URI commençant par `http://`. Ce sont ces URI qui permettent la navigation entre les différentes bases de données comme s'il s'agissait d'un seul espace de données. Par exemple, on peut passer d'un ensemble de données sur des personnes à l'ensemble des cours dont ces personnes sont les auteurs au moyen d'un lien RDF « `est-auteur-de` ». Cela permet de construire des applications ou des agents informatiques qui opèrent sur le web de données global, quelles que soient les technologies sous-jacentes à chaque serveur et les terminologies ou vocabulaires utilisés.

2.2 Le déréférencement des URI

Dans le web de données liées, les URI qui référencent des objets ou des concepts doivent être déréférencées. Via le protocole HTTP, les ordinateurs clients pourront obtenir une description de l'objet, que celui-ci soit un document HTML classique, ou encore un objet ou un concept du web de données. Le déréférencement des URI conduit à des représentations HTML à l'intention des personnes et à des représentations RDF pour le traitement par des machines. Cela se fait par un mécanisme du protocole HTTP appelé *content negociation*. L'idée centrale est que les clients HTTP envoient un code HTTP (des *headers*) avec chaque requête qui indiquent quelle sorte de représentation est requise pour le retour par le serveur, soit du HTML ou du RDF.

Quand les URI identifient des objets du monde réel ou des concepts, il est essentiel de distinguer ces objets et ces concepts des documents web qui les décrivent. On utilise alors des URI différentes, ce qui permet de faire des assertions différentes dans un cas comme dans l'autre; par exemple, la date de création d'un appareil peut être différente de la date de création d'un document le décrivant.

L'exemple suivant montre trois URI à propos d'une même ressource dans DBpedia. Le premier représente la ressource elle-même (un concept), le second la page HTML (dans Wikipedia) de la ressource et la troisième les données RDF concernant cette ressource.

```
http://dbpedia.org/resource/Wildlife_photography
http://dbpedia.org/page/Wildlife_photography
http://dbpedia.org/data/Wildlife_photography
```

Contrairement aux documents du web classique, les objets du monde réel ou les concepts abstraits ne peuvent évidemment circuler eux-mêmes sur le web, mais ils peuvent être représentés dans le web de données. Dans ce cas, on utilise une méthode de déréférencement indirecte. Il existe deux stratégies alternatives pour ce faire, ayant chacune leurs avantages et leurs inconvénients.

La première stratégie consiste, dans un premier temps, à retourner un code « `303 see other` » au lieu de fournir l'objet comme on le ferait si c'était un document. Puis, dans un second temps, on déférence cette URI spéciale pour retourner un document qui décrit l'objet plutôt que l'objet lui-même.

La seconde stratégie évite les deux requêtes inhérentes à la stratégie précédente en utilisant la structure de l'URI composée de la base de l'URI, suivie du symbole # (*hash code*), suivie d'une partie appelée *fragment identifier*. Le protocole HTTP sépare le fragment de la base avant d'interroger le serveur, ce qui empêche le retrait d'un document et indique qu'il peut s'agir d'un objet du monde réel ou d'un concept. On obtient alors une description HTML et une description RDF de l'objet réel ou du concept.

Toutes les URI qui se terminent par # suivi d'un fragment peuvent servir à identifier sans ambiguïté ce type d'objet propre au web de données. Par exemple, les deux personnes présentées à la figure 3 pourront être désignées par les URI suivantes :

```
http://ns.teluq.ca/annuaire#Gilbert_Paquette
http://ns.teluq.ca/annuaire#Josianne_Basque
```

2.3 Le modèle de données RDF et les vocabulaires

Le modèle de données RDF permet de décrire chaque ressource du web par un certain nombre de triplets (sujet, prédicat, objet) où :

- le sujet est l'URI identifiant la ressource;
- l'objet est une valeur, par exemple, un nombre, une chaîne de caractères, une date, ou encore une URI désignant une autre ressource ayant un lien avec le sujet;
- le prédicat (ou relation, ou propriété) est aussi une URI décrivant une relation entre le sujet et l'objet; cette URI est choisie dans un *vocabulaire*, c'est-à-dire une collection d'URI regroupant des concepts et des propriétés abstraites qui permettent de décrire de l'information à propos d'un certain domaine.

Il y a deux types de triplets RDF selon le type de la partie objet.

- Les *triplets de propriétés* (*literal triples*), dont la partie objet est une valeur, décrivent des propriétés de la ressource sujet.

  Voici un exemple de ce type de lien et d'un triplet correspondant qui donne l'adresse courriel de l'auteur de ce texte (référencé dans un vocabulaire tq à propos de la TÉLUQ) au moyen du prédicat #mbox (mail box).

  ```
  (http://ns.teluq.ca/annuaire#Gilbert_Paquette,
   http://xmlns.com/foaf/0.1/#mbox,
   mailto :gilbert.paquette@teluq.ca)
  ```

- Les *liens RDF*, dont la partie objet est l'URI d'une autre ressource, décrivent un lien entre la ressource sujet et cette ressource objet.

  L'exemple suivant indique que l'auteur a un intérêt (prédicat #interest) pour un sujet décrit dans l'URI objet, ici une page sur la représentation des connaissances référencée dans le vocabulaire DBpedia.

  ```
  (http://ns.teluq.ca/annuaire#Gilbert_Paquette,
   http://xmlns.com/foaf/0.1/#interest,
  http://fr.dbpedia.org/page/Representation_des_connaissances)
  ```

Le vocabulaire FOAF (Friend of a Friend) est un vocabulaire particulièrement utile parmi les nombreux disponibles sur le web. Il fournit des concepts et des prédicats pour décrire des personnes. Ce vocabulaire est identifié par l'URI `http://xmlns.com/foaf/0.1/`. En y ajoutant `#mbox` ou `#interest`, on obtient l'URI de ces prédicats du vocabulaire FOAF.

