À la recherche des connaissances du Web...

Serge Abiteboul [1], Fabien Gandon [2], Fabian Suchanek [3]

Le Web contient une masse impressionnante de données, plus ou moins explicites et plus ou moins accessibles aux machines. Nous discutons ici des grandes tendances pour le management de ces données : l'extraction de connaissances du Web, l'enrichissement des connaissances par la communauté des internautes, leur représentation sous forme logique, et leur distribution à travers toutes les facettes du Web. Nous allons montrer comment ces développements rendent les données sur le Web plus sémantiques, plus maniables par les machines, plus accessibles aux applications et donc finalement plus utiles pour l'humain.

Le Web : d'une métaphore documentaire à une architecture de données

Les données numériques tiennent une place de plus en plus importante dans notre société. Depuis les années 60, les logiciels de bases de données se sont imposés pour permettre le partage des données à l'intérieur d'une entreprise ou d'une organisation. Ces données qui étaient isolées dans des centres de calcul sont devenues accessibles partout dans le monde avec l'arrivée d'Internet, le réseau des *réseaux de machines*, puis du Web (la « Toile »), d'abord perçu comme un *réseau de contenus*, puis au fil de ses évolutions comme un *réseau d'individus* et finalement comme un *réseau de ressources* [9] en général et numériques en particulier. Le Web est donc devenu plus

1. Serge Abiteboul est directeur de Recherche à Inria, à l'ENS Cachan, et membre du Conseil national du numérique.

2. Fabien Gandon est responsable de Wimmics, équipe de recherche commune Inria, I3S (CNRS, Université de Nice), et représentant d'Inria au W3C (consortium international de standardisation du Web).

3. Fabian Suchanek est Maître de conférences à Telecom ParisTech.

complexe, s'enrichissant en facettes diverses comme le multimédia, les données, les connaissances, les objets communicants, les services, etc.

La métaphore la plus utilisée pour expliquer le Web à sa naissance est celle de la bibliothèque universelle. La « toile d'araignée » mondiale est alors perçue comme un gigantesque système documentaire où les pages sont des textes entre lesquels on navigue en suivant des liens et que l'on peut marquer d'un signet comme on le ferait à l'intérieur d'un livre.

Moins de 20 ans après sa naissance, en 2008, le Web comptait déjà plus de 1000 milliards de pages et, chaque mois, les internautes réalisaient des dizaines de milliards de recherches. Surtout, on pense que le monde numérique double tous les 18 mois et le trafic sur Internet est déjà chaque année supérieur à tout ce que nous pourrions stocker en utilisant tous les supports, tous les disques disponibles. Nous baignons aujourd'hui dans un monde numérique qui s'insère dans notre quotidien au travers de milliards d'objets communicants qui viennent chacun contribuer et consommer ses ressources numériques.

Au-delà de cette toile sur laquelle nous « surfons » chaque jour, lorsque l'on parle d'« architecture du Web », on se réfère maintenant aux standards qui définissent l'infrastructure technologique dont il émerge, en perpétuel renouvellement. Trois notions fondamentales sont au cœur de son architecture originelle :

– L'**URL**, pour « Uniform Resource Locator », est un identifiant qui donne un chemin d'accès pour obtenir une ressource du Web. Par exemple « http ://www.inria.fr/ » est l'URL de la page d'accueil d'Inria. Cette adresse identifie la page et permet d'y accéder.

– Le protocole **HTTP**, pour « HyperText Transfer Protocol », permet, à partir d'une adresse URL, de demander une représentation de la ressource identifiée (par exemple une page Web) et localisée par cet URL, et d'obtenir en retour soit les données de cette représentation, soit des codes d'erreur indiquant un problème rencontré. Nous avons tous déjà rencontré la célèbre « erreur 404 » qui indique que la page demandée n'a pas été trouvée.

– **HTML**, pour « HyperText Markup Language » est le format de données utilisé pour représenter les pages Web. C'est un langage de balisage permettant d'écrire de l'hypertexte, d'où son nom. HTML, dérivé du Standard Generalized Markup Language (SGML), permet de spécifier la mise en forme des pages et d'inclure des ressources multimédias comme des images. Il est souvent utilisé conjointement avec des langages de programmation comme JavaScript et des formats de présentation comme CSS.

Très vite les pages HTML sont passées du statut de documents statiques à celui de textes générés dynamiquement par des applications, par exemple pour décrire le résultat d'une requête par mots-clés dans un moteur de recherche ou de celui d'une requête SQL dans un annuaire d'entreprise. Ces contenus dynamiques constituent ce que l'on nomme le « Web profond » (on dit aussi invisible, ou caché, et « Deep

Web » en anglais). Ce Web a la particularité d'être créé à la demande en réponse à des questions des internautes, par appel à des programmes et des données qui viennent ainsi tisser de la toile au vol ; s'il est accessible en ligne, il n'est en général pas indexé par des moteurs de recherche.

Les informations disponibles sur la toile sont devenues d'énormes gisements de connaissances à découvrir, à *valoriser*. L'analyse de données a été un domaine très actif, quasiment depuis les débuts de l'informatique, sous divers noms comme fouille de données ou business intelligence. Du fait de l'accroissement des capacités des disques et des mémoires, et des puissances de calcul avec des clusters pouvant aller jusqu'à des milliers de machines, du fait aussi de l'explosion des données disponibles, l'analyse de données pour en extraire de la valeur est devenue une industrie florissante. Et c'est sous le nom de « Big data » (en français, « Grosses données » clairement moins glamour) qu'elle se développe aujourd'hui. Comme si le défi posé n'était pas déjà suffisant, ces grandes masses de données prennent une complexité supplémentaire au contact du Web en atteignant des volumes inouïs, en utilisant les langages et l'architecture du Web pour se lier entre elles, formant une nouvelle extension du Web appelée « Linked Open Data ».

Nous allons présenter ici plusieurs développements actuels du Web. Nous allons montrer comment ces développements rendent le contenu du Web plus *sémantique* et donc plus accessible aux machines. Nous discuterons notamment l'extraction de connaissances du Web, l'enrichissement des connaissances par la communauté des internautes, leur représentation sous forme logique, et leur distribution à travers des systèmes.

