

COMPRENDRE LES STANDARDS DU WEB DE DONNÉES

[Bruno Menon](#)

A.D.B.S. | « I2D - Information, données & documents »

2016/2 Volume 53 | pages 32 à 34

ISSN 2428-2111

DOI 10.3917/i2d.162.0032

Article disponible en ligne à l'adresse :

<https://www.cairn.info/revue-i2d-information-donnees-et-documents-2016-2-page-32.htm>

Distribution électronique Cairn.info pour A.D.B.S..

© A.D.B.S.. Tous droits réservés pour tous pays.

La reproduction ou représentation de cet article, notamment par photocopie, n'est autorisée que dans les limites des conditions générales d'utilisation du site ou, le cas échéant, des conditions générales de la licence souscrite par votre établissement. Toute autre reproduction ou représentation, en tout ou partie, sous quelque forme et de quelque manière que ce soit, est interdite sauf accord préalable et écrit de l'éditeur, en dehors des cas prévus par la législation en vigueur en France. Il est précisé que son stockage dans une base de données est également interdit.



WEB DE DONNÉES, UN ENJEU POUR LES BIBLIOTHÈQUES



Bruno MENON est maître de conférences à Paris 8 où il enseigne dans le master Gestion stratégique de l'information. Son parcours antérieur inclut un poste de documentaliste à Sciences-Po puis d'ingénieur R&D dans le secteur du traitement automatique des langues. Ses thèmes de recherche, au laboratoire Paragraphe, concernent l'organisation et la représentation des connaissances.

bruno.menon@univ-paris8.fr

1. Trois points de départ sur le Web permettent une exploration de ces travaux : <https://www.w3.org/standards/semanticweb> ; <https://www.w3.org/2013/data/> et <https://www.w3.org/2001/sw>

2. Ce schéma, ainsi que les citations non référencées, proviennent du W3C.

3. C'est l'Internet Engineering Taskforce (IETF) qui prend en charge la standardisation des URIs/IRIs, mais c'est le W3C qui orchestre leur usage dans le web de données.

Comprendre les standards du web de données

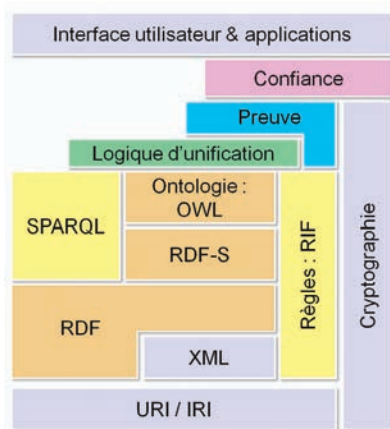
[contexte] Pour publier sur le Web des données accessibles, exploitables, combinables entre elles et interprétables, il faut les exprimer et les décrire dans des formats et des modèles homogènes, stables, explicites et consensuels : bref, il faut se conformer aux standards – ou « langages » – élaborés par le World Wide Web Consortium. L'architecture dessinée par ces standards peut apparaître complexe au premier abord, mais il importe d'en saisir la cohérence. En voici une vue d'ensemble.

Le projet du web sémantique supposait à l'origine de définir des moyens pour adosser les ressources du Web à des descriptions structurées, formalisées et standardisées de ces mêmes ressources, les métadonnées. Assez vite, l'infléchissement vers le web de données conduira à mettre au point des formats standard et à préconiser des bonnes pratiques permettant d'exposer sur le Web des jeux de données exploitables « programmiquement », compatibles, voire liées entre elles – les *linked data* – et interopérables. Les technologies mobilisées dans ce cadre pour l'encodage des données ont fait l'objet dès 1998 d'un ensemble de recommandations du W3C (World Wide Web Consortium) qui ont été peu à peu amplifiées et actualisées¹. Le schéma en couches du web sémantique, familièrement appelé *layer cake*, ici dans une version de 2006², montre comment ces technologies s'articulent et signale la volonté d'aboutir à une architecture cohérente.

Arrêtons-nous sur le rôle des principales briques de cette architecture, en commençant par RDF, qui en est le socle.