Le modèle de données RDF présente les avantages suivants :

1. En utilisant des URI HTTP comme identifiants uniques pour les données et les termes de vocabulaire, le modèle RDF est intrinsèquement construit pour être utilisé à l'échelle globale du web, permettant à tous de référencer n'importe quoi : documents, objets du monde réel, concepts et prédicats d'un vocabulaire.
2. Les ordinateurs clients peuvent utiliser toutes les URI d'un graphe RDF sur le web pour obtenir de l'information. Chaque triplet faisant partie du web de données liées, un graphe RDF peut servir de point de départ pour explorer cet espace global de données.
3. Le modèle de données RDF permet de tracer des liens RDF entre les données de différentes sources. L'information de ces sources peut être facilement combinée en joignant les ensembles de triplets en un seul graphe. L'exemple ci-dessus permet de joindre dans un même graphe de l'information exprimée dans un vocabulaire comme FOAF pour les personnes et un vocabulaire comme DBpedia pour décrire le contenu de Wikipedia, l'encyclopédie du web.
4. En combinant un modèle de données RDF avec des langages de schémas comme RDFS ou OWL, on peut lui ajouter des structures au niveau de complexité voulue. C'est ce que nous verrons dans les modules 3 et 4 de ce cours.

2.4 Les formats de sérialisation

Il est important de souligner que RDF n'est pas un format de données, mais un modèle de données sous forme de graphe regroupant des triplets (sujet, prédicat, objet). Pour publier des données sur le web, il faut utiliser un format de données. La transformation d'un modèle comme celui de la figure 3 s'appelle la *sérialisation*. Le résultat est un fichier selon un certain format. Il existe plusieurs de ces formats pour le web de données.

On peut énoncer les triplets comme dans l'exemple précédent, avec les URI des trois éléments, ce qui requiert trois lignes de code par triplet. Ce format appelé N3 est très encombrant bien qu'efficace pour enregistrer par ordinateur de longues listes de données. Voilà pourquoi le W3C a standardisé deux autres formats de sérialisation, le RDF/XML et le RDFa.

Un autre format appelé Turtle, encore plus pratique, est aussi en voie de standardisation. Dans ce cours, nous allons l'utiliser très abondamment, ainsi que dans une moindre mesure le format RDF/XML. Le format RDF/XML[6] est standardisé et largement employé pour publier des données liées sur le web. Il est cependant un peu difficile à lire et à écrire par des humains.

Le format RDFa[7] est aussi standardisé. Il est conçu pour imbriquer des triplets RDF directement dans le code des documents HTML. Cela signifie que le contenu d'une page, auquel on ajouté du code RDFa, permet d'exposer des données structurées sur le web. Ce format est utile dans des contextes où l'on veut modifier le gabarit HTML pour capter la structure qui nous intéresse.

Le format Turtle est un format texte qui supporte des préfixes pour les espaces de noms (comme FOAF ou DBpedia) et d'autres abréviations, ce qui en facilite la lecture et l'écriture à la main. Il a été soumis au W3C pour standardisation[8].

Il existe aussi un format RDF/JSON[9] pour la sérialisation utilisant le JavaScript Object Notation. Ce format est utile, car beaucoup de langages de programmation, tels JavaScript et PHP, fournissent un support à JSON. Cela évite aux programmeurs web d'installer des librairies logicielles additionnelles pour traiter les données RDF. On s'attend à ce que des efforts soient entrepris pour standardiser également ce format.

Le choix d'un format de sérialisation dépend évidemment de l'utilisation qu'on veut en faire et du contexte technologique de son utilisation. Il existe par ailleurs des traducteurs entre les différents types de format. Pour notre part, nous construirons les modèles RDF dans GMOT et dans Protégé et nous présenterons leurs sérialisations uniquement dans les formats RDF/XML et Turtle.

2.5 Les types de liens RDF entre les espaces de données

Le quatrième principe énoncé au début de cette section stipule que des liens peuvent pointer sur différentes sources de données sur le web. Ces liens externes sont fondamentaux par leur fonction d'intégration ou de « colle sémantique » entre les différents espaces du web de données. Techniquement un lien externe est un triplet dont le sujet est une URI d'un espace de noms, et dont le prédicat ou l'objet, ou les deux, sont une URI d'un autre espace de noms, liant ainsi les deux ensembles de données (*datasets*).

---

[6] Manola, F. et Miller, E. (dir.). (2004). RDF primer. W3C Recommandation. Récupéré le 10 octobre 2013 de : http://www.w3c.org/TR/rdf-primer/

[7] Adida, B., Birbeck, M., McCarron, S. et Herman, I. (dir.). (2013). RDFa Core 1.1 – Second Edition: Syntax and processing rules for embedding RDF through attributes. W3C recommendation. Récupéré le 10 octobre 2013 de : http://www.w3.org/TR/2013/REC-rdfa-core-20130822/

[8] Beckett, D. et Berners-Lee, T. (2011) Turtle – Terse RDF Triple Language. W3C Team Submission. Récupéré le 10 octobre 2013 de : http://www.w3. org/TeamSubmission/turtle/

[9] Keith, A. (2008). Rdf in json. Dans *Proceedings of the 4th Workshop on Scripting for the Semantic Web.* Tenerife, Espagne. Récupéré le 10 octobre 2013 de : http://ceur-ws.org/Vol-368/paper16.pdf

Le déréférencement des URI fournit une description des ressources liées fournies par les serveurs où elles résident. Ces descriptions fournissent généralement des liens RDF internes ou externes additionnels qui pointent sur de nouvelles URI et ainsi de suite. C'est ainsi que la toile se tisse entre les descriptions des ressources individuelles, encore une fois, que celles-ci soient des documents web, des objets du monde réel ou des concepts abstraits. De la sorte, le web de données liées peut être exploré en utilisant un navigateur de données liées ou en lançant des recherches à l'aide d'un moteur de recherche.

