Documentation et bibliothèques

Les bibliothèques numériques sont-elles solubles dans le Web sémantique ?
Are Digital Libraries the Solution for the Semantic Web ?
Las bibliotecas digitales, ¿son asequibles en la Web semántica ?

Michel Gagnon

Volume 59, Number 3, July–September 2013

Bibliothèques numériques

URI: https://id.erudit.org/iderudit/1018846ar
DOI: https://doi.org/10.7202/1018846ar

See table of contents

Publisher(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (print)
2291-8949 (digital)

Explore this journal

Cite this article

Gagnon, M. (2013). Les bibliothèques numériques sont-elles solubles dans le Web sémantique ? *Documentation et bibliothèques*, *59*(3), 161–168. https://doi.org/10.7202/1018846ar

Article abstract

The semantic Web, especially the Linked Open Data network, will radically transform the way in which digital libraries make their documents available and the metadata associated with them. Following a brief description of the principal technologies used by the semantic Web, specifically the RDF data model, the author illustrates how digital libraries can be integrated into the Linked Open Data network. He then examines the crucial problem of extracting the metadata from the textual contents in order to undertake this integration.

DOCUMENTATION BIBLIOTHÈQUES

Les bibliothèques numériques sont-elles solubles dans le Web sémantique ?

Michel Gagnon
Professeur agrégé
École Polytechnique de Montréal
michel.gagnon@polymtl.ca

Résumé | Abstracts | Resume

Le Web sémantique, et plus particulièrement le réseau Linked Open Data, est appelé à transformer radicalement la manière dont les bibliothèques numériques rendent accessibles leurs documents et les métadonnées sur ceux-ci. Dans cet article, après une brève présentation des principales technologies du Web sémantique, et plus particulièrement du modèle de données RDF sur lequel il repose, nous montrons comment les bibliothèques numériques peuvent s'intégrer au réseau Linked Open Data. Puis nous abordons le problème crucial de l'extraction, à partir de contenus textuels, des métadonnées nécessaires à cette intégration.

Are Digital Libraries the Solution for the Semantic Web ?

The semantic Web, especially the Linked Open Data network, will radically transform the way in which digital libraries make their documents available and the metadata associated with them. Following a brief description of the principal technologies used by the semantic Web, specifically the RDF data model, the author illustrates how digital libraries can be integrated into the Linked Open Data network. He then examines the crucial problem of extracting the metadata from the textual contents in order to undertake this integration.

Las bibliotecas digitales, ¿son asequibles en la Web semántica ?

La Web semántica y, en especial, la red Linked Open Data, tiene como objetivo transformar de forma radical la manera en que las bibliotecas digitales permiten el acceso a sus documentos y a los metadatos que se encuentran en estos documentos. En este artículo, realizamos una breve presentación de las principales tecnologías de la Web semántica, y especialmente del modelo de datos RDF, en el cual se basa. Luego, explicaremos de qué forma las bibliotecas digitales pueden integrarse en la red Linked Open Data. Finalmente, abordaremos el problema más significativo de la extracción, a partir de contenidos textuales, de los metadatos necesarios para esta integración.

Introduction

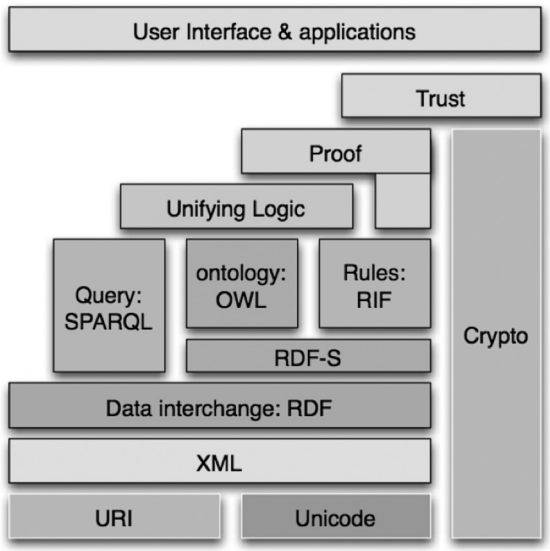
Les bibliothèques numériques, par la grande quantité de documents qu'elles contiennent et leur accessibilité à un très grand nombre d'usagers, posent de grands défis technologiques : traitement des contenus, divulgation, stockage, accès aux documents, etc. Étant donné leur importance grandissante, il est pertinent de se demander dans quelle mesure elles peuvent profiter des technologies du Web sémantique. Dans cet article, nous tentons d'apporter une réponse à cette question.

Le Web Sémantique est un projet dont l'idée a été lancée en 2001 (Berners-Lee *et al.* 2001) et qui a pour but de créer un Web de données qui se superposera au Web actuel. Il s'agit en quelque sorte d'une projection du Web actuel, conçu pour une navigation par l'humain, dans une représentation plus formelle, manipulable directement par la machine, et dans laquelle la sémantique des informations contenues est clairement établie, facilitant ainsi l'interopérabilité entre les différents services offerts par le biais du Web.