Le Layer cake

Source : W3C



Un mille-feuille technologique

Représentation des données : RDF

RDF (Resource Description Framework) se voulait initialement un modèle général de métadonnées, permettant de décrire les caractéristiques des documents du Web. Ce modèle postule que décrire une ressource, c'est renseigner à son sujet une ou plusieurs caractéristiques, c'est-à-dire des propriétés munies de leurs valeurs. L'élément de base de cette description est donc un triplet (ressource, propriété, valeur), que l'on appelle assertion RDF. On dit que la ressource décrite en est le sujet, la propriété renseignée le prédicat et la valeur de cette propriété l'objet. Le sujet et le prédicat sont identifiés par des URIs (Uniform Resource Identifier) ou des IRIs (Internationalized Resource Identifier). L'objet peut être une ressource identifiée par un URI/IRI ou une valeur littérale, comme un nombre ou une chaîne de caractères.

Rappelons ce que sont les URIs/IRIs, qui servent à identifier les ressources³. À côté de leur forme la plus connue, l'URL (Uniform Resource Locator), élément essentiel du Web, on trouve l'URN (Uniform Resource Name) qui identifie les ressources sans nécessairement donner leur chemin d'accès, tout en respectant un certain format. L'IRI peut comporter des caractères Unicode, donc issus de n'importe quel système d'écriture, à la différence de l'URI qui n'admet qu'un sous-ensemble des 128 caractères de l'Ascii, mais qui est plus facile à manipuler dans le contexte actuel des protocoles du Web.

Au fil du temps, du fait de l'extrême généralité de ce modèle, la portée de RDF a évolué : en 1999, on y voyait un « *mécanisme de description des ressources* » constituant une « *base pour traiter les métadonnées* » ; en 2004, on le définit comme un « *cadre pour représenter les informations sur le Web* », conçu pour « *permettre à n'importe qui de faire des déclarations sur n'importe quelle ressource* ». Une recommandation de 2014 précise que « *tout peut être une ressource, y compris des objets physiques, des documents, des concepts abstraits, des*

nombres et des chaînes de caractères ». RDF peut donc être utilisé pour exposer sur le web de données la description de pages ou sites web, de documents, d'images fixes ou animées, mais aussi de personnes, de produits, d'événements, de lieux ou de concepts *via* leurs avatars dans le monde virtuel et grâce au pouvoir de nommer que confère l'utilisation des URIs/IRIs pour désigner ou dénoter les ressources. Créer un URI/IRI revient à donner existence à une ressource qui pourra ainsi jouer son rôle dans des assertions RDF : tout ce qui est identifié par un URI/IRI peut être sujet ou objet d'une assertion.

À un même sujet peuvent être associés plusieurs couples (prédicat, objet). Le sujet d'une assertion peut aussi devenir objet d'une autre assertion. L'objet d'une assertion peut devenir sujet d'une autre assertion, s'il est identifié par un URI/IRI. Il peut aussi être objet de plusieurs assertions. Le prédicat, puisque muni d'un URI/IRI, peut devenir sujet et/ou objet d'assertions à son tour. Tout cela fait que le modèle se prête à la construction de graphes où les sujets et les objets sont des nœuds et les prédicats des arcs. Au-delà de la représentation graphique qui est ainsi suggérée, cette notion permet de s'appuyer sur la théorie des graphes pour développer les outils informatiques permettant l'exploitation des jeux de données RDF ou *datasets*.

La notation (ou sérialisation) de RDF peut être réalisée en XML, mais d'autres formats, plus compacts ou mieux adaptés à certains usages ont fait leur apparition. Un certain nombre d'entre eux ont été entérinés par le W3C : Turtle pour une meilleure lisibilité, N-Triples et N-Quads pour plus de simplicité, JSON pour une intégration plus facile à des programmes.

XML, on le rappelle ici brièvement, est un format normalisé de documents numériques d'usage très général, qui se veut indépendant des plates-formes informatiques. Il permet, en suivant des règles simples et largement diffusées, de mettre en évidence la structure des documents et d'y étiqueter l'information qu'ils contiennent. Avec ce format, les frontières entre documents et bases de données s'estompent : une base de données peut être vue comme un document et inversement.