À la recherche de la sémantique perdue sur le Web

Jusqu'à maintenant, nous avons parlé du Web comme d'une collection de pages HTML. Ces pages ont une grande valeur pour les humains, mais beaucoup moins pour les machines. Là où un humain voit un article sur Wikipédia, ou un avis sur un produit sur eBay, la machine ne voit, a priori, que des chaînes de caractères. Les machines ne comprennent pas qu'un article sur Wikipédia parle d'une personne et de sa biographie, où qu'une page sur TripAdvisor parle d'un voyage dans une île idyllique. Les machines comprennent facilement la notion de mot. Cela explique la base du fonctionnement des programmes parmi les plus populaires du Web, les moteurs de recherche. L'internaute propose des mots et la machine renvoie des pointeurs vers des pages du Web qui contiennent ces mots. Si on pose comme question le mot « Java », on obtient les pages qui mentionnent ce mot, qui parlent peut-être de l'île d'Indonésie, de la danse, du café, ou du langage de programmation. Un programme qui ne connaît que des listes de mots n'est pas capable de comprendre que le mot « Java » a plusieurs significations, et que, pour un humain, une page est beaucoup plus que la somme de ses mots. Avec ce type de recherche par mot-clé, il est très

compliqué de trouver, par exemple, tous les politiciens français qui ont un doctorat, tous les présidents de démocraties qui sont mariés à des chanteuses, ou tous les films qui sont diffusés dans les cinémas autour de ma ville. Le Web contient bien sûr les réponses à toutes ces questions. Mais les réponses sont distribuées sur plusieurs pages, et cachées dans des textes qui ne sont, pour une machine, que des listes de mots.

Pour aller plus loin, la machine a besoin de « comprendre » le sens du contenu des pages. Elle a besoin de construire une représentation de connaissances du Web. Si, par exemple, on disposait d'une base de connaissances des présidents du monde, de tous les chanteurs et chanteuses célèbres du monde, et des relations matrimoniales sur la planète, un programme pourrait répondre à la question « Quel président est marié à une chanteuse ? ». Le challenge est bien de construire une telle base de connaissances.

Une première approche est manuelle. La base « Cyc [4] » est un exemple d'une telle approche. Elle a collectionné des milliers d'informations élémentaires dans un format accessible par des machines. Cyc peut donc répondre à des questions précises. Par contre, l'étendue de ses connaissances est limitée. On réalise vite qu'il n'est pas facile de rentrer toutes les informations importantes du monde dans une base de connaissances surtout si le nombre des contributeurs est limité. Une première façon de passer à l'échelle est d'automatiser cette acquisition et d'ouvrir au monde entier les moyens de contribuer.

Extraction d'information de Wikipédia

Pour construire des bases de connaissances plus grandes que des bases construites manuellement, on va chercher à transformer le contenu du Web automatiquement dans une base de connaissances. Il s'agit essentiellement d'analyser les textes et d'en « extraire » des connaissances. La tâche est complexe parce qu'elle met en jeu la compréhension de la langue naturelle (par opposition aux langages artificiels à la base des programmes informatiques) avec toutes ses subtilités. Les extracteurs de connaissance font des erreurs et il est difficile de leur en vouloir : ils partent de textes qui fourmillent d'imprécisions et d'erreurs et de faits qui, même lorsqu'ils sont établis, peuvent être controversés et complexes comme « la République de Macédoine a été admise aux Nations unies sous le nom provisoire d'Ancienne République yougoslave de Macédoine ». L'intégration des connaissances de plusieurs sources est aussi délicate, comme l'est la vérification des connaissances obtenues. L'extraction d'information à grande échelle a véritablement commencé avec Wikipédia.

Wikipédia, une encyclopédie libre et gratuite, est un bel exemple d'édition coopérative. Un grand nombre d'internautes collaborent pour développer une encyclopédie. Tout le monde peut participer. Le nom de « wiki-pédia » vient du fait que cette

4. http://www.cyc.com/

encyclopédie repose sur le principe des wikis qui permet à une communauté d'utilisateurs de créer et de maintenir collaborativement les pages d'un site Web. Et cette communauté est impressionnante puisque en 2013, avec plus de 400 millions d'utilisateurs mensuels, Wikipédia est le cinquième site Web le plus visité au monde avec plus de 23 millions d'articles dans de multiples langues. Le chapitre français, lancé en 2001, contient aujourd'hui plus d'un million trois-cent mille articles, ce qui en fait le troisième chapitre pour ce qui est de ce critère. Il est facile d'imaginer la cacophonie résultant des erreurs, des incompétences, des désaccords, des intérêts personnels. La tâche semble impossible. Pourtant, si la qualité de son contenu est parfois contestée, il est passionnant de voir la place considérable prise si rapidement par Wikipédia dans la diffusion des connaissances. Le recours à une foule d'auteurs a permis de dépasser la notion classique d'encyclopédie et d'atteindre une couverture bien plus large des sujets qui intéressent les internautes. Nous trouvons de tout dans Wikipédia depuis la biographie de Clémence Castel, une héroïne de Koh-Lanta, jusqu'à la preuve du Lemme de l'Étoile, un résultat fondamental de théorie des langages. Les erreurs y sont nombreuses mais il y en a aussi beaucoup dans les encyclopédies traditionnelles malgré leur spectre bien plus étroit. Une étude qualitative [5] a ainsi montré que la qualité ne cesse d'augmenter et que sur beaucoup d'articles, elle est maintenant comparable aux encyclopédies classiques tout en étant incomparablement plus diversifiée, plus réactive, plus vivante. Un article qui a des références avérées et est publié depuis longtemps, et donc exposé à la critique, devient mécaniquement plus fiable. Il existe ainsi un cercle vertueux de maintenance qui exerce une pression sur la qualité et la condamne à l'amélioration permanente. D'une certaine façon si un utilisateur voit une erreur sur Wikipédia et ne la corrige pas, il en devient tout aussi responsable que celui qui l'a commise.

Wikipédia est un excellent point de départ pour développer une *vraie* base de connaissances. Plusieurs projets ont commencé à extraire des informations de Wikipédia. Les plus connus sont YAGO, DBpedia, et Freebase.

Typiquement, une page encyclopédique de Wikipédia est dédiée à un sujet, par exemple la Tour Eiffel. Une page peut avoir un contenu très riche, par exemple en détaillant toute l'histoire et les caractéristiques du monument. Plus ce contenu est riche et plus il a de chances de receler des connaissances, qu'il s'agisse, par exemple, de la hauteur ou des coordonnées GPS du monument. C'est notamment le cas de la « boîte d'information » présentée à droite des articles de Wikipédia qui contient des informations déjà formatées. L'idée d'un système comme YAGO ou DBpedia fut donc d'extraire ces données, notamment celles de la boîte d'information. Les connaissances extraites sont stockées et maintenues de façon structurée permettant des traitements

5. http://blog.wikimedia.org/2012/08/02/seven-years-after-nature-pilot-study-compares-wikipedia-favorably-to-other-encyclopedias-in-three-languages/

automatiques. Ainsi, la liste des monuments avec leurs données structurées peut permettre de classer ces derniers par hauteur ou de calculer un itinéraire touristique dans une ville.

