Pourquoi les moteurs de recherche ont-ils besoin du web de données ?

Bruno Bachimont, Fabien Gandon, Gautier Poupeau, Bernard Vatant, Raphaël Troncy, Stéphane Pouyllau, Ruth Martinez, Michèle Battisti, Manuel Zacklad

Pourquoi les moteurs de recherche ont-ils besoin du web de données ?

Le moteur de recherche est aujourd'hui présent à tous les niveaux du système d'information de l'entreprise (SIE) : de l'intranet au Web, en passant par les applications de gestion de bases de données. Il est cependant parfois marginalisé, voire peu valorisé, dans les interfaces hommes-machines. Pire, l'éternelle présentation des résultats « en liste » paginée − qui reste souvent le parent pauvre de la réflexion ergonomique d'un tel outil – ne favorise plus aujourd'hui l'accès aux informations des documents, peu ou pas visibles sous cette forme. Cela est en partie dû à l'histoire du Web et à son impact sur nos pratiques. L'évolution du moteur de recherche est intimement liée aux documents qu'il indexe et aux méthodes qu'il utilise pour proposer aux utilisateurs des résultats de recherche « toujours plus pertinents ». Dans mon approche, je laisserai de côté le moteur de recherche du SIE, plus spécialisé et répondant à des besoins particuliers, afin de faire un peu de prospective sur les gains potentiels, pour les moteurs de recherche, du développement du web de données.

Le modèle du moteur de recherche « documentaire »

L'invention du Web et son expansion rapide, dans les années quatre-vingt dix, a vu l'émergence du moteur de recherche « documentaire » qui permet de trouver des documents, au sens de fichiers individualisés. Ces documents, qui ont été placés « dans le Web », sont enchâssés dans les pages HTML et parfois reliés entre eux selon les principes de l'hypertexte. Ainsi, en rupture avec les fonctions de recherche présentes dans les bases de données traditionnelles, travaillant principalement sur de l'information structurée, les outils de recherche ont affronté un Everest : trouver de l'information dans un monde mixte − à la fois structuré (métadonnées) et non structuré (texte intégral) − et utilisant une multitude de formats. Pourtant, la plupart des documents numériques sont structurés. Ouverte ou pas, la structuration des documents s'appuie le plus souvent sur XML pour les documents textuels et sur des formats de codage pour les documents multimédias qui embarquent eux aussi des métadonnées structurées (par exemple les métadonnées International Press Telecommunications Council (IPTC) pour les images). Cette structuration donne un cadre technique aux documents afin de les rendre exploitables par des applications.

Si la structuration technique d'un document textuel peut être utilisée pour faire des traitements comme des conversions ou des mises en page, nous restons là dans une utilisation à vocation informatique des structures des documents qui trouvent rapidement leurs limites pour une utilisation par un moteur de recherche devant répondre à des besoins de plus en plus sémantiques. Structurer un document sur le plan sémantique implique l'utilisation de normalisations et de référentiels communs, comme c'est le cas pour la structuration technique. Les professionnels de l'information et les éditeurs de logiciels ont traduit sous la forme de DTD (Document Type Definition), puis de schémas XML, les besoins de structuration des producteurs d'informations. Mais les outils de moteurs de recherche ont peu utilisé cette structuration pour améliorer leurs performances, du moins en apparence et sans doute en réaction à la complexité du Web faisant intervenir divers formats. Le choix de faire confiance au « tout algorithmique » pour indexer le texte intégral a détourné l'attention des documents pouvant être structurés à la fois sur le plan technique, sur le plan documentaire − par la mise en place ou l'extraction de métadonnées − et sur

le plan sémantique grâce à l'introduction des principes du web sémantique proposés par le W3C. Ainsi, tous les documents allaient être consommés de la même façon par les moteurs de recherche généralistes.