Il est tout naturel de se demander dans quelle mesure les technologies du Web sémantique peuvent être utiles au développement des bibliothèques numériques. Comment l'accès aux bibliothèques numériques peut-il en bénéficier ? Quel sera leur apport à l'intégration des bibliothèques numériques provenant de sources diverses ? On peut imaginer plusieurs scénarios où le Web sémantique vient à la rescousse des bibliothèques numériques : recherche de document basée sur des concepts plutôt que sur des mots-clés, personnalisation prenant en compte le profil de l'usager, sites offrant de manière transparente l'accès à des documents provenant de diverses bibliothèques numériques, classification automatique des documents selon une ontologie spécifique[1]. À terme, on peut espérer voir se généraliser les « bibliothèques numériques sémantiques » (Kruk *et al.* 2009).

1. Une ontologie représente de manière formelle et non ambiguë les concepts d'un domaine et leurs relations.

Figure 1
Technologies du Web sémantique (Bratt 2006)



Le Web sémantique

Le Web sémantique est un ensemble de technologies visant à rendre le contenu des ressources du Web utilisable par la machine, grâce à un système de métadonnées formelles. Il s'agit en quelque sorte d'une couche de données interreliées qui s'ajoute au Web original. Pour établir la liaison entre les données, on fait appel à des vocabulaires partagés qui établissent de manière non ambiguë les concepts pertinents (voir Antoniou *et al.* 2012, pour un texte introductif).

Comme l'illustre la figure 1, le Web sémantique est développé en différentes couches. On remarque que toutes les technologies reposent sur XML et les URI (Uniform Resource Identifier). Un URI est un identificateur unique pour chaque entité (aussi appelée *ressource*) que l'on veut décrire sur le Web de données. Un URI s'apparente à un URL, mais n'est pas nécessairement accessible sur le Web.

La couche suivante, RDF (Resource Description Framework) (Manola *et al.* 2004), est un modèle de données qui permet de représenter les relations entre les différentes entités décrites dans le Web de données, sous forme de triplets $\{S, P, O\}$, où S (le sujet) et O (l'objet) sont deux entités et P une propriété liant S à O. Par exemple, on pourra représenter en RDF l'information spécifiant que Michel Gagnon travaille à l'École Polytechnique, où le sujet et l'objet correspondent à *Michel Gagnon* et École Polytechnique, respectivement. Ces deux entités seraient reliées par le prédicat *travailler*. S'ajoute à cela la couche RDF-S (RDF Schema), qui permet d'ajouter une hiérarchie de types. Pour accéder aux données représentées en RDF, on peut utiliser le très puissant langage de requête SPARQL (Prud'hommeaux *et al.* 2008), qui permet d'extraire l'information, de combiner des données provenant de sources hétérogènes et de construire de nouveaux graphes RDF à partir de l'information obtenue.

RDF est très utile pour représenter des relations entre des entités, mais il est limité à plusieurs égards. Notamment, il ne permet pas d'exprimer la négation, ni de définir des concepts de manière précise. Par exemple, il ne permet pas de définir une *mère* comme étant une *femme* qui a au moins un *enfant*. C'est pour cette raison que le langage OWL (Web Ontology Language) a été développé (Hitzler *et al.* 2011). Il permet d'exprimer formellement les restrictions qui définissent un concept. Il augmente ainsi les capacités d'inférence, principalement pour déterminer la classe à laquelle appartient une entité, ainsi que les propriétés qui en découlent. Mais comme la classification n'est qu'une forme de raisonnement, il faut en complément un langage de règles. Par exemple, OWL n'est pas suffisant pour exprimer le fait que le copyright d'un document échoit 50 ans après le décès de l'auteur. Ceci ne peut être exprimé que par une règle de la forme suivante :

SI

document(x) **et** auteur(x,a) **et** décès(a,d) **et** dateCourante(c) **et** $c - d > 50$

ALORS

libreDeDroits(x)

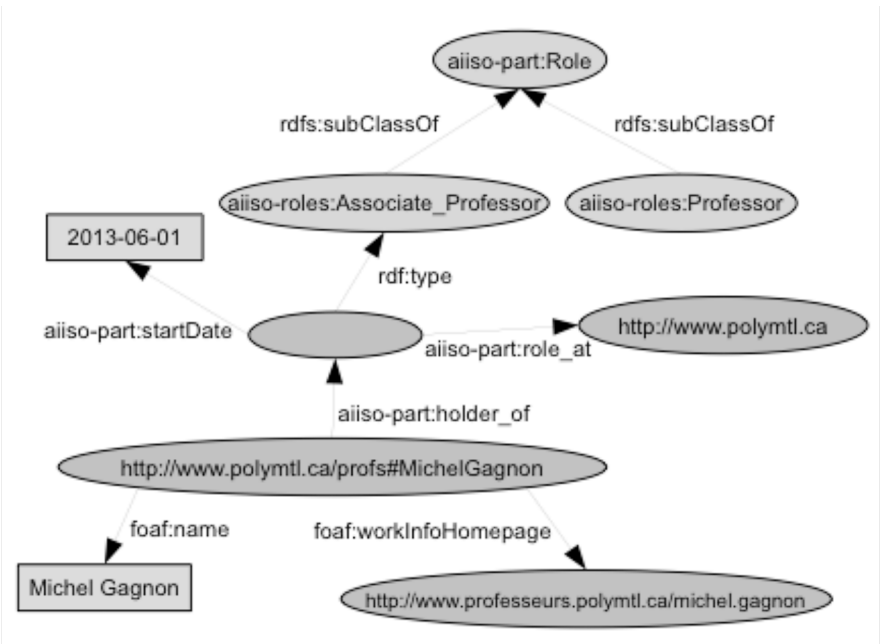
Essentiellement, cette règle exprime que si a, l'auteur d'un document x, est décédé à une date d et que l'écart entre cette date et la date courante c est supérieur à 50 ans, alors ce document est libre de droits. Le langage RIF (Rule Interface Format) a donc été créé pour représenter formellement ce genre de règle (Morgenstern *et al.* 2010).