Les données de la boîte d'information sont connectées à ce que l'on appelle une ontologie c'est-à-dire, dans ce cas, essentiellement une hiérarchie de classes, dans laquelle chaque objet a sa place. Par exemple, la Tour Eiffel est placée dans la classe des « Tours de Paris », qui est une sous-classe de « Tours », qui est une sous-classe des « Monuments ». Cette hiérarchie est transitive : chaque tour de Paris est une tour, chaque tour est un monument, et donc chaque tour de Paris est un monument. Cette propriété permet des requêtes plus abstraites. Si l'on cherche, par exemple, des monuments qui ont une hauteur supérieure à 300m, on va trouver la Tour Eiffel, même si elle n'a pas été définie comme « Monument » mais comme « Tour de Paris ». Les projets ont suivi des approches différentes pour organiser les objets dans cette hiérarchie. DBpedia a créé sa propre taxonomie de manière collaborative. YAGO a créé la sienne en partant de WordNet, un dictionnaire de la langue anglaise, et des catégories de Wikipédia.

Les bases de connaissances ont grandi de manière impressionnante. En 2013, DBpedia contenait presque 4 millions d'objets. En plus, DBpedia, qui avait commencé avec l'anglais, extrait maintenant des informations des chapitres de Wikipédia dans d'autres langues. Aujourd'hui, DBpedia [6] est disponible en 119 langues. Tous les chapitres ensemble contiennent près de 30 millions d'objets décrits par près de 2,5 milliards de faits. YAGO [7] contient lui 10 millions d'objets et plus de 100 millions de faits sur ces objets. Des évaluations manuelles ont confirmé que les données extraites avaient un faible taux d'erreur (moins de 5%). On possède donc déjà de très riches bases de connaissances d'assez bonne qualité malgré la jeunesse du domaine.

6. http://wiki.dbpedia.org/Datasets
7. http://yago-knowledge.org

Ces travaux de recherche commencent d'ailleurs à être visibles du grand public. En 2010, Google a acheté Freebase, une autre base de connaissances extraites de Wikipédia. Depuis, Freebase a été transformée et renommée en « Knowledge Graph », la base de connaissances qui nourrit la recherche sur Google. Si l'on cherche « Eiffel Tower » sur Google, le moteur de recherche ne nous montre pas seulement toutes les pages Web qui contiennent ces mots. Il nous montre aussi la Tour Eiffel propre, avec une photo, des informations sur son architecture, et d'autres monuments à proximité. Grace à sa base de connaissances, Google a donc « compris » qu'on ne cherche pas les mots « Tower » et « Eiffel » dans les pages, mais bien le concept de la Tour Eiffel.

Il y a maintenant un véritable enjeu d'enrichissement pérennisé et conjoint de l'encyclopédie Wikipédia et des connaissances extraites par différentes initiatives. Il s'agit d'enrichir le contenu de l'encyclopédie des connaissances obtenues, d'aider à détecter les manques, les erreurs aussi bien dans les pages que dans les connaissances extraites. Il s'agit aussi à terme de fournir des nouveaux outils à la communauté de Wikipédia qui est, à l'origine, la source de toutes ces connaissances et de leurs applications.

Extraction de connaissances à partir du Web

Wikipédia est une source très riche d'informations, mais elle n'est certainement pas la seule source potentielle de connaissances. Les pages d'Amazon, les sites personnels, les blogs et les news, par exemple, contiennent des tas de connaissances ignorées de Wikipédia. Pour cela, plusieurs approches ont commencé à extraire de l'information de tout le Web.

Une des premières approches est « KnowItAll ». Ce système parcourt le Web et collectionne des bouts de phrases de la forme « sujet, verbe, objet ». Par exemple, KnowItAll va trouver les mots « les égyptiens », « ont construit » et « les pyramides » dans la phrase « Les égyptiens ont construit les pyramides ». L'approche est assez brutale. Elle va trouver de nombreux triplets qui ne correspondent pas à des faits du monde réel. Par exemple, la base de données va aussi contenir le triplet « Les extraterrestres ont construit des pyramides ». Par contre, KnowItAll va permettre de répondre à des requêtes de la forme « Qui a construit les pyramides ? », ce que les moteurs de recherche du Web actuels ne peuvent pas faire. Ainsi, KnowItAll (et ses successeurs [8]) nous donnent une première idée grossière d'une base des connaissances globale du Web.

Le projet « WebTables » a repris cette vision d'une base de données globale pour l'appliquer aux tableaux et aux listes que l'on trouve sur les pages Web et qui contiennent souvent beaucoup d'information. En recherchant de telles informations de manière systématique, WebTables arrive à construire une belle base de connaissances à partir de millions de listes et de tableaux trouvés sur le Web.

8. http://openie.cs.washington.edu/

D'autres approches se basent sur un principe différent : Il est beaucoup plus facile de faire grandir une base de connaissances existante que de créer une nouvelle base à partir de rien. Les systèmes NELL [9], SOFIE [10], et PROSPERA [11] partent de cette idée. SOFIE et PROSPERA parcourent le Web pour enrichir la base de connaissances YAGO. Les informations obtenues les aident à trouver des connaissances nouvelles. Prenons comme exemple la phrase « Elvis Presley est né en 1935 » trouvée sur le Web. A priori, cette phrase n'est qu'une chaîne de caractères pour le système. Mais YAGO a déjà la connaissance (obtenue de Wikipédia) que « Elvis Presley » et « 1935 » sont dans une relation particulière, « year-of-birth ». Comme ces deux éléments apparaissent dans la phrase, le système conclut que « X est né en Y » peut donner l'année de naissance d'une personne. On appelle cette structure : un modèle (« pattern » en anglais). S'il trouve maintenant sur le Web la phrase « Obama est né en 1961 », le système va comprendre que cette phrase spécifie aussi une date de naissance. Il va donc ajouter ce fait à la base de connaissances.