Il y a trois sortes de liens RDF externes qu'il importe de distinguer.

- Les *liens relationnels* à partir d'un sujet pointent vers des objets dans d'autres sources de données, par exemple d'autres personnes, le lieu où vit une personne, sa date de naissance, ses publications, ses centres d'intérêt. Ces liens permettent, de proche en proche, de lier une source de données à une deuxième, puis de là à une troisième, établissant des connexions dans un réseau potentiellement illimité de données qui peuvent être traitées par les applications.
- Les *liens d'identification* pointent sur des URI utilisées par d'autres sources de données, mais qui identifient le même objet du monde réel ou le même concept abstrait. Cela permet d'obtenir des descriptions additionnelles de l'entité en provenance d'autres sources de données, notamment des descriptions dans d'autres langues. Ces liens ont une importante fonction sociale, car ils permettent différentes visions et différentes descriptions d'un même objet. Cela se fait à l'initiative des utilisateurs plutôt que par une agence centralisée qui nierait la dynamique distribuée du web. Les URI qui décrivent un même objet sont appelées « alias ». Par convention, on utilise le prédicat `http://www.w3.org/2002/07/owl#sameAs` pour lier une URI à un de ses alias.
- Les *liens de vocabulaire* pointent sur les définitions des concepts et des prédicats d'un vocabulaire ou encore de ces définitions vers d'autres définitions de termes reliées dans d'autres vocabulaires. Les liens de vocabulaire servent aussi à décrire les propriétés des données (les métadonnées), permettant aussi aux applications de comprendre et d'intégrer les données de plusieurs vocabulaires.

Le web de données liées favorise une double approche pour traiter les représentations hétérogènes des données. D'un côté, on recommande l'adoption et la réutilisation des termes des vocabulaires les plus répandus comme DC (Dublin Core), FOAF, DBpedia ou Geobase.

D'un autre côté, on recommande de faire en sorte d'auto-décrire les ensembles de données, c'est-à-dire d'y insérer des méta-informations à propos de l'ensemble de données telles que ses créateurs, la date de création, la licence permettant d'utiliser les données. Sur ce certain point, un vocabulaire RDF sert à décrire les licences Creative Commons[10] qui sont très répandues.

[10] Site web : http://creativecommons.org/

Plus généralement, on utilise de façon standard le voID (Vocabulary of Interlinked Datasets)[11], un vocabulaire spécialisé pour décrire les vocabulaires utilisés dans un ensemble de données ainsi que ses sous-ensembles, son sujet, le créateur des métadonnées, la date de création, etc. En intégrant ces informations dans l'ensemble de données lui-même, une application qui trouve un vocabulaire inconnu sur le web pourra trouver toutes les méta-informations requises pour traduire les données sous une forme qu'il peut traiter.

Si, plus tard, un éditeur de données découvre un autre vocabulaire standard qui contient des termes équivalents, il devrait ajouter un lien RDF de type #sameAs ou encore utiliser des triplets qui mettent en relation les termes à l'aide des méta-vocabulaires tels que le RDFS[12] (RDF Schemas), SKOS[13] (Simple Knowledge Organization System) ou OWL[14] (Ontology Web Language). Nous étudierons ces systèmes dans les modules 3 et 4 de ce cours.

3. Topologie et architecture du web de données

Nous avons souligné plus haut que le web de données est une couche additionnelle imbriquée dans le web de documents classique. Sa construction a commencé en janvier 2007 par l'identification des premiers ensembles de données disponibles sous des licences ouvertes, puis par leur conversion en RDF et leur diffusion sur le web.

3.1 Topologie du web de données liées

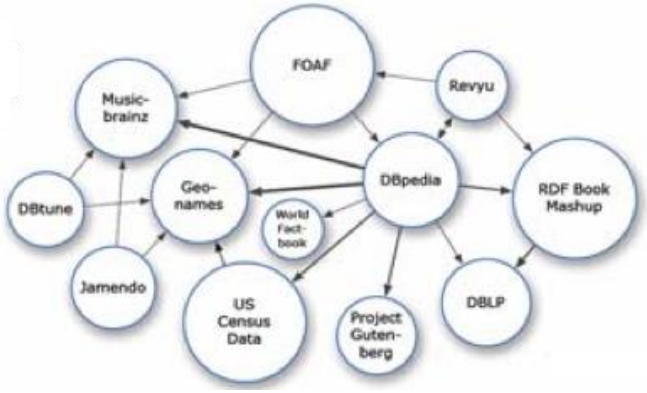

Figure 4 Le web de données en mai 2007.