Il reste actuellement à établir une logique unificatrice pour assurer l'intégration des langages SPARQL, OWL et RIF, qui représentent chacun une perspective sur les connaissances représentées. Il faudra aussi éclaircir les mécanismes de preuve formelle qui seront utilisés pour effectuer les raisonnements nécessaires afin de tirer profit de la masse de connaissances appelée à se développer. Et, finalement, au sommet de tout cela, nous devrons identifier les mécanismes permettant de prendre en compte la crédibilité des sources de données. En effet, la valeur d'une inférence dépend fortement de la crédibilité des informations utilisées. Ces dernières couches du Web sémantique sont encore à un état embryonnaire, mais plusieurs travaux de recherche s'y attaquent.

Le modèle de données RDF

La figure 2 fournit un exemple de données représentées en RDF. On y représente l'auteur de cet article,

Figure 2
Exemple de données en RDF



représenté par l'URI suivant : <http://www.polymtl.ca/professeurs#Michel.Gagnon>.

On y indique que son nom est Michel Gagnon et qu'il assume, depuis 1er juin 2013, le rôle de professeur agrégé à l'École Polytechnique, cette dernière étant représentée par l'URI <http://www.polymtl.ca>. On fournit aussi un lien vers sa page personnelle. L'emploi occupé par Michel Gagnon est représenté par un nœud vide (l'ovale sans étiquette textuelle au centre de la figure 2), ne contenant aucun URI pour l'identifier. Il s'agit ici d'une caractéristique importante de RDF, qui permet de représenter une entité non pas par un identificateur unique, mais par les propriétés qui la lient à d'autres entités.

On remarquera que la représentation utilise un vocabulaire provenant de plusieurs sources, indiquées par les préfixes *rdf*, *rdfs*, *foaf*, *aiiso-roles* et *aiiso-part*. Les deux premiers sont des propriétés déjà définies dans les spécifications de RDF et RDF Schema. Le troisième est le vocabulaire FOAF[2], très utilisé, et qui sert à décrire des personnes. Les deux derniers sont des sous-ensembles de AIISO[3] (Academic Institution Internal Structure Ontology), un vocabulaire pour représenter la structure interne d'une institution académique.

Ce graphe RDF pourrait être représenté de la manière suivante dans la notation Turtle (il s'agit d'une des diverses manières de représenter un graphe RDF) :

```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix aiiso-roles: <http://purl.org/vocab/aiiso-roles/schema#>.
@prefix aiiso-part: <http://purl.org/vocab/participation/schema#>.
@prefix foaf: <http://xmlns.com/foaf/0.1/>.

<http://www.polymtl.ca/profs#MichelGagnon>
foaf:name "Michel Gagnon";
foaf:workInfoHomepage <http://www.professeurs.polymtl.ca/
michel.gagnon>;
aiiso-part:holder_of
    [ rdf:type aiiso-roles:Associate_Professor;
    aiiso-part:role_at <http://www.polymtl.ca>;
    aiiso-part:startDate "2013-06-01"^^xsd:Date ].

aiiso-roles:Associate_Professor rdfs:subClassOf aiiso-part:Role.
aiiso-roles:Professor rdfs:subClassOf aiiso-part:Role.
```

Dans cette représentation, on utilise le point-virgule pour séparer les propriétés définies pour une même entité (dans ce cas-ci l'entité correspondant à Michel Gagnon est désignée par l'URI <http://www.polymtl.ca/profs#MichelGagnon>). Entre crochets [ ], on retrouve toutes les propriétés qui décrivent le nœud vide.

En résumé, RDF se distingue par sa simplicité, sa flexibilité et son extensibilité, qui favorisent la distribution des descriptions dans des documents différents répartis sur le Web. Le fait que les propriétés prennent autant de place dans le modèle RDF n'est d'ailleurs pas étranger à cela. Pour favoriser son utilisation, plusieurs notations ont été proposées, et nous avons à notre disposition le langage de requête SPARQL qui permet de tirer profit efficacement des données représentées en RDF. Pas étonnant qu'à ce jour il y ait déjà un très grand nombre d'outils pour manipuler, explorer, stocker et créer des données RDF. C'est en RDF

2. <http://www.foaf-project.org/>.
3. <http://vocab.org/aiiso/schema >.

qu'a pris forme la principale incarnation du Web sémantique, soit le réseau Linked Data, dans lequel on retrouve déjà beaucoup de données relatives aux bibliothèques numériques.

