Journée d'étude FULBI sur le Web de données

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 18 janvier 2013

https://www.bortzmeyer.org/fulbi-web-donnees.html

La FULBI <http://www.fulbi.fr/> est une association de bibliothécaires et autres professionnels de l'information et de la documentation. Elle organise tous les ans une journée d'études <http://www.fulbi.fr/?q=content/les-journ%C3%A9es-d%C3%A9tude-de-la-fulbi> sur un thème donné et le thème de la réunion du 17 janvier 2013 <http://www.fulbi.fr/?q=content/journee-fulbi-du-17-j était le Web des données.

Il y avait plusieurs exposés très intéressants. Emmanuelle Bermès <http://www.figoblog.org/> (Centre Pompidou) a joué le rôle de la formatrice : expliquer le Web des données pour ceux et celles qui ne seraient pas déjà dedans. Son accroche était que l'usager est comme un petit enfant, très impatient, et qu'il faut lui donner l'information très vite, et lui apporter là où il est (allusion probable au fait que certains professionnels de l'infodoc avaient snobé le Web pendant longtemps, préférant leurs catalogues rigides et leurs systèmes informatiques non-interopérables. « Le Web est un modèle d'interopérablité. On ne s'en rend plus compte parce qu'on est habitué, mais c'est rare en informatique »).

Ensuite, pour accéder à cette information, il faut qu'elle ait pu être digérée par les machines. « Vous regardez une page Wikipédia, instinctivement, vous la structurez, vous repérez le menu, le titre, etc, le tout en analysant la mise en page, ce que la machine ne peut pas faire ». D'autre part, face à un formulaire de recherche, le moteur de recherche est face à un mur, il ne sait même pas quelle question poser. Il faut donc donner accès au catalogue. Le Web de données vise à répondre à ces deux questions, structuration de l'information et accès au catalogue, permettant d'énumérer toutes les données (problème dit du Web profond).

Au passage, l'oratrice avait promis d'expliquer la différence entre Web de données et Web sémantique mais j'ai raté cette explication. Elle note à juste titre que le Web sémantique n'a rien de sémantique : pas d'IA, pas de traitement automatique des langues.

Emmanuelle Bermès a longuement parlé des URI (une de ses spécialités, pour laquelle elle avait fait un excellent exposé à JRES <http://2007.jres.org/planning/paper69c5.html?pid=163>), puis de RDF (mais ne vous inquiétez pas RDFS/OWL et SPARQL étaient aussi cités). Les URI sont vitaux en RDF car chaque entité (personne, lieu, concept, etc) est identifiée par un URI (qui doit être

pérenne : contrairement aux vendeurs de vent comme ceux des DOI, elle affirme que les URI peuvent être pérennes). « Le bonheur, l'amour (et le malheur aussi) ont des URI ».

Autres technologies mentionnées et qui semblaient intéressantes : Open Graph Protocol et le site de dépôt de schémas schema.org <http://schema.org/>.

(Deux avis personnels : pour le problème du Web profond, le plus simple est de traduire automatiquement toute la base de données en HTML avec liens, et on n'a plus qu'à laisser faire le moteur de recherche. Et, pour RDF, c'est au Web ce qu'est IPv6 au réseau : des années qu'il existe et qu'il va vraiment décoller « l'année prochaine ». J'ai posé la question des microformats à l'oratrice, mais elle ne les voit pas vraiment comme une alternative à RDF, car ils permettent moins de choses.)

Lionel Maurel <http://scinfolex.wordpress.com/>, lui, a parlé (sans diapos) de la question du droit des données publiques. Je vais essayer de résumer mais attention, je ne suis pas juriste donc les erreurs dans ce résumé sont les miennes, pas celles de Lionel Maurel.

L'"*Open Data*" n'a pas de cadre juridique particulier pour l'instant. Le cadre juridique de l'accès aux informations publiques est la loi du 17 juillet 1978 (celle qui créait la CADA). Les bases non publiques dépendent du droit des bases de données (qui ressemble au droit d'auteur). La loi de 1978 pose le principe de la réutilisabilité, par défaut. L'administation n'a pas le droit de refuser la réutilisation des données (y compris à des fins commerciales). Ce principe est très fort mais, en pratique, est limité par des tas de choses :

— documents secrets (exemple Défense Nationale),
— données produites par les EPIC comme l'INA ou l'IGN (justement ceux qui ont des données...),
— données personnelles,
— l'État peut faire payer, au tarif qu'il décide (pas de régulateur pour le prix),
— défense d'altérer les données (je trouve ce point très contestable, car il revient à donner un droit de veto contre toute modification),
— le droit d'auteur garde toujours le dessus (une bibliothèque publique ne peut pas distribuer en "*Open Data*" les livres récents qu'elle vient de numériser).