On peut utiliser un peu différemment les métaphores linguistiques usuelles et parler de cet ensemble de technologies comme d'une langue

Spécification des modèles de données : RDF Schema et OWL

Pour guider et uniformiser la description des ressources en RDF, il est souvent intéressant de spécifier (c'est-à-dire à la fois de contraindre et de documenter) les moyens descriptifs utilisables, ou « vocabulaires ». On peut le faire avec RDF Schema (ou RDFS), qui permet de déclarer des propriétés, de les organiser en hiérarchie et de spécifier leur signature en termes de domaine et de portée. Le domaine d'une propriété est l'ensemble des types de ressources, ou classes, qui peuvent fournir les sujets des assertions où cette propriété est le prédicat ; et sa portée est l'ensemble des classes qui peuvent fournir les objets de ces assertions. Cela suppose de pouvoir déclarer des classes et leur hiérarchie, ce que RDF Schema permet aussi. RDF Schema se situe à la charnière entre la modélisation de données (il y a une parenté évidente avec la modélisation objets) et la représentation des connaissances : une hiérarchie de classes est le premier pas vers une ontologie⁴ et autorise certaines formes de raisonnement automatisé, surtout des inférences simples sur les classes. On parle alors de *lightweight ontologies* ou ontologies légères.

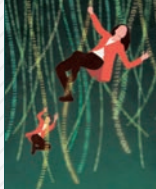
OWL (Web Ontology Language) prend le relais de RDF Schema là où l'on a besoin de rendre compte plus en détail de la structuration conceptuelle d'un domaine d'application, en définissant formellement les classes au moyen d'expressions logiques et en enrichissant la description des propriétés. On peut décrire des objets individuels ou instances en les rangeant dans les classes et en spécifiant les valeurs des propriétés pertinentes dans leur cas. Parmi les primitives de OWL figurent des moyens de déclarer équivalentes des classes ou des propriétés, et de tenir pour identiques des individus ; ce qui est précieux pour assurer l'interopérabilité des applications. Des mécanismes de raisonnement permettent d'enrichir automatiquement les ontologies OWL en opérant des inférences sur les définitions assertées des classes et propriétés.

Raisonnement sur les données : RIF (Rule Interchange Format)

Les parties de l'architecture consacrées à la logique et au raisonnement sont moins développées. Dans l'article de 2001 qui a popularisé l'idée du web sémantique⁵, on peut lire : « *Pour que le web sémantique fonctionne, les ordinateurs doivent avoir accès à des collections structurées d'informations et d'ensembles de règles d'inférence qu'ils peuvent utiliser pour parvenir à un raisonnement automatisé* ». Une règle d'inférence, typiquement de forme « si P alors Q », sert à produire une nouvelle proposition à partir

4. En technologie de l'information, une ontologie est l'ensemble des définitions logiques des concepts d'un domaine et des relations qu'ils entretiennent.

5. T. Berners-Lee T., J. Hendler et O. Lassila. « The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities ». *Scientific American*, mai 2001, p. 29-37. www.scientificamerican.com/article/the-semantic-web



1 WEB DE DONNÉES, UN ENJEU POUR LES BIBLIOTHÈQUES

6. Beaucoup d'inférences sont déjà réalisables à partir des informations exprimées en RDF Schema ou en OWL.

7. Respectivement : le *Répertoire d'autorités matière* utilisé par de nombreuses bibliothèques francophones, le *Thésaurus pour la description et l'indexation des archives locales anciennes, modernes et contemporaines* du ministère français de la Culture, le thésaurus de l'Organisation des Nations Unies pour l'agriculture et l'alimentation.

//// d'assertions existantes⁶. Il existe une grande variété d'approches de ces règles, fondées sur différentes formes de logique : le W3C en cite une dizaine. Après avoir entamé la spécification de son propre langage de règles, SWRL (Semantic Web Rule Language), il a finalement choisi avec RIF de ne pas prendre parti et de spécifier un formalisme « pivot » pour noter les règles issues de ces différentes approches, tout en assurant l'interfaçage avec OWL et RDF.

