HAL open science

# Une toile de fond pour le Web : lier les données et lier leurs vocabulaires sur la toile, pour un Web plus accessible aux machines

Fabien Gandon

▶ **To cite this version:**

Fabien Gandon. Une toile de fond pour le Web : lier les données et lier leurs vocabulaires sur la toile, pour un Web plus accessible aux machines. Annales des Mines - Enjeux Numériques, 2020. hal-02748219

**HAL Id: hal-02748219**

**https://inria.hal.science/hal-02748219v1**

Submitted on 3 Jun 2020

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

# Une toile de fond pour le Web : lier les données et lier leurs vocabulaires sur la toile, pour un Web plus accessible aux machines

Par **Fabien GANDON**
Directeur de Recherche
Inria

Attention…Top ! Je parle toutes les langues. J'ai plus de trois milliards d'utilisateurs directs. Je peux être privé ou public. Je m'étends sur tous les réseaux du monde. Je suis derrière votre réservation de vacances, dans votre téléphone, vos échanges avec une assurance ou votre livre électronique… je suis … je suis… je suis… Le Web, la toile ou l'ouaibe, pour nos cousins d'outre-Atlantique. Nous avons toujours cherché dans l'Histoire des moyens efficaces de collecter et d'accéder aux masses d'informations que nous créons, c'est notamment la raison d'être des bibliothèques. Ce fut aussi la motivation de Tim Berners-Lee lorsqu'il proposa en 1989 de tisser un système d'hypertexte global [F] au CERN pour mieux partager les informations dans un campus où plusieurs milliers de personnes se croisent avec de multiples spécialités et instruments.

Aujourd'hui encore, il est frappant de voir à quel point le Web est à la fois très connu et mal connu, comme en témoigne la confusion tenace entre les termes Web et Internet que l'on rencontre encore bien trop souvent. Malgré le fait que leurs inventeurs respectifs aient reçu deux prix Turing bien distincts, respectivement en 2004 et en 2016 pour deux inventions bien différentes, Internet et Web sont encore trop souvent utilisés de façon interchangeable.

Redisons-le : Internet permet l'interconnexion des réseaux d'ordinateurs et objets connectés en général. Il fournit une infrastructure de communication qui supporte au-dessus d'elle de nombreuses applications comme : la messagerie électronique (mail), la téléphonie et la vidéophonie… et le Web, cet hypermédia distribué qui devient l'architecture logicielle majoritaire des applications sur Internet. Une autre information encore moins connue à propos du Web est que depuis la fin des années 1990, il n'est plus uniquement consulté et utilisé par nous, les humains, mais aussi par les machines, notamment dans ses versions que l'on appelle Web de données et Web sémantique.

## Un réseau hypermédia de ressources

Dès 1996, Tim Berners-Lee fait une relecture de l'architecture du Web [E] en insistant sur trois concepts-clés : l'adressage (les adresses Web ou URL et URI de la forme http://inria.fr), le protocole de transfert de données pour le Web (HTTP) et le mécanisme de négociation de contenu. Ce dernier est un mécanisme du protocole HTTP qui permet à un serveur Web de fournir pour une même adresse URL différentes représentations d'une même ressource en fonction de ce qu'il sait sur celui qui l'interroge. Par exemple, s'il connaît les langues parlées par celui qui accède à une adresse, il pourra préférer lui servir une version de la réponse dans une de ces langues. Cette négociation de contenu a lieu à chaque fois que nous accédons au Web et sans même que nous le voyions.

Les possibilités de la négociation de contenu vont plus loin que cet exemple et, d'une certaine façon, déclassent la célèbre « page Web » en nous donnant la possibilité de négocier auprès d'un serveur Web différents types de formats de réponse. Ainsi le Web n'est pas limité à une toile de documents

mais offre la possibilité de servir et lier tout et n'importe quoi. En effet, comme les URI permettent d'identifier tout type de ressources (une page, une image, une personne, un produit, une molécule, etc.) et pas uniquement les contenus du Web, on peut dès lors utiliser la toile et ses langages pour décrire et lier tout ce que l'on sait identifier dans le monde. Le Web affirme ainsi son indépendance à un modèle ou une structure de données et le langage HTML des pages Web redevient donc juste un prérequis pour un navigateur [T]. Il aura permis dans un premier temps de fournir un format uniforme de documents hypertextuels et de documentariser le réseau de ressources que devient le Web. La toile est prête à échanger beaucoup d'autres choses que des pages.