---

[11] Alexander, K., Cyganiak, R., Hausenblas, M. et Zhao, J. (2010). Describing linked datasets with the VoID vocabulary. Editor's draft. W3C. Récupéré le 10 octobre 2013 de :

http://www.w3.org/2001/sw/interest/void/

[12] D. Brickley D. et Guha, R. V. (dir.). (2004, février). RDF vocabulary description language 1.0: RDF schema. W3C recommendation. Récupéré le 10 octobre 2013 de : http://www.w3.org/TR/rdf-schema/

[13] Miles, A. et Bechhofer, S. (dir.). (2009, août). SKOS simple knowledge organization system. Reference. Récupéré de : http://www.w3.org/TR/skos-reference/

[14] McGuinness, D. L. et van Harmelen, F. (dir.). (2004). OWL web ontology language overview. W3C recommendation. http://www.w3.org/TR/2004/REC-owl-features-20040210/

Cette possibilité étant ouverte à tous (d'où le terme *open data*) le nombre d'ensembles de données et de vocabulaires disponibles a crû très rapidement. La figure 4 indique qu'en mai 2007, on ne comptait que 12 ensemble de données, notamment DBpedia, FOAF et Geonames. À la fin de 2010, le web de données regroupait 203 ensembles de données regroupant près de 27 milliards de triplets et près de 400 millions de lien RDF, comme il est indiqué sur la figure 5.

Dans le graphe de la figure 5[15], chaque nœud représente un ensemble de données et les liens entre eux signifient qu'un ensemble de données utilise en partie le vocabulaire d'un autre nœud. Ainsi, un ensemble de données comme DBpedia est l'objet d'un très grand nombre de liens, car son vocabulaire est souvent réutilisé. La description de chaque ensemble de données peut être consultée dans le site CKAN [16], , et également dans le site Linked Open Vocabularies (LOV)[17] qui fournit une animation graphique et une variété de liens entre les vocabulaires ouverts liés.

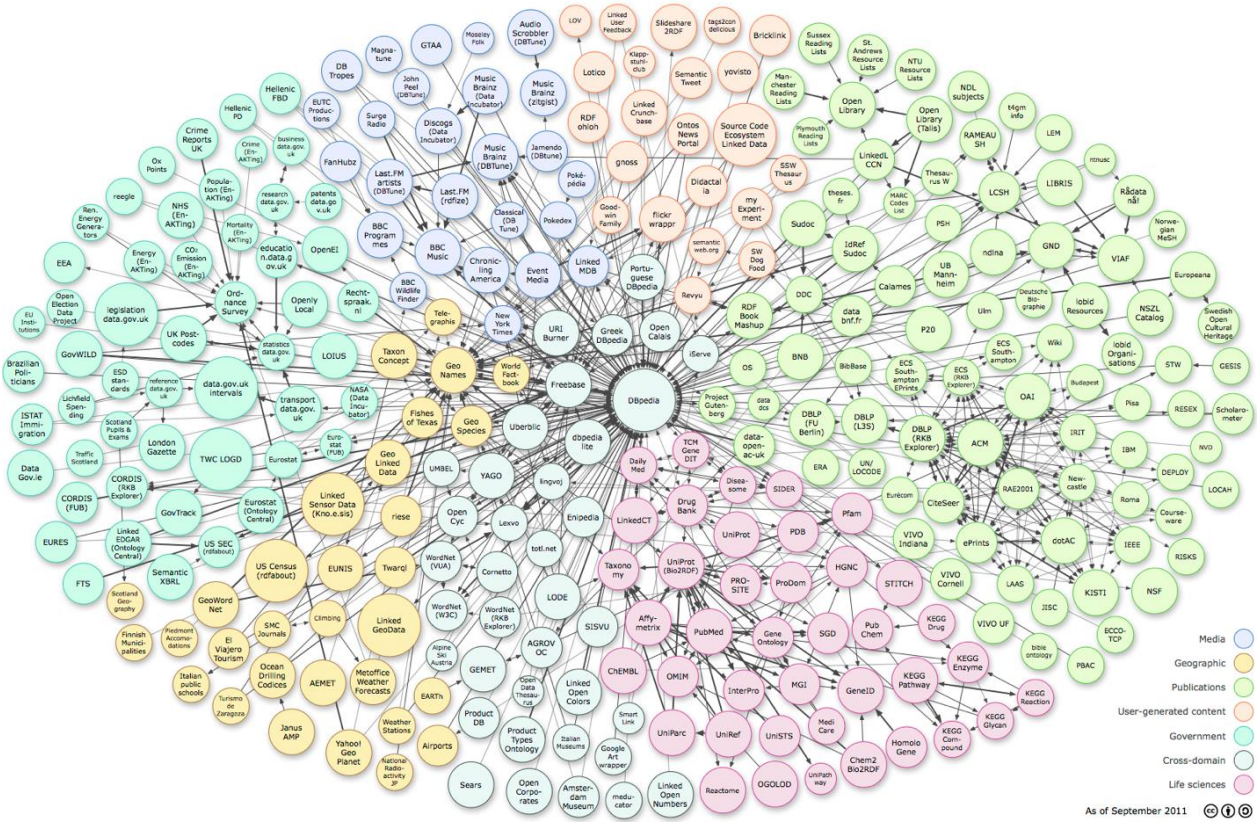

Figure 5 Le web de données en mai 2010.

[15] Disponible en divers formats à l'adresse suivante : http://lod-cloud.net

[16] Site CKAN : http://www.ckan.net/group/lodcloud

[17] Site web : http://lov.okfn.org/dataset/lov/

Les ensembles de données dans CKAN sont classés par catégories comme il est indiqué dans le tableau 1 qui donne pour chacune le nombre de noeuds, de triplets et de liens RDF.