L'incarnation actuelle du Web sémantique : le réseau Linked Data

Le réseau Linked Data (Heath *et al.* 2011) est un ensemble de données RDF interreliées qui couvrent un très grand nombre de domaines, et dont les sources sont très variées. En septembre 2011, il contenait déjà plus de 31 milliards de triplets. Par exemple, on y trouve les données du recensement des États-Unis (environ un milliard de triplets) et plusieurs données issues de différents gouvernements (le Royaume-Uni est un des premiers à avoir décidé de divulguer ses données en format RDF). Les types de données qu'on y retrouve incluent des statistiques liées aux recensements, les budgets de l'administration, les relevés météorologiques et ceux liés aux catastrophes naturelles, des informations sur les logements subventionnés par l'État, etc. Une grande quantité de données dans le secteur de la bibliothéconomie sont également présentes (voir le tableau 1). On y trouve aussi DBpedia, qui est en quelque sorte une version RDF de Wikipedia, produite en glanant dans les pages de Wikipedia les informations structurées qui s'y trouvent (noms des personnes, dates de naissance, informations géographiques et géospatiales, etc.). DBpedia contient plus de un milliard de triplets RDF et est devenu une référence dans le réseau Linked Data. La majorité des dépôts de données qui s'y trouvent établissent des liens avec des entités de DBpedia.

Il est important de noter que l'intérêt du réseau Linked Data est non seulement l'ensemble de liens que chaque dépôt établit avec au moins un autre dépôt, mais aussi le fait que chaque entité qu'il contient doit être représentée par un URI déréférençable, c'est-à-dire qu'elle pointe sur un serveur Web qui nous fournira une description RDF de cette entité si on y accède. La combinaison de ces deux caractéristiques fait en sorte de favoriser l'enrichissement d'informations de manière automatisée : il suffit de connaître l'URI de l'entité en question. À titre d'illustration, prenons un extrait des données fournies par la Bibliothèque nationale de France sur *L'œuvre au noir*, roman écrit par Marguerite Yourcenar :

```
<http://data.bnf.fr/ark :/12148/cb119466521#frbr :Work>
rdf :type rda :Work ;
rdfs :label «L'œuvre au noir» ;
dc :creator <http://data.bnf.fr/ark :/12148/
      cb11929535g#foaf :Person> ;
dc :date «1968» ;
dc :description «Roman historique»@fr ;
dc :language <http://id.loc.gov/vocabulary/iso639-2/fre> ;
dc :subject <http://dewey.info/class/800/> ;
dc :title «L'œuvre au noir»@fr.

<http://catalogue.bnf.fr/ark :/12148/cb35647589k>
rdf :type rda :Manifestation ;
dc :date «1994» ;
dc :title «L'œuvre au noir» ;
rda :workManifested <http://data.bnf.fr/ark :/12148/
      cb119466521#frbr :Work>.

<http://catalogue.bnf.fr/ark :/12148/cb400716495>
rdf :type rda :Manifestation ;
dc :date «2005» ;
dc :description «10 disques compacts (11 h 10 min)» ;
dc :publisher «[DL 2005] Paris. - Livraphone. - [France]
```

Tableau 1
Données de Linked Data en bibliothéconomie

Dépôt de données	Description	Nombre approx. de triplets (en millions)
Library of Congress	Vocabulaires et normes promulguées par la Bibliothèque du Congrès	Plusieurs millions
Open library	Catalogue de livres avec informations sur les différentes éditions publiées	400
lobid. Bibliographic Resources	Métadonnées sur des ressources bibliographiques (livres, articles, documents PDF)	668
VIAF : The Virtual International Authority File	Unification de fichiers d'autorité de diverses institutions, parmi lesquelles on retrouve la Bibliothèque du Congrès, la Deutsche Nationalbibliothek et la Bibliothèque nationale de France.	200
British National Bibliography	Métadonnées sur des livres publiés	89
RAMEAU	Langage d'indexation matière utilisé par la Bibliothèque nationale de France et diverses autres bibliothèques	1,7
Europeana Linked Open Data	Métadonnées sur les 25 millions d'objets du portail Europeana, un catalogue d'objets numériques disponibles en ligne (images, textes, sons, vidéos)	117
Bibliothèque nationale de France	Catalogue complet de la bibliothèque	6,3
Idref	Notices d'autorité de Sudoc (Système Universitaire de Documentation)	20

```
    [distrib. les Belles lettres]» ;
  dc :title «L'œuvre au noir : roman» ;
  rda :workManifested <http://data.bnf.fr/ark :/12148/
      cb119466521#frbr :Work>.
```

On remarque que l'œuvre est représentée par l'URI <http://data.bnf.fr/ark :-/12148/cb119466521#frbr :Work> et qu'elle est reliée par le prédicat dc :creator à l'URI qui représente Marguerite Yourcenar, soit <http://data.bnf.fr/ark :/12148/-cb11929535g#foaf:Person>. Cette œuvre a fait l'objet de plusieurs éditions (des «manifestations») dont deux sont décrites ici, soit un livre et une version audio en disques compacts. Remarquons la présence de deux liens vers des entités externes aux données de la BNF : le code selon la classification Dewey et la langue, qui est une entité du dépôt de données de la Library of Congress.