Les faits que le système ajoute peuvent apparaître plusieurs fois dans le corpus. Le système peut ainsi découvrir de nouveaux modèles, s'en servir pour apprendre de nouveaux faits, et ainsi de suite. On obtient donc un véritable cercle vertueux entre l'extraction d'information et l'apprentissage des nouveaux modèles. C'est une idée décrite déjà en 1999 par Sergei Brin, un des fondateurs de Google. Evidemment, les langues naturelles sont imprécises, les mots peuvent être ambigus (Paris est un nom de ville, un prénom, le nom d'un système informatique, etc.) et surtout le Web contient des faits erronés. Le système peut être amené à faire des erreurs. Il faut arriver à un compromis entre un apprentissage trop agressif (qui découvrira beaucoup de faits en introduisant beaucoup d'erreurs) et un apprentissage trop prudent (qui minimisera les erreurs mais manquera des faits). C'est l'éternel compromis entre deux mesures de qualité pour les systèmes de recherche d'information : le rappel (dans quelle mesure le système trouve les résultats qu'il devrait trouver) et la précision (dans quelle mesure les résultats ramenés par le système sont corrects).

Une tâche complexe est donc bien de vérifier l'information. Par exemple, supposons que le système trouve le fait qu'Elvis est mort en 528. La base de connaissances sait déjà qu'Elvis Presley est né en 1935. Donc le système conclut que le mot « Elvis » dans ce contexte ne peut pas faire référence au roi du Rock and Roll, parce que dans ce cas, Elvis serait mort avant qu'il ne soit né. Le système va donc chercher d'autres significations du mot « Elvis » et peut découvrir Saint Elvis, un évêque qui a vécu au 6ème siècle. En vérifiant que cet Elvis là est né moins de 100 ans avant 528, le système trouve que 528 peut être la date de décès de l'évêque Elvis. Le système ajoute donc le fait que Saint Elvis est mort en 528. Par ce mécanisme,

9. http://rtw.ml.cmu.edu/rtw/
10. http://www.mpi-inf.mpg.de/yago-naga/sofie/
11. http://www.mpi-inf.mpg.de/yago-naga/prospera/

le système peut vérita-blement comprendre et interpréter les phrases, un peu comme son modèle, l'humain.

Le système NELL utilise la même idée du cercle vertueux entre l'extraction d'information et l'apprentissage des nouveaux patterns. NELL combine plusieurs extracteurs, et donc les informations extraites par un d'entre eux peut aider les autres. Le principe de NELL est de continuer sans jamais s'arrêter. NELL , ce qui signifie d'ailleurs « Never Ending Language Learner », extrait des informations sans interruption depuis 2010 et a créé une base de données qui contient déjà des millions d'objets du Web.

Si le Web reste très largement dominée par HTML et le texte, les bases de connaissances de demain sont déjà en construction à partir de l'énorme ressource que constitue la masse de documents textuels. Et demain ? *À côté* des documents textuels, il faut s'attendre à voir proliférer des millions de bases de données ou de connaissances, de toutes tailles et de toutes natures, interconnectées entre elle.

Encore plus de connaissances grâce aux internautes

Nous considérons dans cette section comment trouver encore plus de connaissances notamment en utilisant des informations sur les internautes, comme ce qu'ils aiment, ce qu'ils achètent, ou en les faisant participer directement.

L'évaluation de l'expertise

Une technique essentielle pour évaluer la qualité d'une information est de déterminer la qualité de la source, la confiance que nous pouvons avoir dans les informations que cette source fournit en général. Pour illustrer ce type de techniques, nous mentionnerons un travail récent sur la corroboration [3]. Imaginons un système où des internautes introduisent des connaissances. Ils peuvent se tromper. Pourtant, s'ils ne faisaient que spécifier des connaissances positives comme « Alice possède une 2CV », rien ne pourrait empêcher le système de croire *tout* ce que disent les internautes, y compris toutes leurs erreurs. Pour que le système puisse commencer à douter, il faut que des internautes se contredisent et, pour cela, qu'ils se mettent à publier des informations négatives comme « Alice ne possède pas de BMW ». En général, les internautes ne veulent pas perdre de temps à entrer explicitement de telles informations notamment parce que la liste des informations fausses est bien au-delà de l'accessible. Pourtant, les internautes publient des informations négatives *sans le savoir*. Par exemple, le fait qu'« Alice est née à Romorantin » indique qu'elle n'est pas née à Sèvres, du fait d'une « dépendance fonctionnelle », c'est-à-dire d'une loi que doivent satisfaire les données (ici, la loi qui spécifie qu'une personne ne peut pas être née à deux endroits distincts).

Dans le travail cité précédemment, nous utilisons les informations incluant des négations provenant de dépendances fonctionnelles. Nous estimons la véracité des

connaissances, en déduisons les taux d'erreurs de chaque internaute, ce qui nous procure une meilleure estimation de la véracité des connaissances, d'où des taux d'erreurs plus précis pour chaque internaute, etc. Nous continuons ce processus jusqu'à atteindre un point fixe. Ce genre de technique illustre bien comment il est possible de dégager collectivement des connaissances.

Comme la notation, l'évaluation de l'expertise a sa place sur le Web. C'est en particulier le cas pour ce qui concerne les informations publiées par la presse. Des blogs, comme celui de Maître Eolas pour les affaires juridiques, font maintenant autorité. De simples internautes sont de plus en plus amenés à remplacer les journalistes, comme récemment en Tunisie ou en Syrie. Cela ne rend que plus crucial le besoin de croiser les informations, de les vérifier. Nous pouvons imaginer que demain des programmes participeront à déterminer les réputations en termes d'information dans cet espace-temps du Web qui donne le tournis.

La recommandation

Un système comme Meetic utilise les données fournies par ses clients pour organiser des rencontres, pour les apparier. Un système comme Netflix recommande des films. Pour ce faire, ces systèmes réalisent typiquement des analyses statistiques dans le cadre classique très général de la fouille de données. Ils essaient de mettre en évidence des « proximités » entre clients dans Meetic ou entre clients et produits dans Netflix. Ils peuvent regrouper des personnes parce qu'elles partagent les mêmes goûts même si elles ne se sont jamais rencontrées, ou découvrir des affinités inattendues entre produits. L'exemple souvent cité est que les clients de couches-culottes achètent statistiquement beaucoup de bières. Les classifications des clients et des produits s'enrichissent donc mutuellement et participent ainsi à établir de nouvelles proximités entre individus et produits.

De telles analyses sont réalisées à très grande échelle, par exemple par Amazon ou Google. Réaliser des analyses statistiques de qualité, sur des volumes de données de plus en plus grands, est un des défis du domaine de la gestion d'information.

La notation

Le Web ne grandit pas seulement par la création de pages, mais aussi par la notation et l'enrichissement des pages existantes. L'internaute est invité à noter d'autres internautes, des services, des produits, et participe ainsi à la construction d'une connaissance collective. Par exemple, eBay permet aux acheteurs de donner leur avis sur les vendeurs de sa plateforme (et réciproquement). Cela conduit à une fantastique incitation à fournir un excellent service au risque, sinon, d'être mal noté et de perdre des clients. Les systèmes fourmillent qui utilisent les avis de leurs utilisateurs, comme ViaMichelin pour les restaurants ou AlloCiné pour les films. Notons que, dans ces deux cas, les critiques qui notaient jusqu'alors les restaurants ou les films perdent une forme de monopole.

