

Linked data, enjeu(x) et devenir

Les technologies du *Linked data* offrent de nombreuses possibilités à tous les producteurs et utilisateurs de données. Michael Jeulin, Aline Le Provost et Yann Olivier, qui connaissent bien la question à l'Abes, font le point sur ces perspectives élargies.

Les bibliothèques pratiquent depuis longtemps l'échange de données catalographiques. Elles savent que toute collaboration implique l'adoption de normes communes faisant en sorte que ce qui est échangé soit utile à tous : un socle commun bibliographique transportable et relativement fiable. Bien que robuste, ce système présente néanmoins un inconvénient majeur, celui de supporter difficilement les extensions et les spécialisations.

Le web de données permet, au contraire, d'exprimer et de partager des données produites de façon hétérogène. L'enjeu descriptif s'attachant autant au vocabulaire utilisé qu'à la chose décrite, le point de vue sur la ressource décrite devient aussi important et accessible que la ressource elle-même. Grâce aux identifiants uniques d'une part et aux ontologies d'autre part, l'interopérabilité est devenue une réalité prometteuse. Chaque métier, chaque organisme, chaque communauté peut produire et partager un référentiel (IdRef pour les chercheurs français, RNSR pour les structures de recherche, MeSH pour l'indexation matière médicale...) ou un vocabulaire (Bibframe, PRESSoo...), qui, dès lors que les expertises sont reconnues, feront foi et seront réutilisés.

L'autre grand intérêt du web de données est de faciliter l'accès aux contenus. L'ajout de métadonnées descriptives dans les pages HTML est un premier pas que beaucoup de sites ont aujourd'hui franchi. Les moteurs de recherches généralistes indexent ces données et les exploitent pour fournir à leurs utilisateurs de plus en plus d'information structurée – qui n'a jamais rencontré le *knowledge graph* de Google ?

Les possibilités sont pourtant loin d'être atteintes, en particulier dans le domaine de la publication scientifique. Si une partie des éditeurs fournit des notices en XML, ces dernières restent difficiles à exploiter et doivent être retraitées ; c'est ce que le hub de métadonnées s'attache à faire. Et s'il faut saluer les efforts de certains groupes commerciaux comme Nature, qui a publié en *Linked data* l'ensemble des titres de sa plateforme, on peut espérer la généralisation de ces efforts, notamment du côté des acteurs de l'*open access*.

L'exposition native en RDF d'un nombre croissant de sources permettrait de faire émerger de nouveaux outils de découverte, qui sachent vraiment intégrer des données hétérogènes, au lieu de juxtaposer des notices XML.

DE L'EXPOSITION AU « TRIPLESTORE »

Mais le web de données est bien plus qu'un moyen d'exposer ou d'exporter les données. Les travaux de la BNF, du Centre International de l'ISSN et de l'Abes donnent à voir des cheminements similaires : la modélisation et l'exposition ont été d'abord l'occasion d'apprendre en marchant, de se familiariser avec la structure en réseau des graphes, très différente de celle des bases de données relationnelles. Chemin faisant, on a commencé (et ce n'est pas rien pour des bibliothécaires !) à se défaire du concept de *notice*, pour mieux constater que celle-ci est elle-même un document, juste une vue parmi d'autres sur des triplets décrivant une ou plusieurs ressources représentées par des URI.

Mais s'arrêter là aurait un goût de trop peu. Après avoir proposé du RDF pour les autres, la suite logique est d'essayer d'en exploiter soi-même toutes les potentialités : pour cela, il faut les charger dans une base RDF interne, et se mesurer à la puissance et à la souplesse du langage SPARQL. Celui-ci est un outil d'administration avec lequel les bibliothécaires ou documentalistes peuvent explorer et manipuler très finement leurs métadonnées, effectuer des croisements et des inférences complexes, interroger des sources hétérogènes, internes ou externes, qui peuvent y coexister à volonté. Et tout cela, en s'affranchissant, dans une certaine mesure, du recours aux informaticiens.

Il permet ensuite d'utiliser au maximum la logique des alignements pour connecter de façon cohérente toutes ces sources, et *in fine* de s'appuyer sur eux pour enrichir ou compléter ses propres contenus. Le SPARQL Endpoint de démonstration élaboré par l'Abes donne un aperçu, partiel, de ce genre d'« atelier ».

Cette articulation ne va cependant pas de soi, car les formats actuels de production de métadonnées (Marc, EAD...) obéissent à des logiques et des structures qui n'ont pas été pensées pour le web de données. Le « *triplestore* » est l'environnement où l'on peut construire et tester de nouveaux modèles hors des contraintes des bases natives, en particulier pour accompagner la transition bibliographique vers les modèles FRBR, ou FRBRoo, et pour livrer un cadre commun à la documentation, aux archives et aux musées. C'est là que des initiatives comme Bibframe, PRESSoo ou le projet « Records in context » du Conseil international des archives donneront toute

[1] <http://bibframe.org/tools/editor/#>

leur mesure, ou devront évoluer encore. Enfin, le web de données ouvre de nouveaux horizons aux professionnels de l'IST, en leur donnant l'opportunité de valoriser leurs compétences traditionnelles, comme la création de classifications et de thésaurus, la normalisation et la structuration des informations. Bref, à ce stade, la base RDF tend à devenir une usine de retraitements de métadonnées provenant de multiples bases de production et de référentiels, pour y retourner ensuite, enrichies. C'est déjà beaucoup, mais ce n'est pas forcément tout. Plusieurs questions se posent : comment ces données sont-elles produites par les utilisateurs professionnels en amont ? A quoi ressembleront les interfaces qui permettront à l'utilisateur final d'explorer et d'interroger ces données ?