Tableau 1 Statistiques sur le web de données 2010

Domaine	Ensembles de données (Nb de nœuds)	Nb de triples	%	Nb de liens RDF	0,02
Inter-domaines	**20**	1,999,085,950	7.42	29,105,638	7.36
Géographie	**16**	5,904,980,833	21.93	16,589,086	4.19
Gouvernement	**25**	11,613,525,437	43.12	17,658,869	4.46
Media	**26**	2,453,898,811	9.11	50,374,304	**12.74**
Bibliothèques	**67**	2,237,435,732	8.31	77,951,898	**19,71**
Sciences de la vie	**42**	2,664,119,184	9.89	200,417,873	**50.67**
Données des usagers	**7**	57,463,756	0.21	3,402,228	**0.86**
TOTAL	203	26,930,509,703		395,499,896	

Comme on peut le voir sur la figure 5, DBpedia est un ensemble de données interdomaines particulièrement important parce qu'il établit un grand nombre de liens avec d'autres ensembles de données. Les données RDF de DBpedia sont extraites automatiquement des pages de Wikipedia, l'encyclopédie du web, par une application qui crée automatiquement une URI de DBpedia fondée sur l'URI de l'article dans Wikipedia.

Par exemple, l'article sur la ville de Montréal ayant pour URI `http://fr.wikipedia.org/wiki/Montréal` (voir figure 6) sera transformé en l'URI `http://dbpedia.org/resource/Montréal` qui désigne non pas la page, mais la ville elle-même dans l'ensemble de données DBpedia. Les triplets RDF qui décrivent les propriétés de cette ville seront générés en extrayant de l'information des différentes sections de l'article correspondant dans Wikipedia, en particulier des boîtes d'informations que l'on retrouve dans la plupart des articles.

DBpedia n'est évidemment pas le seul ensemble de données qui permet de lier l'information de divers domaines plus spécialisés. Par exemple, l'ensemble de données Geonames publie un ensemble de données liées de plus de huit millions de lieux. L'ensemble LinkedGeoData fournit de l'information sur 350 millions de données spatiales. En liant ces ensemble de données et d'autres, notamment les données sur les personnes dans FOAF, on obtient déjà un noyau important du web de données liées. D'autres ensembles de données sur la musique, les programmes de radio et de télévision, les journaux, des archives et des données gouvernementales, des bibliothèques telles que la Librairie du Congrès, ou encore les données de e-commerce ou provenant des médias sociaux, procurent une quantité impressionnante d'informations sur le web de données liées, en format RDF.

Figure 6 Une source de DBpedia.

3.2 Architecture du web de données liées

Ces données sont rendues disponibles sur le web par une variété d'outils et de méthodes de publication. On peut décrire l'architecture du web de données selon les outils utilisés dans les applications ainsi que les protocoles et les langages.

La figure 7 présente les principales composantes d'une architecture-cadre pour les applications du web de données liées[18]. Les éléments tels que l'infrastructure, le protocole HTTP, ainsi que les outils de sécurité et les interfaces utilisateurs sont les mêmes que pour toute application web. Les composants spécifiques du web de données (en bleu sur la figure) visent à permettre la publication des données en format RDF et leur utilisation.

[18] Graphique adapté de : Domingue, J., Fensel, D. et Henler, J. A. (dir.). (2011). *Handbook of semantic web technologies*. Berlin : Springer Berlin – Heidelberg, p. 52.

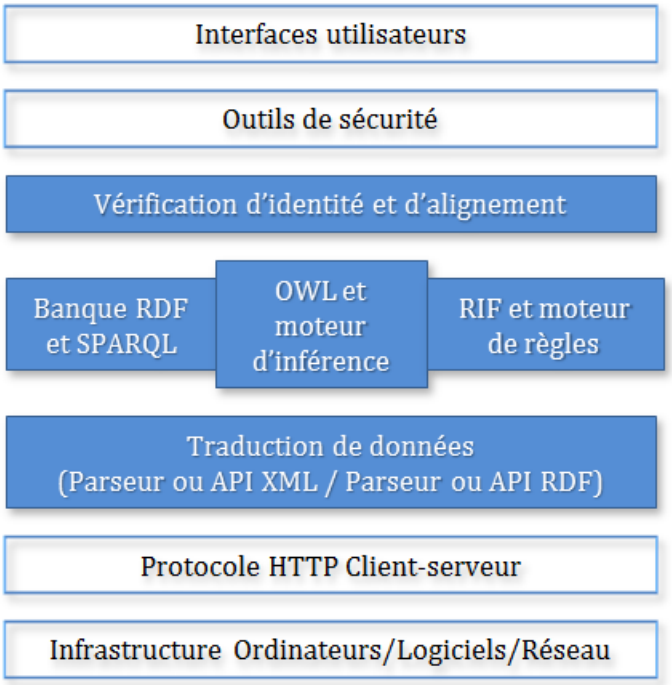

Figure 7 Architecture du web sémantique.

Le contenu retourné par le protocole HTTP doit d'abord être déréférencé au moyen de parseurs XML ou des API XML dans le cas de données non-RDF ou par des parseurs RDF ou des API RDF. Par la suite, les triplets RDF obtenus sont enregistrés dans une banque de triplets RDF (*triple store*). SPARQL, un langage d'interrogation analogue à SQL permet de traiter le contenu d'une banque RDF pour en extraire ou ajouter de l'information. Les langages d'ontologie comme RDFS ou OWL permettent de structurer davantage une banque RDF et de déduire des données nouvelles à l'aide d'un moteur d'inférence. RIF, un format standard pour exprimer des règles, et un moteur de traitement des règles ajoutent des fonctionnalités de traitement encore plus étendues. Comme les données proviennent de plusieurs sources, des outils de vérification d'identifié et d'alignement sont nécessaires.

3.3 Méthodes de publication

La figure 8 présente de façon intégrée les méthodes de publication les plus répandues d'informations sur le web de données liées, à partir de données structurées dans d'autres formats ou de données textuelles. Cela peut se faire à l'aide d'une variété d'outils et des méthodes de stockage qui dépendent du format originel.