Maintenant, si on demande au serveur de la BNF les données concernant l'URI <http://data.bnf.fr/ark:/12148/cb11929535g#foaf:Person>, afin d'obtenir des informations sur Marguerite Yourcenar, on obtient les triplets suivants :

```
<http://data.bnf.fr/ark:/12148/cb11929535g#foaf:Person>
rdf:type foaf:Person;
rda:countryAssociatedWithThePerson <http://id.loc.gov/
      vocabulary/countries/fr>;
rda:dateOfBirth "08-06-1903";
rda:dateOfDeath "17-12-1987";
rda:languageOfThePerson <http://id.loc.gov/vocabulary/iso639-2/
      fre>;
rda:placeOfBirth "Bruxelles";
rda:placeOfDeath "Mount-Desert (Me.)";
foaf:gender "female";
foaf:name "Marguerite Yourcenar";

<http://data.bnf.fr/ark:/12148/cb11929535g>
rdf:type skos:Concept;
foaf:focus rdf:resource="http://data.bnf.fr/ark:/12148/
      cb11929535g#foaf:Person"
owl:sameAs <http://dbpedia.org/resource/Marguerite_Yourcenar>,
  <http://viaf.org/viaf/105154898>,
  <http://www.idref.fr/027201694>;
skos:altLabel «Marguerite de Crayencour (1903-1987)»;
skos:prefLabel «Marguerite Yourcenar (1903-1987)».
```

On remarque ici que la description de l'entité correspondant à Marguerite Yourcenar utilise le vocabulaire FOAF pour spécifier qu'il s'agit d'une personne de sexe féminin dont le nom est Marguerite Yourcenar. On remarque aussi la présence d'un concept abstrait, dont l'URI est <http://data.bnf.fr/ark :/12148/cb11929535g>, et qui est lié à l'entité qui représente Marguerite Yourcenar. Il s'agit en quelque sorte de la représentation « informationnelle » de l'auteure, alors que l'URI <http://data.bnf.fr/ark:/12148/-cb11929535g#foaf:Person> représente plutôt l'auteure elle-même. C'est cet avatar abstrait qui est utilisé pour établir des liens avec la représentation de Marguerite Yourcenar dans d'autres dépôts de données, soit DBpedia, VIAF et Idref (ces liens sont indiqués par la relation owl:sameAs).

Chacun de ces URI pointe sur une ressource du Web qui fournit des données sur Marguerite Yourcenar. DBpedia nous permet de savoir, entre autres, que Grace Frick était sa compagne de vie, qu'elle a reçu le prix Erasmus et qu'elle a été influencée par Chen Ran et Simonetta Greggio. De plus, DBpedia identifie explicitement Marguerite Yourcenar comme étant un écrivain. Ainsi, si on demande à DBpedia de nous fournir toutes les entités qui représentent des écrivains, c'est-à-dire dont le type est <http://dbpedia.org/-ontology/Writer>, on obtiendra une liste qui contiendra l'URI de Marguerite Yourcenar.

VIAF nous fournit les différents noms et formes du nom sous lesquels Marguerite Yourcenar est identifiée dans différents fichiers d'autorité (NLA, PTNTB, VLACC, etc.). Finalement, Idref établit des liens vers plusieurs notices bibliographique concernant Marguerite Yourcenar (autres œuvres, préfaces de livres, interviews, traductions, etc.).

On voit donc que la standardisation de la représentation des informations en RDF et l'existence de liens entre les différences sources de données permettent de produire des outils automatiques profitant d'un substantiel enrichissement des données.

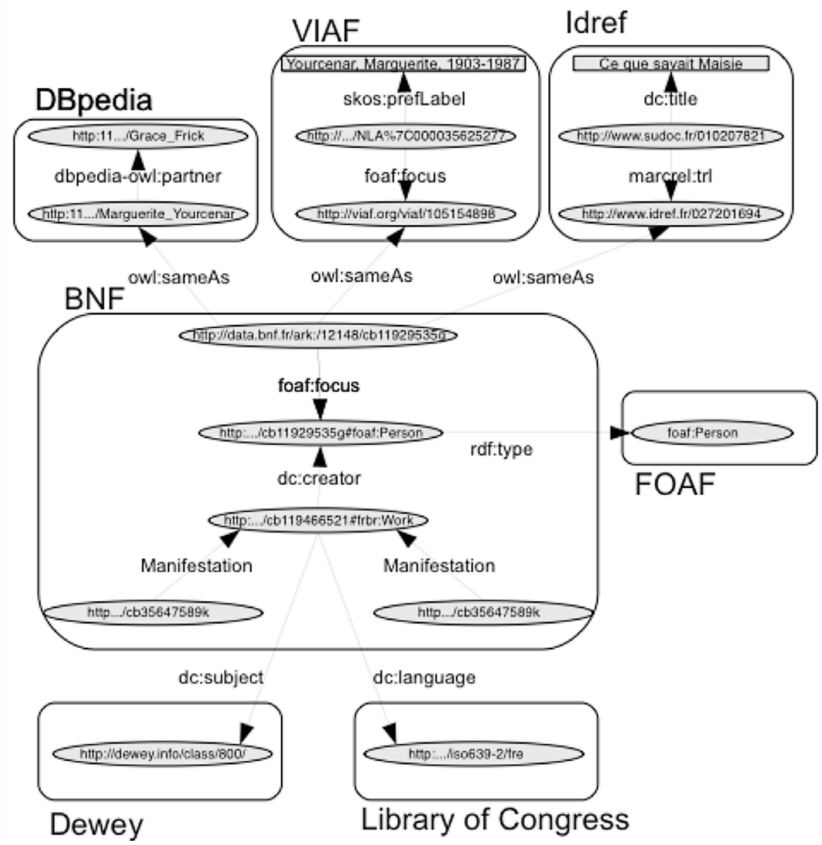
La figure 3 illustre les liens entre différents dépôts de données que nous venons de citer (à noter que Idref utilise le code *trl* de la norme MARC pour indiquer que Marguerite Yourcenar est la traductrice du roman *Ce que savait Maisie*).