## Un Web plus accessible aux machines

En 1996, le langage PICS permet de standardiser le filtrage des contenus inappropriés, notamment pour les enfants. Incidemment, PICS ouvre aussi l'idée d'étiqueter les contenus avec des données pour les machines. Le Web s'ouvre à la notion de métadonnées en général ; commence alors une évolution vers un Web de documents et de données structurées.

Dans cette évolution, le langage de feuilles de styles CSS est une étape importante qui marque le début de la séparation du fond et de la forme sur le Web. La notion de feuille de style permet de sortir et séparer la mise en forme de la structure du document et d'utiliser une même mise en forme pour plusieurs documents, ou inversement de faire varier la mise en forme d'un même document. Peu après, le Web connaît une nouvelle évolution avec le standard XML permettant de créer et gérer ses propres structures de documents et données. Prolongeant cette évolution, Tim Berners-Lee publie en 1998 une feuille de route pour ce qu'il appelle le Web sémantique [AK]. Elle est dans la continuité de sa présentation de 1994 [S] et aussi de son article de la même année [T]. On peut lire dans cet article qu'il souhaite une évolution des objets du Web, qui sont à l'époque essentiellement des documents destinés aux humains, vers des ressources avec une sémantique plus orientée vers les machines pour permettre des traitements plus automatisés. La feuille de route de 1998 [AK] ouvrira la voie à tous les travaux sur le Web de Données et le Web sémantique et aux standards qui en découlent (RDF, RDFS, SPARQL, OWL, etc.).

## Web sémantique : quand le lien fait sens

Si la notion de Web sémantique date de 1998 et que l'article grand public le plus connu date de 2001 [BZ], elle peut, vingt ans après, s'expliquer en deux grandes étapes : premièrement, la notion de données liées et le Web de données et, deuxièmement, la notion de schémas liés et le Web sémantique.

### Données liées et Web de données

Le principe des données liées est de créer des liens entre les données, comme on crée des liens entre des pages et, par extension, de créer des liens entre les bases de données, comme on crée des liens entre les sites.

La première étape est d'utiliser les identifiants standardisés par le Web pour identifier les sujets et relations de ces données. Ainsi, je peux forger une adresse Web (URI) pour identifier une berline du parc automobile de mon entreprise : http://www.mon-entreprise.fr/voiture/berline-n3
On imagine bien que la voiture concernée n'est pas accessible par le Web à cette adresse mais que cette adresse peut la représenter dans des données qui la décrivent. Ce même identifiant pouvant être réutilisé dans plusieurs sources de données, il permet alors de faire des liens entre ces données et entre ces sources : on parle de données liées.

De même, je peux forger une adresse Web (URI) pour identifier une relation entre une personne et une voiture : http://www.mon-entreprise.fr/voiture/estConducteurDe

A nouveau, ceci fournit un identifiant pour représenter l'ensemble des occurrences d'une relation de conducteur entre une personne et une voiture. Cet identifiant peut être réutilisé par autant de jeux de données qui le souhaitent et ainsi favoriser une interopérabilité entre les applications consommant ou produisant ces données.

Dans ces deux exemples, les URI utilisent le protocole HTTP, *i.e.* elles commencent par http://... On parle d'alors d'URI HTTP. On ne les qualifie pas d'URL car contrairement à une adresse (par exemple http://fabien.info) elles ne correspondent pas à une ressource (par exemple une page Web) disponible sur la toile. Par contre, le fait d'utiliser le protocole HTTP permet d'avoir un mécanisme par défaut pour découvrir ce que ces URI identifient en accédant à l'adresse qu'ils donnent pour obtenir des données à leur propos. C'est ce que l'on appelle la déréférenciation.