Les internautes ne notent pas seulement des vendeurs et des films, mais aussi des villes ou attractions (TripAdvisor), des produits (Amazon), des pièces de musique (Deezer), ou des pages Web (Reddit.com, Delicious.com). Ces systèmes de notation ont aussi leur place au niveau global du Web. Par exemple, le système Delicious demande aux internautes d'associer des mots-clés (de la sémantique) aux pages. Encore plus génériques, les mécanismes « I like » de Facebook, « +1 » de Google ou « InShare » de LinkedIn permettent aux utilisateurs de ces plateformes de noter n'importe quelle ressource identifiée sur le Web, générant des quantités de nouvelles données pour ces grands acteurs sur une variété elle aussi grandissante de sujets.

Pour conclure avec la notation, il faut souligner que le Web en tant que toile d'araignée peut être vu comme un gigantesque système de notation. Quand une page référence une autre page, cela peut être compris comme une notation (positive ou négative) pour cette page. Le moteur de recherche Google s'est d'ailleurs imposé en étant le premier à analyser ce graphe gigantesque avec un algorithme appelé PageRank. PageRank calcule la popularité des pages en utilisant ce système de notation, le but étant de donner comme réponses les pages les plus populaires d'abord. À ce propos, il a été dit qu'une société de service délivrait *volontairement* de mauvais services à certains de ses clients pour que ceux-ci en parlent sur le Web et augmentent ainsi la popularité, donc la visibilité, de la société en question. Même si cette information non vérifiée n'est peut-être qu'une des légendes du Web, le fait que la popularité ignore le sens des références est dérangeant. En analysant les liens du Web suivant un système de notation plus riche (avec des notes négatives), ce biais pourrait être corrigé.

Le crowdsourcing

Nous utiliserons ici le terme anglais « crowdsourcing » parce qu'un accord sur sa traduction semble encore manquer et parce que nous n'avons été conquis par aucune des traductions, comme « externalisation ouverte », proposées sur le Web. Il s'agit de publier sur le Web des problèmes que des programmes ne savent pas bien résoudre ; des internautes proposent alors des réponses, typiquement moyennant finance. Des systèmes comme le Mechanical Turk [12] d'Amazon permettent de tels contacts. Les compétences de la foule ont été utilisées par exemple pour rechercher sans succès l'un des plus célèbres chercheurs du domaine des bases de données, Jim Gray, disparu avec son yacht au large des îles Farallon. Les internautes devaient observer des photos satellites à la recherche d'indices. En utilisant un jeu vidéo, Foldit, des internautes sont en revanche arrivés à décoder la structure d'une enzyme proche de celle du virus du sida [3]. Ils ont réalisé ce qui bloquait experts et ordinateurs, comprendre comment cette enzyme se repliait dans un espace en trois dimensions pour construire sa structure. Le jeu se marie ici au réseau, dans le plus pur esprit des réseaux sociaux.

12. Référence au *Turc mécanique*, un automate joueur d'échecs de la fin du 18e siècle, en réalité un canular.

L'originalité de tels dispositifs est que l'individu se retrouve au service d'un système informatique qui l'utilise, par exemple, pour compléter sa base de connaissances ou résoudre des contradictions dans les informations. Cela s'apparente à un nouvel espace de calcul où machines et humains se retrouvent, appelé « human-based computing », qui tend à développer une nouvelle symbiose entre eux. D'une certaine façon Wikipédia préfigure ce nouvel espace puisque l'on oublie souvent qu'un grand nombre d'actes d'éditions sur l'encyclopédie sont effectués par des robots en charge de maintenir son contenu, par exemple en détectant un article n'ayant pas de références ou un ajout qui relève du spam. Et au-delà des données et traitements ce sont les algorithmes eux-mêmes qui peuvent faire l'objet d'une externalisation au Web. Ainsi, la maintenance des logiciels et des plateformes de DBpedia est elle-même le résultat de l'action collective d'une communauté ouverte. Non seulement la plateforme est libre et les codes sources sont ouverts, mais de plus les collections de règles d'extractions tout comme les vocabulaires utilisés pour structurer les données extraites sont maintenus de façon collaborative, là encore dans des wikis. Ainsi la configuration et la maintenance d'une plateforme DBpedia font-elles aussi l'objet d'une activité sociale à laquelle chacun peut participer, par exemple pour ajouter ou modifier des règles d'extractions et ainsi augmenter ou améliorer les données extraites.

Wikidata

Un des projets les plus nouveaux dans le domaine de l'édition collaborative est Wikidata. En effet, en réponse à l'évolution des utilisations et réutilisations de Wikipédia, la Fondation Wikimedia, en charge notamment de l'encyclopédie, a lancé un nouveau projet appelé Wikidata. Il s'agit d'une base de connaissances libre qui peut être lue et éditée aussi bien par des personnes que par des applications. Elle est destinée à fournir des données dans toutes les langues aux projets de Wikimedia et permettre un accès centralisé aux données. Wikidata reprend l'idée de Wikipédia, mais cherche à créer une base de connaissances au lieu d'une encyclopédie. Les internautes sont appelés à fournir des faits individuels pour participer à la construction d'une grande base de données. Cette base de données vise à englober toutes les entités de Wikipédia (comme les chanteurs, politiciens, universités, villes et organisations du monde) et des faits concernant ces entités. Différent en cela de Wikipédia, Wikidata contient cette information sous forme tabulaire. Il va donc être possible de poser des requêtes sémantiques sur les données. En 2013, un an après la naissance du projet, Wikidata contenait déjà 13 millions d'objets et 18 millions de faits. Les deux projets Wikidata et Wikipédia collaborent, et l'idée est de migrer la partie structurée de Wikipédia (les boîtes d'information, les liens entre les langages, et les listes) dans Wikidata. Ce projet témoigne bien de l'importance que les données structurées sont en train de prendre dans la communauté de Wikipédia et sur le Web.

La représentation et distribution des connaissances

Le Web classique est destiné à des humains qui y trouvent l'information qui les intéressent dans des textes, en s'appuyant sur les moteurs de recherche et le butinage (« surfing »). Comme déjà mentionné, les machines préfèrent plus de structures, des données plus typées. C'est pourquoi la tendance est de faire évoluer le Web vers un Web de données, un Web sémantique, à terme un Web des connaissances. Les connaissances (dans des langages formels) y jouent un rôle primordial mais il s'agit avant tout de faciliter la publication de ces connaissances et leur intégration à l'échelle du monde entier — plutôt que d'avoir des milliards d'îlots de connaissance sans liens aucuns. Donc les standards jouent un rôle essentiel dans ces Web en devenir, dans ces Web qui sont déjà là.