D'autres restrictions à ce principe de publication sont plus consensuelles :

— citer la source,
— pas d'**obligation** de publier activement. Si les données sont sur du papier dans le sous-sol, l'État n'est pas obligé de les numériser.

Lionel Maurel a résumé cela en disant que le droit permettait à un établissement public de faire presque tout ce qu'il voulait. Ceux qui sont dynamiques et ouverts peuvent publier. Les rétrogrades fermés ne sont pas obligés de publier.

Et les licences ? Celles du logiciel libre ou les CC n'étaient pas parfaitement adaptées. La ville de Paris a utilisé ODBL (licence "*copyleft*"), plus adaptée aux bases de données mais qui ne mentionne pas la loi de 1978. La licence IP du Ministère de la Justice français est, elle, explicitement adaptée à la loi de 1978. Idem pour la Licence Ouverte, d'Etalab. Et, pendant ce temps, Creative Commons a créé CC0 (équivalent au domaine public).

À noter que l'État (mais pas les collectivités locales) est, lui, obligé de publier depuis une circulaire Fillon, et via Etalab. Parmi les exceptions à cette obligation (le droit, c'est comme le français, c'est plein d'exceptions) :

— Le cas où les données sont vendues (Légifrance),
— Les données « culturelles », ce qui permet au Louvre de ne pas diffuser les siennes.
— Le cas de la recherche publique n'est pas clair...

Enfin, une dernière faiblesse du dispositif légal français, il n'y a pas d'obligation d'un format ouvert ou même interopérable. C'est ainsi que la grand majorité des données sur <http://data.gouv.fr/> sont dans un format purement Microsoft.

Au niveau européen, la directive PSI est en cours de révision (et pourrait être publiée dans les mois qui viennent). Elle prévoit l'obligation de formats ouverts, et une limite au prix qu'on peut faire payer (le coût marginal).

Lionel Maurel a terminé par un hommage à Aaron Swartz, militant des données ouvertes.

Troisième exposé que j'ai beaucoup apprécié, celui de Stéphane Pouyllau (CNRS) sur son expérience d'informaticien au service des chercheurs en SHS pour les aider à publier leurs données (pas les articles, les données qui sont à la base de ces articles). À l'heure actuelle, en SHS, les données sont typiquement en vrac sur portable du chercheur, ou, dans le meilleur des cas, classées selon ses principes à lui et donc peu réutilisables (l'orateur a dit cela plus diplomatiquement). Chaque laboratoire ou MSH a développé un outil de documentation, utilisé uniquement en interne. Il y a donc bien trop d'outils. D'une manière plus générale, il y a peu d'interaction entre les chercheurs (qui aiment bien développer un outil NIH) et les professionnels de la documentation. (≪ On n'a pas pensé à leur demander ≫ a dit un chercheur interrogé, après que l'orateur lui ait demandé pourquoi diable il avait utilisé Drupal pour bâtir un catalogue bibliographique, au lieu d'outils plus adaptés.)

L'orateur a plaidé pour une structuration (il disait plutôt ≪ normalisation ≫) des données à la base. Vaste chantier.

Enfin, j'ai aimé l'exposé de Romain Wenz sur <http://data.bnf.fr>, le service de distribution de données structurées de la BNF, avec une jolie démonstration en utilisant le pirate/écrivain/cartographe Exquemelin (sans doute le seul pirate dans le catalogue BNF...) Les données incluent notamment des liens entre auteurs, œuvres... data.bnf.fr, c'est ≪ Facebook pour les morts ≫. (Voici la fiche d'Exquemelin <http://data.bnf.fr/12052075/alexandre-olivier_exquemelin/>.)

L'orateur a insisté sur l'importance de connaître son public, qui fait ainsi souvent un usage déroutant des données mises à sa disposition. C'est ainsi que 25 % des lecteurs du Roman de la Rose sur Gallica déclaraient qu'ils cherchaient des idées de déguisement (le livre est riche en illustrations de costumes).

Par contre (attention, à partir d'ici, je suis négatif), l'exposé de Caroline Goulard (Dataveyes) était nettement moins intéressant : mauvaise oratrice (cafouillage à la mise en route de la vidéo de pub, vidéo elle-même sans intérêt, récriminations contre le navigateur qui lui semblait trop lent, etc), vocabulaire approximatif ("*open data*" pour parler de toute entreprise qui donne accès à ses données, même avec une licence ultra-restrictive), et très peu de visualisation de données, ce qui était pourtant le titre (prometteur) de son exposé. Par contre, j'y ai découverte les jolies visualisations du New York Times comme celle sur le budget états-unien <http://www.nytimes.com/packages/html/newsgraphics/2011/0119-budget/>.