Le Web de données met donc en relation des sources de données plus ou moins grandes en reposant sur l'architecture classique du Web et en l'étendant. Comme ce ne sont donc plus uniquement des pages qui sont liées sur le Web, mais des identifiants de ressources arbitraires, se pose la question de ce que l'on doit obtenir lorsque l'on accède à de tels identifiants. Lorsqu'un identifiant est consulté, les serveurs répondent en fournissant des données décrivant la ressource sans que celle-ci soit nécessairement sur le Web (par exemple, une voiture, une espèce animale, une protéine, un auteur...) et en s'adaptant à celui qui interroge les données. Ainsi, pour un même identifiant et grâce au mécanisme de négociation de contenu, un utilisateur devant son navigateur recevra une page Web en HTML pour sa lecture, là où un agent logiciel recevra des données à intégrer à sa base de données.

On identifie classiquement cinq étapes et critères incrémentaux de qualité pour la publication de données ouvertes liées cinq étoiles sur le Web :

| | |
|---|---|
| ★ | les données sont sur le Web sous licence libre |
| ★★ | *idem* + les données sont explicites et structurées |
| ★★★ | *idem* + les données sont dans un format non propriétaire |
| ★★★★ | *idem* + des URI HTTP sont utilisés pour identifier sujets, objets et types de relations |
| ★★★★★ | *idem* + les données sont liées à d'autres données |

L'évolution du Web documentaire vers le « Web des données » repose sur ces principes et standards, permettant de tout identifier et de tout décrire sur la toile, et de tisser ainsi un graphe de données mondial. En appliquant ces principes, là où on avait avant un Web de documents (les pages Web) essentiellement à consommation humaine, à partir des années 2006-2007, on ajoute un Web reliant des bases de données de toutes tailles et sur tous les sujets (Figure 1), essentiellement à consommation des machines, qui peuvent les parcourir, suivre les liens pour trouver de nouvelles sources et naviguer et chercher ce Web de données comme nous naviguons et cherchons sur les pages du Web. L'appellation « Web de données » insiste donc sur la possibilité d'ouvrir nos silos de données de toutes tailles, depuis notre agenda jusqu'aux immenses bases géographiques, et de les échanger, de les relier, de les composer selon nos besoins.