La clé de voûte de l'architecture du Web de données et du Web sémantique est la même que celle du Web classique : le nommage et la référence par des URI, pour « Uniform Resource Identifier ». La différence avec le Web des documents textuels est que les URI du Web sémantique ne dénotent pas des documents, mais des objets. On aura donc un URI pour Elvis Presley, un URI pour la France, un URI pour la note FA, un URI pour François Hollande, un URI pour le Poivre, un URI pour la Seine, et ainsi de suite. Ces URI sont liés les uns aux autres. Contrairement aux liens classiques du Web standard, ces liens sont des liens *sémantiques*. La Seine et la France seront reliées par le lien « coule » (La Seine coule en France), et François Hollande et la France par le lien « est le président de ». Il ne s'agit donc plus juste d'offrir un simple moyen de passer d'une page à une autre, mais de procurer un mécanisme suffisamment général pour décrire les liens sémantiques les plus riches entre objets.

Le Web sémantique qui se dégage des recommandations de standards du W3C (consortium de standardisation de l'architecture du Web) dessine une architecture informatique pour l'interconnexion universelle de sources de données et donc forme la base du développement de bases de connaissances à l'échelle du monde. On parle de « Web de données » et de données liées (« Linked data ») parce que cela sert de base pour définir à l'échelle du Web des objets, des relations entre ces objets, leurs attributs, à la manière des bases de données relationnelles utilisées depuis des années massivement dans les entreprises, mais cette fois de façon ouverte et interconnectée dans une architecture Web. On parle de « Web sémantique » par ce que, *à côté des données*, on publie des connaissances formelles qui expliquent ces données, leurs vocabulaires, leurs schémas, et le sens de leurs relations. Ces idées, déjà esquissées dans les années 90 par Tim Berners Lee, étendent les principes fondateurs du Web établis pour des documents textuels à d'autres ressources comme des bases de données et des bases de connaissances. Des techniques développées dans le domaine des bases de données, deviennent applicables mais demandent à être adaptées, et trouvent aussi de nouveaux défis pour le passage à l'échelle de volumes gigantesques, pour

la gestion de la distribution, de l'hétérogénéité des formats, de la fiabilité, la confidentialité, etc. Pour les bases de connaissances aussi il s'agit d'adapter les langages, les algorithmes et les théories considérées jusqu'alors, pour s'adapter aux spécificités du Web (son échelle, sa distribution, sa dynamique, son incertitude, son monde ouvert, etc.). Le Web sémantique est ainsi porteur de nouveaux problèmes, comme l'intégration de nombres impressionnants de sources de données, l'inférence sur ces gros volumes de données, la traçabilité de l'information, etc.

Les standards du Web sémantique

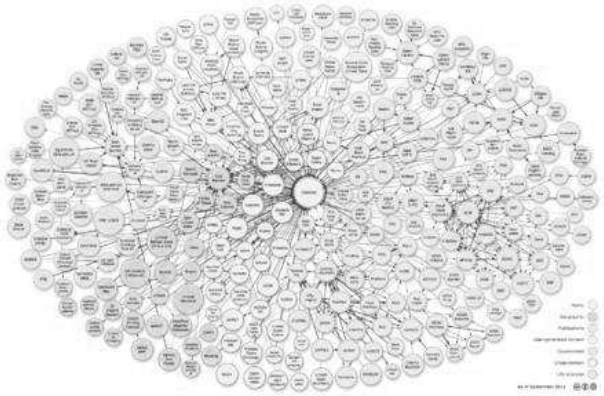
RDF. Le langage RDF (« Resource Description Framework ») est le socle des standards du Web sémantique. RDF fournit à la fois un modèle et plusieurs syntaxes pour publier des données à propos de tout et n'importe quoi sur le Web. Il permet de décrire les objets, comme par exemple l'auteur, la date de création, le titre et les droits de diffusion d'un film. C'est avec RDF que l'on décrit, représente et relie des ressources à échanger sur le Web.

HTTP. Le protocole HTTP offre un mécanisme dit de négociation de contenu. Ainsi, pour un appel à une même adresse URL, un navigateur Web recevra une version HTML en français ou en anglais en fonction de sa configuration tandis qu'un autre logiciel se verra proposer une version RDF qu'il aura la possibilité d'intégrer à sa propre base de données. L'architecture du Web de données est de ce fait complètement imbriquée dans celle du Web classique.

SPARQL. Au-dessus de RDF, le langage SPARQL permet d'interroger à travers le protocole HTTP un serveur distant. Il vise notamment à interroger un graphe de données RDF, peut-être un jour à l'échelle mondiale. Là encore la complexité et les volumes de données posent des défis considérables pour assurer les performances de logiciels qui interrogent et répondent sur ces données.

L'initiative du « Linked Open Data » (données ouvertes liées) encourage la mise en ligne de données libres. Il s'agit bien sûr de publier des données libres mais aussi de les lier entre elles, de faciliter leur utilisation, leur intégration, en utilisant principalement URI, HTTP, et RDF. L'utilisation de SPARQL répond aux scénarios d'interrogation sur des ressources ou sous-graphes précis et le standard LDP en préparation apportera une interface de programmation simplifiée de l'accès aux données. On cherche donc à lier les bases de connaissances dans un graphe global de connaissances. Pour lier les bases de connaissances, on se sert de plusieurs approches. Certaines approches s'intéressent à lier les schémas des données entre eux ; ainsi, le concours annuel « Ontology Alignement Evaluation Initiative » permet aux meilleurs algorithmes de se comparer chaque année. Les approches utilisent par exemple des similarités structurelles, d'usage, linguistiques, etc., pour détecter des liens d'équivalence ou de spécialisation. D'autres approches utilisent en plus la recherche de clés ou de dépendances fonctionnelles (ex. numéro de sécurité sociale) pour aligner les données elles-mêmes et détecter ainsi que deux URI représentent

par exemple la même personne. Le succès de Linked Open Data peut être mesuré : en septembre 2011, les données atteignaient déjà 31 milliards de triplets RDF dans des bases interconnectées sur le Web par 504 millions de liens RDF. En mars 2012 ce sont 52 milliards de triplets RDF qui étaient disponibles.