Les informations stockées dans des bases de données relationnelles peuvent être publiées relativement facilement sans modifier le lieu et le format de stockage. On peut le faire au moyen d'enveloppes (*wrappers*) qui consistent à définir une correspondance entre BD relationnelles et graphes RDF. Certains systèmes de gestion de contenu web (CMS) comme Drupal intègrent cette fonctionnalité en option.

Lorsque des données structurées existent derrière une API propriétaire, comme Flickr, Amazon ou autre, la situation est un peu plus compliquée. Une enveloppe sur mesure doit être construite selon les spécificités de l'API. Cependant, il est possible de modulariser les enveloppes pour réutiliser les composants d'un ensemble de données à l'autre.

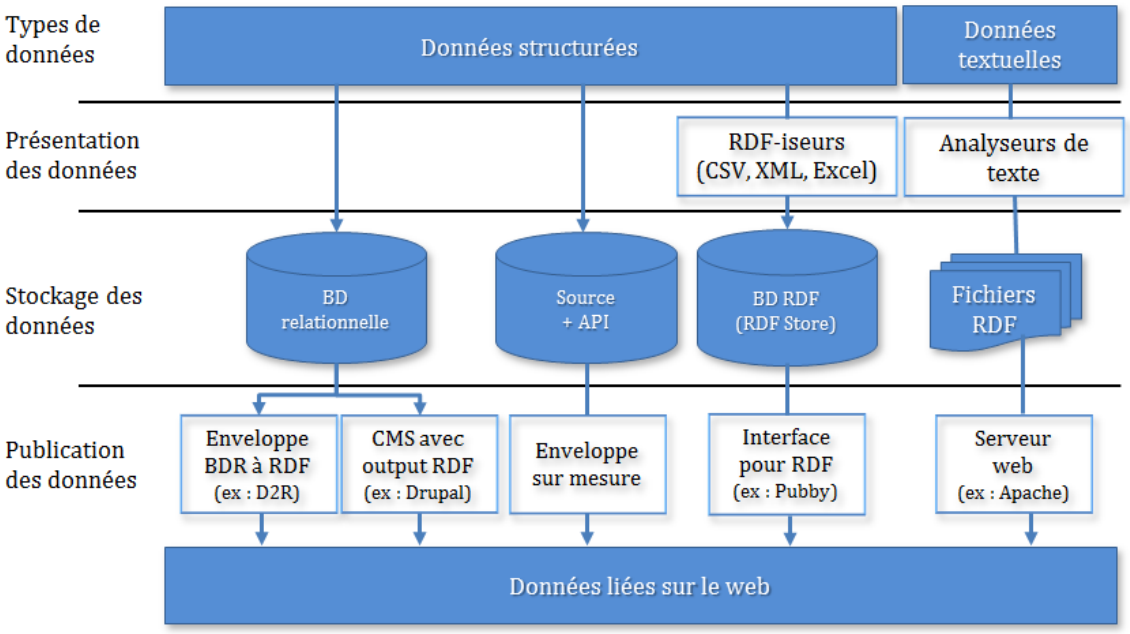

Figure 8 Publication de données sur le web de données liées.

Outre ces cas de données dynamiques, l'information structurée peut aussi résider dans des fichiers statiques CSV, des tableurs Excel, des fichiers XML ou de simples fichiers obtenus de bases de données. Dans ce cas, il faudra d'abord préparer les données en les convertissant en RDF pour leur stockage dans une banque RDF (*triple store*) ou pour leur diffusion dans des fichiers RDF. Il existe plusieurs parseurs ou RDF-iseurs pouvant être utilisés à cette fin[19].

Le cas le plus difficile est celui des données non structurées contenues dans des documents textuels en langue naturelle, par exemple des rapports dans des fichiers texte, des articles de journaux imprimés ou des pages HTML. Dans ce cas, il faudra prétraiter les documents à l'aide d'outils d'extraction de concepts abstraits représentés par leur URI. Des outils tels que Calais, Ontos ou DBpedia Spotlight[20] permettent d'utiliser ces URI pour annoter les documents. Ces annotations, publiées sur le web avec les documents permettent aux applications d'améliorer les fonctions de navigation ou de recherche d'informations.

[19] Voir http://esw.w3.org/ConverterToRdf ou http://simile.mit.edu/wiki/RDFizers

[20] Sites web : http://www.opencalais.com/, http://www.ontos.com/o_eng/index.php?cs=1, et https://github.com/dbpedia-spotlight/dbpedia-spotlight

4. Application du web de données liées

En terminant, nous allons présenter quelques applications qui opèrent sur le web de données liées. On peut les subdiviser en deux catégories : les applications génériques qui consomment des données potentiellement dans tous les domaines et les applications spécifiques d'un domaine. Parmi les applications génériques, on retrouve les navigateurs sémantiques et les moteurs de recherche sémantiques.

4.1 Les navigateurs sémantiques

Tout comme les fureteurs (Firefox, Safari, etc.) permettent aux utilisateurs de naviguer entre les pages HTML en suivant les hyperliens, les *navigateurs sémantiques* permettent la navigation entre les sources de données en suivant les liens RDF.

Par exemple, un utilisateur peut voir la description RDF d'une ville comme Montréal dans DBpedia, puis, de là, en suivant le lien vers la cofondatrice, Jeanne-Mance, obtenir la biographie du personnage, son lieu de naissance, la région de France, etc. Ainsi, l'utilisateur peut commencer son exploration dans un ensemble de données et progressivement traverser le web en suivant les liens RDF plutôt que les liens HTML.

Il existe plusieurs navigateurs sémantiques tels que Disco, Tabulator ou Marbles[21].