Annotation sémantique des documents numériques

Nous avons constaté, dans la section précédente, la richesse des informations disponibles sur le Web sémantique, dont une grande partie est pertinente pour une utilisation en bibliothéconomie. Il nous faut maintenant aborder une question importante : comment peut-on établir les liens entre la masse de documents dans les bibliothèques numériques et ces données représentées en RDF ?

Dans le cas où les métadonnées associées aux documents sont déjà disponibles de manière formelle, on peut envisager des outils qui effectuent la traduction de ces données vers le format RDF. Cette tâche de traduction est compliquée par le fait que le vocabulaire des métadonnées d'une bibliothèque numérique n'est pas nécessairement le même que celui utilisé dans le réseau Linked Data. Il faudra alors établir une correspondance, en s'appuyant éventuellement sur des outils d'appariement automatique. Pour un survol des différentes méthodes d'appariement, on peut consulter (Shvaiko *et al.* 2013). À noter que les approches qui ont été proposées jusqu'à maintenant sont basées sur des appariements terminologiques (on cherche un terme similaire dans l'autre vocabulaire) et structurels (l'appariement tient compte des liens avec les autres concepts).

Figure 3
Liens entre différents dépôts de données contenant des informations sur Marguerite Yourcenar



Mais ces métadonnées laissent dans l'ombre une bonne partie du contenu textuel des documents. Pour intégrer les bibliothèques numériques au Web sémantique, nous avons donc intérêt à extraire les informations pertinentes de ce contenu textuel.

À l'heure actuelle, nous n'avons pas encore de solution définitive pour ce problème, mais plusieurs travaux proposent des pistes intéressantes. La plupart des annotateurs sémantiques actuellement disponibles établissent un lien entre une expression dans le texte et l'entité correspondante dans DBpedia. L'idée, telle que proposée par (Charton *et al.* 2011) et (Mendes 2011), consiste en deux étapes : (1) dans un premier temps, on identifie les expressions candidates dans le texte, c'est-à-dire susceptibles d'être liées à une entité conceptuelle pertinente, et ensuite (2), on identifie, parmi toutes les entités qui pourraient être associées à une expression du texte, la plus pertinente.

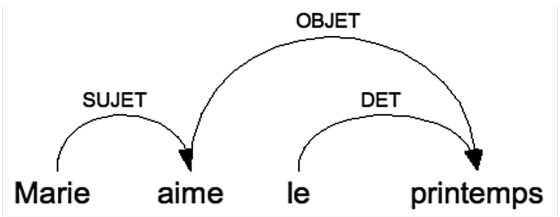
Supposons qu'un texte contienne *Mozart* parmi les expressions candidates. On trouve dans DBpedia plusieurs entités qui pourraient y correspondre :

<http://dbpedia.org/resource/Wolfgang_Amadeus_Mozart>
<http://dbpedia.org/resource/Leopold_Mozart>
<http://dbpedia.org/resource/Mozart_(crater)>
<http://dbpedia.org/resource/Mozart_Programming_System>
...

Pour identifier l'entité la plus pertinente, on compare deux ensembles de mots : ceux qui se trouvent avant et après l'expression *Mozart* dans le texte (l'étendue des mots pris en compte avant et après est une valeur fixée d'avance) et ceux qui se trouvent dans la fiche Wikipedia correspondant à l'entité (rappelons-nous que chaque entité de DBpedia correspond à une entrée dans Wikipedia). On choisit l'entité qui maximise une valeur de similarité. On peut consulter (Huang 2008) pour une présentation de mesures de similarité couramment utilisées. Par exemple, une mesure simple, appelée Jaccard (Jaccard 1901), consiste à calculer le ratio de mots en commun sur le nombre total de mots contenus dans l'union des deux ensembles.