Isidore crée un accès unifié à des données réparties

Le Web évolue, il devient de plus en plus hétérogène et la mise en place du web de données en est la dernière grande évolution. Le Web est non seulement le support des sites mais aussi un espace (au sens d'environnement de stockage, d'édition et de diffusion des données) dans lequel se construisent des territoires pour y stocker à la fois des documents non structurés sémantiquement et des documents contenant une information « structurée », c'est-à-dire une proposition de qualification sémantique de l'information. Dans ce cadre, les outils de recherche doivent s'adapter, changer profondément afin de tirer parti de ces espaces structurés, ouverts et normalisés.
L'utilisation du modèle RDF[1] et des principes du *linked data* ainsi que l'identification des informations par des URI (Uniform Resource Identifier) offrent de nouvelles possibilités pour les moteurs de recherche : l'une des plus évidentes est de pouvoir rapprocher des informations entre elles. Dans le domaine de la recherche scientifique, cela permet d'améliorer principalement l'administration de la preuve scientifique, pour retrouver les relations entre les publications scientifiques et les sources de données (les archives, les fonds documentaires). C'est l'une des ambitions de la plate-forme Isidore[2], développée par le Très Grand Équipement (TGE) Adonis[3] et avec l'aide du Centre pour la communication scientifique directe (CCSD)[4], deux équipes du CNRS. Isidore est une solution de traitement de l'information scientifique pour les sciences humaines et sociales (SHS) qui collecte, normalise, enrichit et diffuse données et documents de la recherche. Ses missions sont multiples : créer un accès unifié à des données réparties, qualifier et relier des données ou encore placer les documents et les données numériques des SHS dans le web de données.

Isidore repose sur les principes du web de données et du *linked data*, permettant ainsi à son moteur de recherche d'indexer des informations reliées à des référentiels métiers (exprimés en Skos/RDF). Par la mise en place d'une chaîne de traitement de l'information proposant des normalisations, des enrichissements sémantiques et des catégorisations automatiques, Isidore offre aux chercheurs la possibilité de suivre l'évolution des disciplines et d'en explorer les marges afin de repérer de nouvelles questions. L'apport des méthodes du web de données (format pivot RDF, référentiels en Skos/RDF, utilisation d'URI) constitue un environnement de base permettant au moteur de recherche de travailler sur une assiette plus large de documents structurés. RDF étant au centre du projet, les données sont aussi réutilisables *via* un SPARQL *endpoint*[5]. Ces méthodes, réintroduisant l'information structurée au cœur des données, offrent aux moteurs l'opportunité de diversifier les modes de représentation de l'information : visualisation, frises chronologiques et temporelles peuvent être proposées en complément des résultats en liste. Cela veut dire aussi que les moteurs de recherche doivent être souvent complétés par des chaînes de traitement en amont.

Si les moteurs utilisent depuis longtemps différents gisements de données, l'information

---

[1] Lire l'article p.XXX.
[2] http://www.rechercheisidore.fr
3 Voir : http://www.tge-adonis.fr
4 Voir : http://www.ccsd.cnrs.fr
5 Un SPARQL *endpoint* est une interface d'interrogation d'une base de données RDF (ou *triple store*) utilisant le langage de requête SPARQL, cf. l'article p.XXX.

structurée en RDF placée directement dans le Web leur permet d'indexer des informations complexes réparties (et non plus seulement des documents) afin de proposer des contenus reliés et enrichis. L'indexation du texte intégral, si celui-ci est enrichi d'une structuration RDF de l'information, peut largement améliorer les capacités d'un moteur en matière de pertinence (validation des informations à l'aide de leurs relations) et d'enrichissement (qualification des informations, expansions sémantiques à l'aide de référentiels structurés). C'est justement la proposition que fait Isidore : collecter, enrichir, donner accès et rendre réutilisables les données.

Des moteurs de recherche « sémantiques »

Le web de données, couplé à des moteurs de recherche capables de tenir compte des principes du *linked data* et d'exploiter les documents et données modélisés avec RDF, prend donc une couleur sémantique. Une couleur seulement, car nous ne sommes qu'au début de cette évolution du Web pour de grandes masses de données, dont une partie se déroule dans le cadre du mouvement de l'*open data*. Si de nombreux projets prennent la voie du web de données et du *linked data*, il faut travailler à ce que ces grandes masses de données libérées, parmi lesquelles les données publiques réutilisables, puissent être exploitées par de véritables moteurs de recherche sémantique, afin de pouvoir en tirer toute la richesse informationnelle.●