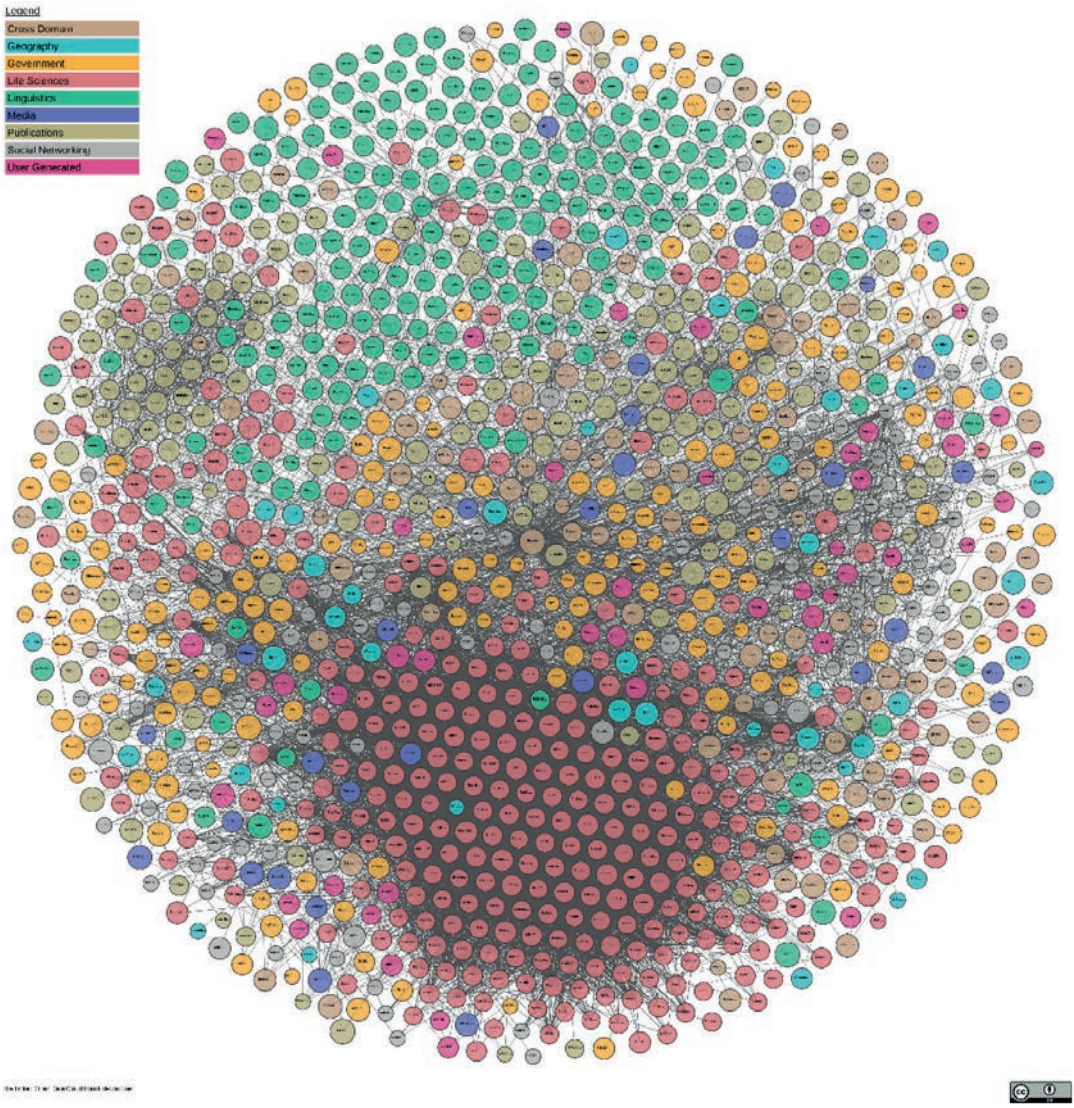


Figure 1 : Le nuage d'une partie des bases de données ouvertes liées du Web le 29/03/2019 par le site lod-cloud.net (The Linked Open Data Cloud, https://lod-cloud.net/).

Pour se représenter l'impact de ce changement et les volumes de données qui deviennent disponibles, nous allons prendre quelques exemples. L'une des bulles de la figure 1 représente DBpedia qui publie sur le Web les données liées que ce site extrait de l'encyclopédie Wikipédia. Au moment d'écrire cet article, cette seule bulle dans le nuage des bases de données ouvertes représente déjà 38 millions de sujets décrits par 3 milliards de données élémentaires (attributs et relations de ces sujets) issus de 125 langues différentes dans Wikipédia et mises ainsi à disposition comme des données structurées ouvertes. Une autre bulle représente le projet Wikidata qui permet de directement saisir et publier des données structurées actuellement à hauteur de 75 millions de sujets décrits par 23 000 utilisateurs. Dans des domaines spécialisés comme la biologie, une des bulles représente Uniprot fournissant 179 millions de données élémentaires, notamment à propos des protéines. Et ce nuage ne rend pas compte de beaucoup d'autres sources de données. Par exemple, les grands moteurs de recherche derrière l'initiative Schema.org nous permettent de

mettre des données structurées dans nos pages Web et des millions de sites le font pour être indexés plus précisément. Autre exemple, mais même technique, tous les sites que vous voyez arborer un bouton « Like » de Facebook incluent dans leurs pages des données structurées descriptives afin de nourrir ce bouton lorsque vous cliquez dessus. Les volumes de données structurées sur le Web ont donc explosé dans les quinze dernières années, sans que l'utilisateur lambda ne s'en aperçoive nécessairement.