Ce genre d'approche affiche de bonnes performances, mais demeure tout de même limité. En effet, pour extraire l'information pertinente d'un texte, on ne peut se contenter de la simple identification des entités. On veut aussi extraire les relations (Bach *et al.* 2007 ; Zouaq 2010). Pour y arriver, on peut utiliser une simple méthode statistique qui consiste à compter le nombre de fois que deux entités se retrouvent dans la même phrase, et inférer qu'une relation existe entre ces deux entités si cette fréquence est supérieure à un certain seuil. Malheureusement, cette approche ne nous

Figure 4
Exemple d'analyse syntaxique



aide pas à identifier la relation entre les entités. On utilisera donc plutôt des méthodes plus sophistiquées, que l'on regroupera en deux catégories : les approches par apprentissage et les approches par règles.

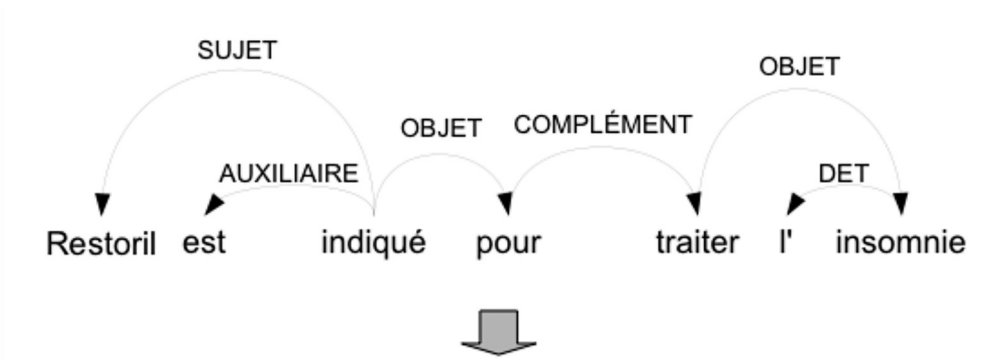
Typiquement, dans les approches par apprentissage, un corpus de textes est annoté manuellement de manière à indiquer quelles phrases expriment la relation qu'on veut identifier. On extrait les caractéristiques intéressantes de ces phrases (mots autour des entités, catégories des mots, structure syntaxique, etc.). On entraîne alors un « classifieur », c'est-à-dire un programme issu des travaux en apprentissage automatique et capable de reconnaître qu'une relation est exprimée par une phrase. Ce programme apprend à reconnaître les caractéristiques des phrases qui expriment une relation donnée, à partir d'un ensemble d'exemples de phrases qui lui sont fournis. On peut aussi adopter une approche semi-supervisée, qui démarre avec quelques exemples de paires d'entités qui sont liées par la relation cible (par exemple : Mozart — Don Giovanni). On cherche ensuite les phrases contenant ces entités et on extrait les attributs pertinents. Puis on cherche de nouvelles phrases qui possèdent ces attributs. À partir de ces nouvelles phrases, on extrait de nouvelles paires d'entités et on recommence le processus jusqu'à convergence, c'est-à-dire lorsqu'on ne découvre plus de nouvelles paires d'entités. Le problème avec les méthodes d'apprentissage est qu'il nous faut des données d'apprentissage pour chaque domaine.

Depuis quelques années, notamment dans le domaine biomédical, on commence à voir apparaître des approches qui effectuent une analyse syntaxique des phrases du document. Une telle analyse syntaxique identifie les relations qui lient les mots de la phrase. Par exemple, l'analyse syntaxique de la phrase *Marie aime le printemps* identifiera *aime* comme le verbe principal de la phrase, auquel *Marie* et *printemps* sont liés par les relations *SUJET* et *OBJET*, respectivement. Le mot *le* est lié à *printemps* par la relation *DET* (déterminant). On obtient ainsi la structure syntaxique illustrée à la figure 4.

À partir de la structure syntaxique, on peut reconnaître certains patrons qui expriment des relations intéressantes. Il suffit alors de trouver les instances de ces patrons dans le document et d'en extraire le contenu qui sera traduit en RDF.

La figure 5 illustre un exemple d'extraction de données RDF à partir de la structure syntaxique d'une phrase. On suppose ici qu'on sait qu'il s'agit de la description d'un médicament dont l'URI est <http://www.drug/3239>. On remarque que le sujet du verbe *indiqué* a été extrait pour représenter le nom du médicament (relation *rdfs :label*). Suivent trois relations qui décrivent un traitement, représenté par l'URI <http://www.drug/tr142>. On indique que ce traitement a pour cible l'insomnie et utilise le médicament en question. Ces données RDF seront générées par l'application d'un patron qui reconnaît une phrase dont le verbe principal est *indiqué*, qui lui-même a un objet de la forme *pour traiter X*.

Figure 5
Exemple d'extraction de données RDF




```
<http://www.drug/3239> rdfs:label "Restoril" .
<http://www.drug/tr142> voc:médicament <http://www.drug/3239> ;
      rdf:type voc:Traitement ;
      voc:cible "insomnie" .
```

Conclusion

Dans cet article, nous avons vu comment le réseau Linked Data constitue actuellement une présence réelle du Web sémantique. Bien que son expressivité sémantique soit plutôt limitée, il contient tout de même un grand nombre de données qui pourront à court terme s'avérer très pertinentes en association aux contenus des bibliothèques numériques. Plus celles-ci seront présentes dans ce Web de données, plus il sera possible de développer des applications utiles profitant de l'enrichissement des métadonnées afin de proposer un meilleur service aux utilisateurs.