QUELLES INTERFACES POUR LES DONNÉES RDF ?

Le catalogueur de demain n'aura pas pris RDF en LV1. Il serait absurde de l'imaginer « cracher » du RDF brut... aussi absurde que d'imaginer le catalogueur d'aujourd'hui saisir du Marc au kilomètre... ce qu'il fait pourtant...

Si on ne veut pas saisir du code, RDF ou Marc, il faut bien concevoir des interfaces permettant de saisir les données sans pour autant en maîtriser la syntaxe qui sert à les échanger ou les stocker. Bref, il faut un formulaire. Mais il y a formulaire et formulaire. Le défi est de concevoir des formulaires orientés « clavier » et non « souris » : le recours à la souris ralentit et exaspère ; il faut pouvoir saisir au kilomètre dans un formulaire... sinon, on préférera saisir du code. C'est un défi qui n'a rien de propre aux données RDF. Par contre, grâce au caractère auto-descriptif des données RDF, le formulaire peut être généré automatiquement : un schéma RDF sait que la propriété « marcel:aut » s'affiche avec le libellé « auteur » en français et prend pour valeur l'identifiant d'une personne, que l'on peut ensuite interroger par son « foaf:name » dans un référentiel de personnes. Ce genre de mécanismes permet de donner plus d'autonomie aux bibliothécaires par rapport aux développeurs et favorise la collaboration entre experts données et « bibliothécaires système ». Le projet Bibframe propose ainsi un éditeur web qui peut servir à produire du RDF/Bibframe, ou n'importe quel autre vocabulaire¹.

Comme pour les interfaces de saisie, les interfaces qui permettent d'interroger ou d'explorer les données RDF sont censées cacher la syntaxe. Mais comment faire pour rendre le RDF invisible tout en mettant entre les mains de l'utilisateur toute la puissance de SPARQL ? Le Graal serait de guider l'utilisateur de sorte que, pas à pas, sans le savoir, intuitivement mais rigoureusement, il construise une requête de plus en plus complexe. C'est exactement l'ambition de SPARKLIS, prototype conçu et développé par Sébastien Ferré, chercheur à Rennes. SPARKLIS est

une interface web qui permet de faire du SPARQL sans le savoir. Elle est la rencontre entre une interface de recherche à facettes, un constructeur de requêtes interactif et une interface en langue naturelle. Voici un exemple de construction pas à pas d'une recherche sophistiquée qui correspond à une requête SPARQL :

step	query
1	Give me something
2	Give me a Writer
3	Give me a Writer that has a nationality
4	Give me a Writer that has nationality Russians
5	Give me a Writer that has nationality Russians and that has a birthDate
6	Give me a Writer that has nationality Russians and whose birthDate is after 1800
7	Give me a Writer that has nationality Russians and whose birthDate is after 1800 and that is the author of something
8	Give me a Writer that has nationality Russians and whose birthDate is after 1800 and that is the author of a Book
9	Give me a Writer that has nationality Russians and whose birthDate is after 1800 and that is the author of a number of Book
10	Give me a Writer that has nationality Russians and whose birthDate is after 1800 and that is the author of the highest-to-lowest number of Book
11	Give me a Writer that has nationality Russians or something and whose birthDate is after 1800 and that is the author of the highest-to-lowest number of Book
12	Give me a Writer that has nationality Russians or Russian_Empire and whose birthDate is after 1800 and that is the author of the highest-to-lowest number of Book

Sparklis - Complex queries and exploratory search on SPARQL endpoints at your fingertips!
 10 W in 2016: nested aggregations, and computational/litering expressions !!

SPARQL endpoint:

Your query and its current focus: **personalist**

give me every **writer**
 whose nationality is
 Russians **or** Russia **U**
 and whose birth date is after 1800
 and that is the author of the highest-to-lowest book
 and for each result
 give me the number of **books**

Sparklis suggestions to refine your query:
 The current focus is **personalist** (click on different parts of the query to change it)

matches all of: matches all of: matches all of:

no concepts no entities 3 modifiers

Results of your query

results 1 - 10 of 13 Show 20 results

	writer	birth date	number of book
1	Shirvan Vrochek	1976-11-01	1
2	Fyodor Dostoyevsky	1821-11-11	11
3	Vincenty Zarynkin	1884-02-01	1
4	Nikolai Ostrovsky	1904-05-29	2
5	Joseph Brodsky	1940-05-24	1
6	Aleksey Konstantinovich Tolstoy	1817-09-05	4
7	Vasily Akayonov	1912-08-20	1

Pour faire joujou avec SPARKLIS et les données du démonstrateur Abes, c'est ici : http://tiny.cc/sparklis_abes

CONCLUSION

Les données des bibliothèques sont des données comme les autres. Avec le web sémantique, les bibliothécaires échappent à une logique de niche, adoptent une approche et des techniques universelles. C'est bon pour nos ressources humaines : nous capitalisons sur nos compétences professionnelles pour leur donner une portée plus large et oser l'expérimentation. C'est bon pour nos marchés : de nouveaux prestataires peuvent nous proposer des produits, des services et du conseil. C'est bon pour le moral : créolisons les données !

MICHAEL JEULIN
jeulin@abes.fr

ALINE LE PROVOST
Le-provost@abes.fr

YANN OLIVIER
nicolas@abes.fr

Experts métadonnées à l'Abes