Pour représenter et échanger ces données à l'échelle du Web, nous avons besoin de définir des standards pour leurs modèles, structures, formats et langages. RDF (pour *Resource Description Framework*) est au Web de données ce que HTML est au Web documentaire : le langage RDF permet de représenter et de relier des données à propos de ressources. RDF s'insère parfaitement dans l'architecture du Web, notamment en utilisant les URI pour identifier les ressources et les types de relations décrits par les graphes qu'il permet de représenter.

Par ailleurs, RDF fournit également un modèle de données servant de fondation à d'autres standards. Ainsi, au-dessus de RDF, le standard SPARQL fournit un langage d'interrogation et de modification des graphes RDF et un protocole pour soumettre de telles requêtes à un serveur distant. Par exemple, sur le site DBpedia (données RDF extraites de Wikipédia), qui est l'une des bases dans le nuage de la figure 1, on peut demander en SPARQL tous les URI des ressources nommées « Paris » en français. A partir des identifiants reçus, on peut à nouveau interroger le site pour avoir des données supplémentaires et ainsi passer de données liées en données liées comme on passerait de page en page. Un autre standard nommé SHACL permet quant à lui de valider des données en capturant des règles auxquelles la structure des graphes RDF doit se conformer pour être valide (par exemple tous les livres doivent avoir un titre). Cette validation est utile pour vérifier les données échangées entre applications, par exemple.

De telles descriptions peuvent provenir de n'importe quelle source sur le Web et être fusionnées avec d'autres. Le terme de « gigantesque graphe global » (*Global Giant Graph*) désigne parfois cette toile de données d'envergure mondiale tissée par des milliers de descriptions distribuées sur le Web déclarant des liens entre des nœuds identifiés par des URI.

## Schémas liés et Web sémantique

Dans cette deuxième étape nous publions en plus sur le Web les schémas de ces données, c'est-à-dire les vocabulaires et les règles qui régissent leurs valeurs, leurs structures, leur utilisation, leur interprétation... bref leur sens, leur sémantique. Ces schémas et leurs termes utilisent eux aussi des identifiants du Web (par exemple une URI identifiant la catégorie « femme ») et des liens pour déclarer des relations entre les notions qu'ils définissent (par exemple, une femme est une personne) leur donnant ainsi un sens et tissant un Web sémantique.

Le « Web sémantique » permet donc la formalisation, la publication et le liage des vocabulaires utilisés dans les descriptions RDF. Ces vocabulaires permettent aux applications d'utiliser plus efficacement les données du Web en reconnaissant les différents types de ressources et de liens qu'elles rencontrent, et en exploitant le sens et les raisonnements qui leur sont attachés. Une application peut ainsi faire la différence entre des ressources nommées « Charles de Gaulle » mais de types différents (l'homme, une rue, une résidence, le poète, l'aéroport, le porte-avions…).

Différents types de modèles sont conçus pour fournir des vocabulaires permettant de décrire notre monde sur le Web, on parle notamment d'ontologies informatiques et de thésaurus. En interrogeant et en raisonnant sur ces modèles informatiques, il est possible d'améliorer des fonctionnalités existantes et d'en proposer de nouvelles. Au-dessus de RDF se dresse ainsi la pile des langages de schémas, ayant une expressivité et un coût de calcul croissants : plus l'on monte dans la pile et plus les définitions logiques du vocabulaire permettent de capturer précisément les

structures et le sens des données, mais aussi plus les raisonnements qu'ils permettent sont coûteux en termes de complexité et donc de temps de calcul. Le premier niveau dit « des schémas légers » est celui de RDFS (*RDF Schema*) permettant de déclarer et de nommer les classes de ressources (comme les livres, les films, les personnes…) et leurs propriétés (comme l'auteur, l'acteur, le titre…) et d'organiser ces types dans des hiérarchies. On appelle aussi ces schémas des ontologies légères. Au-dessus de RDFS, la recommandation OWL (*Ontology Web Language*) permet de représenter formellement les définitions d'ontologies plus lourdes et s'organise en plusieurs fragments d'expressivité plus ou moins étendue, qui permettent des déductions supplémentaires en contrepartie de temps de calculs plus longs.

