L'Abes sur *le web de données*

C'est dans une démarche très progressive, empirique et pragmatique, qu'à partir de 2007, l'Abes a fait ses premiers pas sur le web de données.

Au départ, le principal objectif était d'améliorer le référencement des données par les moteurs de recherche, tout en alimentant une réflexion à plus long terme sur l'interopérabilité et la « réutilisabilité » de ces données, au-delà des formats métiers traditionnels. C'est donc par l'exposition de celles-ci que tout a commencé.

IDENTIFIER : LA PREMIÈRE BRIQUE

La première étape a consisté à identifier nos ressources en tant que telles. Car si nos catalogues sont riches en identifiants, internes ou normalisés (PPN, ISSN, ISBN, NNT…), ils ne permettent pas, à eux seuls, un référencement par les moteurs de recherche puisque ces derniers parcourent essentiellement des pages web en sautant de lien en lien. Il fallait donc, avant tout et pour chacun d'eux, identifier chaque ressource ou notice par des URL ou URI stables, construites sur les identifiants internes et à partir desquelles une redirection permet d'assurer l'affichage de la page de résultats correspondante. Puis il s'agissait de les lister systématiquement dans des « *Sitemaps* », pages à l'usage des robots d'indexation, leur permettant ainsi de les référencer comme des pages web.

Après être descendu au niveau des notices, l'étape suivante devait donner lieu à une meilleure indexation du contenu de ces pages. C'est à partir de cette étape que nous sommes donc véritablement entrés dans le web de données, lequel s'appuie, précisément, sur des URI pour identifier et décrire le contenu de façon à pouvoir être lu par une machine. Dès lors, il était naturel de s'intéresser à ce standard émergent du W3C qu'était RDF.

L'APPRENTISSAGE DE RDF

C'est avec Calames, en 2008, que nous avons commencé à distiller, l'air de rien, des métadonnées en RDF sous sa forme encore la plus répandue : du RDFa, c'est-à-dire des triplets encapsulés dans des balises cachées du code HTML. Ces balises, ignorées par les navigateurs, peuvent être moissonnées par des « parseurs[1] » spécialisés ainsi que par les moteurs de recherche. C'est également une solution de moissonnage des données, alternative à OAI-PMH. C'est d'ailleurs celle retenue par le portail Isidore[2] jusqu'à aujourd'hui.

Avec IdRef, ouvert en octobre 2010, nous avons choisi une exposition distincte de l'interface publique, en exposant des fichiers RDF/XML. Pour les récupérer, nul besoin de parser du code HTML : avec un navigateur, il suffit pour récupérer le fichier d'ajouter à l'URI de la notice l'extension **.rdf**. En voici un exemple : **http://www.idref.fr/033702462.rdf**

Et pour un programme, il est possible avec l'URI seule de demander ce fichier RDF, plutôt que la redirection par défaut vers la page HTML, grâce à la négociation de contenu dans la requête HTTP[3].

Le portail theses.fr, ouvert l'année suivante, a retenu les deux méthodes : RDFa, et RDF/XML. Et finalement, le Sudoc a fait à son tour son apparition sur le web de données. À ceci près que pour ce dernier, nous avons fait une entorse à la standardisation en proposant également dans des pages HTML spécialement destinées aux robots des microdonnées schema.org, promues par les principaux moteurs de recherche.

RDF INSIDE

Les données exposées collent encore de près aux formats de production sur lesquels elles s'appuient, tout en étant incomplètes et parfois bancales. En effet les vocabulaires les plus répandus sont mal adaptés aux données natives, et les vocabulaires « métiers » (ISBD, RDA, FRBR) ne sont pas toujours bien adaptés au web de données. Surtout, les données sont générées dynamiquement à partir des bases de production, auxquelles elles sont par conséquent étroitement asservies.

Mais ces expériences ont permis de monter en compétences et d'être aujourd'hui plus ambitieux.

Avec le hub de métadonnées, nous avons commencé à interroger et à manipuler les données en RDF dans une base autonome, afin d'explorer plus directement le potentiel des graphes et du langage SPARQL. Enfin, RDF est un bon candidat pour un futur entrepôt de métadonnées ouvertes, synchronisées entre elles, mais aussi avec les bases de production, ainsi qu'avec des référentiels et sources externes.

Michael Jeulin
Expert métadonnées, Abes
jeulin@abes.fr

[1] Programmes permettant de récupérer dans une structure de données – XML, RDF- des contenus (de balises par exemple) et de les rendre accessibles.

[2] Plateforme de recherche et d'accès aux données numériques et numérisées du domaine des sciences humaines et sociales. **http://www.rechercheisidore.fr/**

[3] Possibilité, pour une même URI, de proposer plusieurs versions d'un document. Exemple avec l'interface cURL: curl-H «Accept: application/rdf+xml» **http://www.sudoc.fr/157385477**