Un écosystème

Le *layer cake* ne rend pas complètement justice à l'imbrication de ces technologies, et il faut imaginer une sorte de mise en abyme de ses différents niveaux : descriptions de ressources, spécifications de RDF Schema et concepts définis en OWL sont des données représentées en RDF et peuvent donc être formatées (par exemple en XML), échangées et publiées comme telles.

L'exploitation de ces données revient *in fine* aux applications qui y ont intérêt. Il existe néanmoins des outils génériques pour les manipuler, en particulier lorsqu'elles sont au format XML, ce qui favorise grandement l'adoption de cette technologie. Du fait du rôle central joué par RDF dans le web de données, le W3C a aussi défini des moyens spécialement destinés à travailler sur les graphes RDF : SPARQL (à prononcer « sparkle »). Il s'agit d'un langage de requête sur des « bases de données » RDF : les *triple stores* ou entrepôts de triplets. Il est à RDF ce que SQL est aux bases de données relationnelles : il donne la possibilité de rechercher, mais aussi de créer, modifier et supprimer des assertions RDF. SPARQL spécifie également un format de résultat (en XML) et un protocole pour communiquer avec le « guichet » ou point d'accès SPARQL (*SPARQL endpoint*) d'un serveur de données RDF, notamment *via* HTTP.

Prenons l'exemple d'un langage documentaire, exposé sur le web de données : il existe une recommandation du W3C, Skos (Simple Knowledge Organization Systems), qui définit un vocabulaire RDF (ou une ontologie légère), c'est-à-dire les classes et les propriétés permettant de représenter un système d'organisation des connaissances ; autrement dit les termes, concepts et relations d'un vocabulaire d'indexation, d'un thésaurus, d'une classification, etc. Skos est défini au moyen des primitives de RDF Schema et de OWL, mais ses utilisateurs n'ont pas besoin en principe de se reporter à ces technologies, il leur suffit de « traduire » les données en RDF au moyen de ce vocabulaire et de les enregistrer dans l'un des formats licites, par exemple XML. L'avantage de cette démarche est qu'ainsi représenté, le langage documentaire devient exploitable dans de nombreux

Le projet du web des données vise à faciliter la formulation des données en vue de leur partage et de leur échange à travers le Web

contextes avec un minimum de difficultés techniques : on peut le manipuler avec SPARQL, on peut lier ses termes à ceux d'autres ressources Skos, etc. Le répertoire *Rameau*, le *Thésaurus W*, *Agrovoc*⁷, entre autres, ont déjà suivi cette voie.

Métaphores linguistiques

Toutes ces technologies sont fréquemment décrites avec une terminologie inspirée de la linguistique : souvent qualifiées de langages ou de dialectes, elles ou leurs éléments constitutifs sont définis par leur vocabulaire, leur syntaxe, leur sémantique, etc. Cela est légitime dans la mesure où il s'agit d'une habitude ancienne en technologies de l'information, héritée de la logique formelle : en codant l'information pour la machine, on adopte bien les conventions d'un langage informatique.

Mais on pourrait aussi utiliser un peu différemment les métaphores linguistiques usuelles et parler de cet ensemble de technologies comme d'une langue : le projet du web des données vise à faciliter la formulation des données en vue de leur partage et de leur échange à travers le Web. En cela et vu dans sa globalité, il s'agit bien d'un système d'expression et de communication. Les URIs/IRIs en sont les mots et les ontologies conçues en RDF Schema ou en OWL sont les dictionnaires qui permettent d'en élucider la sémantique. RDF, qui régit la construction des assertions, en est la syntaxe. Et XML est l'un des systèmes d'écriture avec lesquels cette langue peut être notée. Les jeux de données RDF sont les textes ou les ouvrages de référence rédigés dans cette langue, que SPARQL permet de consulter et de compiler. La strate des règles d'inférence et celle des interfaces et applications pourraient tenir lieu d'une pragmatique et compléter ainsi l'image d'ensemble. ■