Visualisation de données liées : chaque nœud représente une base de données et chaque arc correspond aux liens entre deux bases de données liées entre elles. Le nœud le plus central est DBpedia, largement utilisé comme référentiel par les autres bases [13].

La liste suivante hiérarchise les différents aspects de ces données libres et donne les bonnes pratiques pour publier des données liées 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 URL sont utilisés pour identifier sujets, objets, relations et y accéder
⋆⋆⋆⋆⋆	idem + les données sont liées à d'autres données

Pour permettre cette publication de données il faut des systèmes de gestion de bases de données liées performants. Par exemple Virtuoso, OWLIM, RDF-3X, FedX, DARQ ou Diplodocus apportent des réponses différentes en améliorant les techniques de stockage et d'indexation, ou la fédération de bases distribuées. Mais au-dessus de ces systèmes il faut aussi assister l'élévation des données (data lifting), c'est-à-dire aider les producteurs ou gardiens des données à préparer celles qui ont vocation à être publiées sur le Web selon ses règles. Par exemple, la plateforme Datalift.org assiste les manipulations de formats, le respect des règles de publication

13. Le diagramme du Linking Open Data, par Richard Cyganiak et Anja Jentzsch. `http://lod-cloud.net/`

mais aussi le contrôle d'accès et le choix des schémas de données comme nous allons maintenant les introduire.

Vers plus de sémantique

Pour s'assurer de l'interprétation et de l'utilisation qui peuvent être faites des données, on publie simultanément des *schémas* si possible formels qui capturent et expliquent le sens, les catégories, les relations, les contraintes des données et de leur structure. Pour cela, on utilise des langages de description de connaissances plus ou moins expressifs. On parle d'*ontologies* ou de schémas et de langages d'ontologies. Tout est dans le compromis. Un langage plus riche permettra des descriptions plus fines, plus détaillées, mais conduira à des raisonnements plus coûteux (voire à l'indécidabilité de l'inférence). Nous mentionnerons ici les deux principaux langages de représentations d'ontologies sur le Web : RDFS et OWL.

RDFS. RDFS est un langage de schémas très limité en expressivité. Il permet de déclarer et de décrire les types de ressources manipulées (appelées classes, comme par exemple les livres, les films, les personnes, etc.) et les types de relations entre ces ressources, appelées des *propriétés*, comme par exemple « a-pour-auteur » ou « a-pour-titre ». RDFS permet aussi d'organiser ces types dans une taxonomie. Comme on a déjà vu, une taxonomie permet de spécifier par exemple que les auteurs sont des personnes. Des raisonnements relativement efficaces peuvent être réalisés avec RDFS même sur de gros volumes de données et de grosses ontologies.

OWL. Le langage OWL, inspiré des logiques de description, est bien plus expressif et pour cela divisé en plusieurs fragments ou profils d'expressivités différentes pour répondre à des applications différentes. Il permet par exemple la définition de classes en énumérant leur contenu ou par union, intersection, complément et disjonction avec d'autres classes. Voir [1, 5, 6] pour plus d'information sur ces langages et sur les langages d'ontologie en général.

Utiliser ces connaissances

Au cœur de l'utilisation de ces connaissances, on trouve la notion d'inférence. Pour l'illustrer, imaginons qu'une donnée indique que Jorge est l'auteur du livre id28 (« Jorge auteur id28 ») et qu'une autre donnée indique que cet ouvrage est une thèse (« id28 isa thèse »). Nous insisterons sur le fait que Jorge, id28, docteur, ouvrage et thèse sont des entités universelles (identifiées par des URI) et que la chaîne de caractères « Jorge » et les autres ne sont utilisées dans ce texte que pour la lisibilité. Nous soulignerons aussi que les deux faits peuvent être publiés par des sources distinctes et qu'une troisième pourra indiquer la loi :

Forall X (« X auteur thèse » implique « X isa docteur »)

En raisonnant sur les données et leurs schémas, un logiciel, comme par exemple CORESE/KGRAM, SESAME ou WebPIE, pourra déduire que Jorge est docteur et par conséquent inclure son nom dans une requête demandant la liste des docteurs,

même si cette donnée n'est initialement listée nulle part. Les résultats de telles inférences viennent enrichir le Web de données, participent à l'intégration des sources, à la sémantique des requêtes, et ce à l'échelle du Web.

Mais les raisonnements que permettent ces données et leurs schémas ne se limitent pas à de la dérivation logique. La structure de lien permet aussi de nouveaux types de raisonnement par exemple plus métriques que logiques. La centralité, la longueur des chemins, et d'autres métriques de l'analyse de réseaux et de graphes, permettent d'autres exploitations de ces données. Dans l'application DiscoveryHub [8] par exemple, le graphe typé des données liées est utilisé par une version adaptée des algorithmes de propagation d'activation dans les réseaux sémantiques afin de fournir à l'utilisateur des suggestions de nouveaux sujets d'intérêt à partir de ses centres d'intérêts connus, de sujets qu'il veut découvrir ou d'information sur son contexte (ex. sa position, un événement auquel il assiste, etc.). Ici le graphe des données liées est utilisé comme un espace métrique typé à explorer pour suggérer et recommander des sujets de par leur proximité.

De la même façon que le Web classique s'est doté à l'origine d'annuaires et de moteurs de recherche pour nous permettre de trouver et retrouver des sites, les mêmes services sont apparus pour le Web de données. Par exemple, LOV est un annuaire thématique des schémas les plus utilisés pour publier des données, et Sindice est un des moteurs de recherche permettant de trouver des jeux de données et des schémas à partir de mots-clés. Des grands acteurs du Web comme Microsoft Bing, Google Search, Yahoo ! et Yandex, se sont mêmes accordés pour partager les mêmes schémas de base à travers l'initiative Schema.org qui fournit des ontologies pour décrire des créations (Livres, Film, Musique, Recette, etc.), des événements (Concert, Conférence, etc.), des organisations, des personnes, des lieux, des services (Restaurant, Commerces, etc.), des produits et leurs notations, etc.

Pour que le plus grand nombre puisse interroger ces connaissances mondiales, se pose le problème de traiter des questions en langue naturelle. Des systèmes comme QAKIS (INRIA), START (MIT) ou DEANNA (Max Planck) permettent d'interroger les données du Web en formulant des questions directement en langue naturelle. Et aussi, le système Watson d'IBM, qui permet de répondre à des questions en langage naturel, a remporté en février 2011 le jeu télévisé *Jeopardy !*. Sa victoire illustre bien les progrès dans ce domaine et le potentiel de l'utilisation des connaissances du Web une fois les problèmes de passage à l'échelle réglés.