Marbles, par exemple, rassemble des informations de différentes sources à propos d'un même objet lorsqu'on spécifie son URI. Marbles présente ces informations sous la forme d'une table qui indique leur provenance au moyen de billes de différentes couleurs. À partir de là, on peut sélectionner une des propriétés de la ressource et obtenir toutes les informations la concernant. De proche en proche, on peut ainsi « traverser » le web de données.

La figure 10 présente une partie du résultat d'une recherche à partir de Marbles pour l'URI http://dbpedia.org/resource/Montréal représentant la ville de Montréal dans DBpedia. On y a rassemblé toutes les données disponibles concernant cet objet, qui proviennent de différentes sources. À partir de là on peut naviguer sur d'autres ensembles de données.

[21] Site web : http://marbles.sourceforge.net/

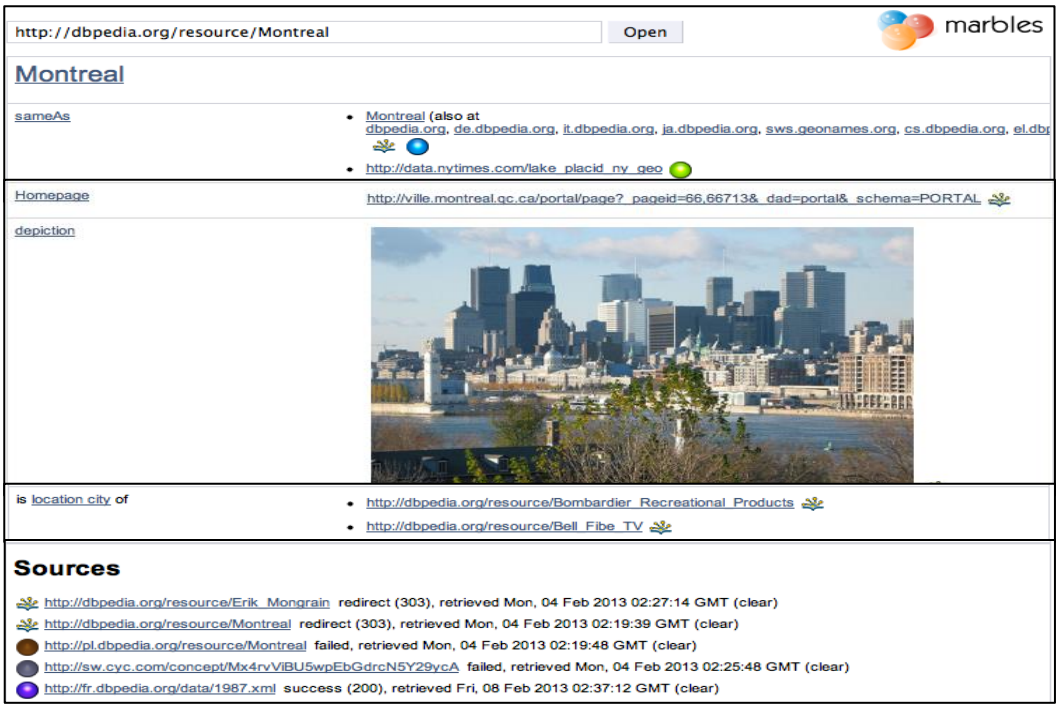

Figure 10 Une vue d'une sortie Marbles pour l'URI de la Ville de Montréal.

4.2 Les moteurs de recherche sémantique

Un certain nombre de moteurs de recherche ont été développés pour interroger le web de données liées à partir des liens RDF. Ils offrent ainsi des fonctionnalités de recherche qui intègrent des données de plusieurs milliers de sources de données. Ces outils démontrent les avantages d'une architecture ouverte liant les sources de données.

Des outils de recherche sémantique tels que Sig.ma, Falcons, SWSE10 et VisiNav offrent une interface utilisateur par mots clefs semblable à celle de Google ou de Yahoo et retournent une liste de résultats. Là s'arrête la similitude puisque ces outils offrent une capacité plus ciblée en exploitant la structure des données, leur sémantique, plutôt que simplement l'occurrence des mots clefs dans les documents.

La figure 11 montre l'utilisation d'un des gabarits de présentation offert par Sig.ma[22] à propos d'une personne, Richard Cyganiak, regroupant des données de 20 sources différentes. Un autre élément intéressant de Sig.ma est son approche pour la qualité des données, laquelle offre aux utilisateurs de choisir les sources de données, ce qui créée automatique des cotes d'utilisation. Un autre moteur de recherche sémantique, VisiNav[23], se concentre sur la réponse à des requêtes complexes, beaucoup plus expressives que celles auxquelles Google et Yahoo peuvent actuellement répondre.

[22] Site web : http://sig.ma/

[23] Site web : http://sw.deri.org/2009/01/visinav/

Figure 11 Une vue des résultats de recherche sémantique dans l'outil Sig.ma.

À titre d'exemple, VisiNav affiche une liste de 54 URL valides à la question « donnes-moi les URI de tous les blogues écrits par des personnes que Tim Berners-Lee connaît ». À une requête comme celle-ci, Google ou Yahoo ne retourne qu'une masse de liens décrivant Tim Berners-Lee.

Il est intéressant de noter que les moteurs de recherche traditionnels tels que Google et Yahoo ont commencé à utiliser les données structurées du web au sein de leurs applications. Google, par exemple, utilise des données RDFa ou d'autres microformats pour décrire des personnes, des produits, des entreprises, des organisations, des rapports ou des événements, ce qui lui permet de fournir des résultats de recherche plus riches et mieux structurés. Par exemple, une recherche sur la date de naissance d'un acteur fournira la date plutôt qu'une liste interminable de pages dans laquelle l'utilisateur doit se débrouiller pour trouver ce qu'il cherche.