Pour y arriver, il reste encore bien des défis à relever. Tout d'abord, chaque bibliothèque numérique doit déterminer le vocabulaire qui sera adopté pour intégrer ses ressources au Web sémantique. Cette démarche exige un compromis entre les caractéristiques particulières du catalogue et l'interopérabilité avec les autres ressources numériques. Parmi les obstacles qui compliquent actuellement cette tâche, on notera l'inexistence d'un véritable consensus sur les vocabulaires à utiliser et la manière de représenter les données, ainsi que la difficulté à générer les données pertinentes, complètes, précises et cohérentes.

Une réflexion s'impose aussi afin de déterminer le genre d'inférences qu'on est appelé à réaliser avec les données en question. Jusqu'à maintenant, l'effort s'est concentré sur la divulgation des données. Une très faible proportion des bibliothèques numériques est présente sur le Web sémantique, et beaucoup se questionnent encore sur la meilleure manière d'y parvenir. Or, il faudra dans un futur proche identifier la manière dont ces données seront utilisées, ce qui aura comme effet bénéfique de jeter un éclairage sur la manière de les produire.

Finalement, nous avons souligné l'importance de développer des extracteurs automatiques afin de produire au moins une première version des données, quitte à faire appel par la suite à des experts pour les nettoyer. Une bonne partie des informations intéressantes se retrouvent à l'intérieur du contenu textuel des documents, et les technologies pour le traitement automatique de la langue sont suffisamment développées pour pouvoir envisager des solutions pour cette tâche dans un avenir proche. ⦿

Sources consultées

Antoniou, Grigoris et Frank van Harmelen. 2012. *A semantic web primer*. Boston, Mass. : MIT Press.

Bach, Nguyen et Sameer Badaskar. 2007. *A review of relation extraction*. Rapport technique. Carnegie Mellon University. <http://www.cs.cmu.edu/~nbach/papers/A-survey-on-Relation-Extraction.pdf> (Consulté le 9 avril 2013).

Bratt, Steve. 2006. *Emerging Web Technologies to Watch*. <http://www.w3.org/-2006/Talks/1023-sb-W3CTechSemWeb/W3CTechSemWeb.pdf> (Consulté le 9 avril 2013).

Berners-Lee, Tim, James Hendler et Ora Lassila. 2001. The semantic web. *Scientific American,* 284 (5) : 34-43.

Charton, Eric, Michel Gagnon et Benoît Ozell. 2011. Automatic semantic web annotation of named entities. In *Canadian Conference on AI*. St-Jean, Terre-Neuve : 74-85.

Heath, Tom et Christian Bizer. 2011. *Linked Data* : *Evolving the Web into a Global Data Space*. San Rafael, CA : Morgan & Claypool Publishers.

Hitzler, Pascal, Markus Krötzsh, Bijan Parsia, Peter F. Patel-Schneider et Sebastian Rudolph. 2011. *OWL 2 Web Ontology Language Primer.* 2nd ed. <http://www.w3.org/TR/2012/REC-owl2-primer-20121211/> (Consulté le 9 avril 2013).

Huang, Anna. 2008. Similarity Measures for Text Document Clustering. *Proceedings of the New Zealand Computer Science Research Student Conference. Christchurch, Nouvelle-Zélande*.

Jaccard, Paul. 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37 : 241-272.

Kruk, Sebastian Ryszard et Bill McDaniel. 2009. *Semantic Digital Libraries*. Berlin, Germany : Springer.

Manola, Frank et Eric Miller. 2004. *RDF Primer.* W3C Recommendation. <http://www.w3.org/TR/rdf-primer/> (Consulté le 9 avril 2013).

Mendes, Pablo N., Max Jakob, Andrés Garcia-Silva et Christian Bizer. 2011. DBpedia spotlight : shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*. New York : 1-8.

Morgenstern, Leora, Chris Welty, Harold Boley et Gary Hallmark. 2010. *RIF Primer (Second Edition).* < http://www.w3.org/2005/rules/wiki/Primer> (Consulté le 9 avril 2013).

Prud'hommeaux, Éric et Andy Seaborne. 2008. *SPARQL Query Language for RDF.* W3C Recommandation. <http://www.w3.org/TR/rdf-sparql-query/> (Consulté le 9 avril 2013).

Shvaiko, Pavel et Jérôme Euzenat. 2013. Ontology matching : State of the art and future challenges. *IEEE Transaction on Knowledge and Data Engineering,* 25 (1) : 158-176.

Zouaq, Amal. 2010. Shallow and deep natural language processing for ontology learning : a quick overview. In *Ontology Learning and Knowledge Discovery Using the Web* : *Challenges and Recent Advances*, sous la direction de Wilson Wong, Wei Liu et Mohammed Bennamoun. Hershey, PA : IGI Global.