Les données une fois publiées permettent d'imaginer bien d'autres applications comme :

– des données scientifiques échangées entre des instituts en fonction des traitements qu'elles requièrent ;
– des GPS culturels qui guident et informent en se renseignant grâce aux connaissances du Web sur les lieux que nous traversons ;

– des données gouvernementales qui permettent aux citoyens de mieux suivre et participer à la vie des territoires où ils vivent ;
– des portails qui enrichissent des collections culturelles en utilisant les connaissances obtenues sur le Web dans des sources comme Le Louvre ou la BBC ;
– des logiciels d'enseignement, par exemple en histoire de l'art, qui ouvrent vers des ressources proposées sur le Web ;
– etc.

En fait on ne peut que rêver en imaginant toutes les applications possibles.

Conclusion

Nous retrouvons dans toutes ces approches la nécessité de vérifier l'information, de séparer le bon grain de l'ivraie. Par exemple, dans Wikipédia, si votre texte est mal ficelé, s'il ne s'appuie pas assez sur des sources, si son intérêt est discutable, il sera probablement débattu, peut-être rejeté. En général, cela conduit à résoudre des problèmes complexes d'analyse de données impliquant un grand nombre de personnes et de gros volumes d'information. L'évaluation de la « qualité » est au cœur du sujet, la qualité d'une information, la qualité d'une source (un internaute, un service) en général. Et, de plus en plus, l'individu est au centre du dispositif, passivement, par exemple via son profil, ou activement, par exemple en spécifiant ce qu'il sait, ce qu'il croit, ce qu'il aime.

Confronté à des systèmes s'attachant à construire une connaissance collective, l'internaute ignore le plus souvent quelles données ont été utilisées et ne comprend parfois pas comment le résultat a été obtenu. S'il est vrai qu'avec les derniers standards du W3C comme PROV-O (ontologie de provenance), le Web se dote des moyens de décrire la provenance des données et de faire de la traçabilité, il est aussi vrai qu'il reste encore beaucoup à faire pour que de telles informations soient disponibles massivement sur le Web. Quand une réponse résulte de l'utilisation d'une grande quantité d'information (bien au-delà de ce qu'il peut appréhender), un utilisateur peut être amené à trouver les réponses proposées surprenantes, magiques, voire inquiétantes. La difficulté d'expliquer les résultats est une faiblesse souvent présente dans les approches que nous venons de discuter et qui en limite les usages, et pour cela font l'objet de recherches, comme Ratio4TA parmi d'autres.

Un autre problème sérieux de ces approches est lié aux atteintes à la confidentialité de l'information. Pour mieux servir leurs utilisateurs, ces systèmes doivent réunir le plus d'information possible sur eux. Un réseau social comme Facebook construit par exemple une base de connaissances sur chacun de ses membres. L'internaute est de plus en plus souvent amené à fournir des informations pour bénéficier de la gratuité de services. Les systèmes vont même jusqu'à s'échanger des informations sur leur clients ; toujours pour mieux les servir ? Cela conduit à des conflits d'intérêts. Un

système de réseau social doit choisir entre le besoin de protéger les données de ses clients (au risque sinon de les perdre) et son avidité naturelle pour les données confidentielles. De son côté, l'internaute aimerait bien que les informations le concernant restent le plus confidentielles possible, mais il est aussi friand de services très personnalisés. Des initiatives comme WebID, WebACL, et plus généralement le Web social, distribué ou fédéré, cherchent des compromis où les données sont accessibles aux applications mais restent hébergées chez l'utilisateur ou un tiers de confiance. Mais là encore les difficultés ne sont rarement que techniques.

En poussant plus loin le désir de communiquer dans l'anonymat, nous atteignons le « darknet », basé sur des protocoles non standards, le cryptage des messages, et un monde de dissidents politiques et peut-être d'activités illégales. C'est par exemple le système Tor (*The Onion Router*), un réseau mondial décentralisé de routeurs organisés en couches, qui permet de transmettre de manière anonyme des flux de données.

La richesse du Web ne cesse d'augmenter, notamment en termes de diversité de ressources. Le Web se tisse maintenant dans notre monde physique et notre quotidien, notamment à travers l'Internet des objets. Mais en retour, la complexité de notre monde vient augmenter celle du Web, le forçant à intégrer, décrire, modéliser et traiter de plus en plus de choses. Ceci se fait au risque de rompre avec la relative simplicité qui avait accompagné les premières heures du Web et pose résolument la question des interactions avec le Web comme un problème de recherche multidisciplinaire.

Pour conclure, oublions temporairement ces problèmes pour nous émerveiller de voir des algorithmes faire surgir des informations disponibles sur le Web des connaissances dont nous n'imaginions pas l'existence.

Références

[1] S. Abiteboul. *Sciences des données : de la Logique du premier ordre à la Toile*. Collège de France, Fayard, 2012. `http://www.college-de-france.fr/site/serge-abiteboul/`.

[2] S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset et P. Senellart. *Web Data Management*. Cambridge University Press, 2011. `http://Webdam.inria.fr/Jorge`.

[3] Seth Cooper et al. Predicting protein structures with a multiplayer online game. Nature 466, 756–760, 2010.

[4] A. Galland, S. Abiteboul, A. Marian, P. Senellart. Corroborating information from disagreeing views. Proc. of the third ACM International Conference on Web Search and Web Data Mining (WSDM'10), 131–140, 2010.

[5] F. Gandon, C. Faron-Zucker et O. Corby. *Le Web sémantique : Comment lier les données et les schémas sur le Web ?* Collection InfoPro, Dunod, 224 pages, 2012.

[6] F. Gandon, S.-K. Han, R. Krummenacher, I. Toma. Semantic Annotation and Retrieval : RDF. In *Handbook of Semantic Web Technologies*, Domingue, John ; Fensel, Dieter ; Hendler, James A. (Eds.), 117–155, 2011.

[7] F. Suchanek, S. Abiteboul, P. Senellart. PARIS : Probabilistic Alignment of Relations, Instances, and Schema. *Proc. of the VLDB Endowment*, Vol. 5, Issue 3, 157–168, 2011.

[8] Nicolas Marie, Myriam Ribiere, Fabien Gandon, Florentin Rodio. Discovery Hub : on-the-fly linked data exploratory search. À paraître dans *Proc. of I-Semantics*, 2013.

[9] Nicolas Delaforge Fabien Gandon et Alexandre Monnin. L'avenir du Web au prisme de la ressource. Séminaire IST Inria : le document numérique à l'heure du Web de données, 229-252, 2012. `http://hal.inria.fr/docs/00/84/38/33/PDF/delaforgegandonmonnin-v2.pdf`.