Dans la continuité du Web de données, le « Web sémantique » met donc l'accent sur la possibilité d'échanger les schémas de nos données et la sémantique associée. Formalisés et publiés selon des standards, ces modèles permettent d'enrichir la gamme des traitements automatiques qui peuvent être appliqués aux données. En ouvrant les données et leurs modèles, le Web de données et le Web sémantique ouvrent l'ensemble des utilisations qu'il est possible d'en faire.

## Un projet littéralement infini

Le Web de données et le Web sémantique sont déjà adoptés et déployés dans beaucoup d'applications. Ils n'en sont pas moins toujours sujets à de nombreux travaux de R&D sur des questions, par exemple, de recherche de plus d'efficacité dans les passages à l'échelle, d'intelligence dans les traitements, ou de robustesse face à l'hétérogénéité, la qualité ou l'incertitude dans les données. Mais de toute façon, pour le Web, il s'agit là d'une direction d'évolution parmi de nombreuses autres.

Si dans les années 1990, le problème de Tim Berners-Lee était de faire imaginer un monde disposant du Web avant que celui-ci n'advienne, nous sommes maintenant dans le cas inverse où les gens oublient ou n'imaginent plus ce que serait un monde sans le Web [M]. Cependant, la défense du Web et de son expansion ouverte reste un enjeu. Il est universellement utile et utilisé mais il reste fragile et son idéal de départ pourrait n'être qu'une parenthèse historique si l'on ne veille pas en permanence à sa préservation, notamment en évitant toute forme de recentralisation, par exemple la centralisation de certaines données par certaines firmes. Il faut toujours garder à l'esprit que le Web n'est pas une réalisation acquise mais, par conception, un interminable projet.

## Pour en savoir plus

Fabien GANDON, « Pour tout le monde : Tim Berners-Lee, lauréat du prix Turing 2016 pour avoir inventé... le Web », *Bulletin de la Société informatique de France,* 1024, Société informatique de France, 2017. https://hal.inria.fr/hal-01623368

Fabien GANDON, "A Survey of the First 20 Years of Research on Semantic Web and Linked Data", *Revue des Sciences et Technologies de l'Information - Série ISI: Ingénierie des Systèmes d'information*, Lavoisier, 2018, https://hal.inria.fr/hal-01935898

## Bibliographie

[E] BERNERS-LEE T. (1996), *The World Wide Web: Past, Present and Future*, August. https://www.w3.org/People/Berners-Lee/1996/ppf.html

[F] "Information Management: A Proposal" (March 1989), the original proposal for the software project at CERN that became the World Wide Web. https://www.w3.org/History/1989/proposal.html

[M] SAVAGE N., “Weaving the Web”, Communications of the ACM, Vol. 60 No. 6, Pages 20-22, 10.1145/3077334 https://cacm.acm.org/magazines/2017/6/217732-weaving-the-web/fulltext

[S] BERNERS-LEE T., “Plenary Talk extracted slides”, First WWW Conference, Geneva 94, https://www.w3.org/Talks/WWW94Tim/

[T] BERNERS-LEE T., CAILLIAU R., LUOTONEN Ari, NIELSEN H. F. and SECRET A. (1994), “The World-Wide Web, Commun”, ACM, August, Vol. 37, n°8, 0001-0782, pp. 76-82, 10.1145/179606.179671, ACM, New York, NY, USA.

[AK] BERNERS-LEE T. (1998), “Semantic web road map”, https://www.w3.org/DesignIssues/Semantic.html

[BZ] BERNERS-LEE T., HENDLER J. and LASSILA Ora (2001), “The semantic web”, *Scientific American*, 284.5 (2001): 28-37.