4.3 Applications spécifiques

Nous présentons maintenant quelques applications du web de données dans des domaines particuliers intéressant une communauté d'utilisateurs.

Dans les sites gouvernementaux comme dat.gov (USA) et data.gov.uk, on retrouve des regroupements (*mashups*) de données qui proviennent d'un appareil gouvernemental. Par exemple, le US Global Foreign Aid Mash up regroupe des données sur les dépenses du United States Agency for International Development (USAID), du département de

l'Agriculture et du département d'État, en les intégrant avec de l'information générale à propos des pays provenant du CIA World Factbook et des articles de journaux du *New York Times*. Ces données, affichées ensemble dans un portail, fournissent à tout utilisateur une vue d'ensemble de la localisation de l'aide américaine et de son évolution dans le temps.

Des applications comme Talis Aspire[24] ou COMÈTE aident les éducateurs à créer et à gérer des listes de ressources (livres, articles, pages web). Talis Aspire est utilisé par quelques milliers d'étudiants à partir d'une interface web qui leur permet d'intégrer, de gérer et d'accéder aux ressources. Les données qu'ils produisent sont stockées en format RDF, ce qui permet de relier des données provenant de différentes sources.

Figure 12 Une recherche sur le web de données dans COMÈTE.

[24] Site web : http://www.w3.org/2001/sw/sweo/public/UseCases/Talis/

COMÈTE[25] est un système développé au Centre de recherche LICEF de la TÉLUQ. L'outil permet de moissonner des banques de ressources éducatives libres (REL), qui sont développées dans différents pays et diffusées par des organismes tels que ARIADNE, European SchoolNet ou MERLOT. Chaque banque documente les ressources selon des normes, des standards et des spécifications variées, principalement le Dublin Core (DC) et le Learning Object Metadata (LOM). COMÈTE moissonne ces métadonnées et les stocke en format RDF. Divers types de recherche sur le web de données et d'affichage des données sont disponibles.

La figure 12 montre le résultat d'une recherche qui a permis de trouver 370 ressources, avec pour chacune une description de la ressource, un accès à la ressource (ici une vidéo) et diverses informations telles que l'auteur et son organisation. Divers icônes donnent accès d'autres ressources du même auteur ou de son organisation.

DBpedia Mobile[26] est une application mobile qui aide les touristes à explorer une ville. L'application s'installe sur un iPhone ou un autre téléphone intelligent. Elle repère la position GPS de l'utilisateur et s'en sert comme point de départ pour trouver des lieux dans DBpedia proche du lieu de l'utilisateur. Celui-ci peut alors naviguer à l'aide des liens RDF vers d'autres sources de données. DBpedia Mobile offre aussi à l'utilisateur de publier sa localisation actuelle, des photos et des commentaires sur le web, les rendant disponibles à d'autres applications, enrichissant ainsi le web de données liées.

Dans le domaine des Sciences de la vie, NCBO Resource Index repose sur des connaissances de plus de 200 ontologies de domaine public permettant aux utilisateurs d'explorer des ressources biomédicales. Un autre exemple, est le Diseasome Map[27], une application qui combine les données de sources variées pour produire un « réseau de désordres et de maladies génétiques » qui met en évidence des associations connues entre désordre et gènes, illustrant l'origine génétique commune de plusieurs maladies.

Une application qui regroupe les données des profiles FOAF permet de produire une carte interactive des chercheurs allemands. Une autre application permet de prévenir les ambiguïtés sur le web social en suggérant des identifiants avec l'aide des données de DBpedia et de Freebase et en organisant les données par sujets et par différents langages.

Le Semantic MediaWiki[28] est une extension du logiciel MediaWiki, utilisé par Wikipedia, qui permet d'ajouter des annotations sémantiques aux pages d'un wiki. Les annotations qui ont été ajoutées peuvent ensuite être utilisées pour réaliser des recherches sémantiques, pour agréger des pages entre elles, structurer leur contenu de différentes manières, par exemple, sur un plan géographique, un calendrier, un graphe, ou exporter ces données pour qu'elles puissent être consommées par des applications tierces.

---

[25] Présentation du projet : http://brer.licef.ca

[26] Site web : http://wiki.dbpedia.org/DBpediaMobile

[27] Site web : http://diseasome.eu/map.html

[28] Site web : http://semantic-mediawiki.org/wiki/Semantic_MediaWiki_-_Page_d%27accueil

Conclusion

Les concepts présentés ici ont été adoptés par un nombre significatif d'éditeurs de données qui, ensemble, ont construit un web de données liées d'une taille impressionnante, démontrant ainsi la faisabilité de l'approche web sémantique et la maturité croissante des outils qui le supportent.

Les entreprises font face à d'énormes défis quant à la gestion de leurs connaissances. Elles maintiennent parfois des centaines de bases de données, qui évoluent séparément selon les départements ou les filiales, ce qui les prive d'une vue d'ensemble sur les connaissances qu'elles détiennent. Maintenant que les technologies web classiques sont devenues incontournables, le web de données a le potentiel de résoudre ces multiples problèmes d'intégration, avec relativement peu d'effort, sans mettre de côté les banques existantes, mais en amenant les données sous un format RDF, ouvrant la voie à de nouvelles applications. De plus, l'évolution des entreprises du web 2.0, telles que Google ou Facebook, vers le web de données permettra l'intégration des données du web social à celles du web sémantique, du web de données ouvrant de fascinantes nouvelles perspectives.