Conférence satellite IFLA 2014
"Le web de données en bibliothèque, du projet à la pratique"

14 août 2014, Bibliothèque nationale de France, Paris

Compte rendu de Gildas Illien

Les participants ciblés par cette manifestation étaient au rendez-vous tant par leur nombre (environ 140 personnes) que par leur origine (diversité géographique et professionnelle, experts du web sémantique ou professionnels motivés, avec 56 organisations étrangères ou internationales représentées).

Ce compte rendu fait la synthèse des tendances et idées importantes qui se sont dégagées des présentations et discussions de cette journée dont l'objectif n'était pas tant de convaincre de l'importance du web sémantique pour les bibliothèques (le public présent en était déjà convaincu) que de partager des retours d'expérience et de bonnes pratiques sur sa mise en œuvre dans des projets concrets.

- ## Des notices aux entités : les autorités en première ligne

Le discours sur l'évolution des catalogues évolue profondément dans le contexte du web sémantique. Conséquence assez logique de l'impact du modèle FRBR d'une part (qui induit un autre découpage et une « autonomisation » des informations de la notice), et de l'insertion des bibliothèques dans l'environnement du web d'autre part, on parle de moins en moins de notices bibliographiques et de plus en plus d'autorités ou « d'entités ». Dans la mise en pratique d'un projet *Linked Data*, l'accent est ainsi mis sur la production, l'exposition et la mise en relation des autorités; l'expression des relations entre ces concepts (selon la syntaxe des triplets RDF sujet/prédicat/objet) se fait à un niveau de granularité plus fin, celui des *éléments de la notice*, et non plus celui de la notice elle-même.
Cette évolution est notamment perceptible dans le discours d'OCLC qui, pour résumer sa stratégie d'innovation, parle désormais de « *Entity-based Data Strategy* » plutôt que de « *Linked Data* ». Interrogé sur ce point, Teodore Fons (OCLC) précise que le *Linked Data* n'est finalement qu'une forme de technologie (d'autres suivront) tandis que **le changement fondamental qui s'opère du point de vue des données de bibliothèques** consiste bien à **déplacer la priorité de la description bibliographique de documents vers l'enrichissement et l'exploitation des données d'autorité ou entités**.

Il a donc été beaucoup question d'autorités au cours de la journée, chacun des grands fichiers d'autorité propres aux bibliothèques étant examiné. La mise en relation de bases d'autorité relatives aux **personnes et collectivités** fait pivot dans la plupart des projets, VIAF (Virtual International Authority File) étant le fichier collaboratif le plus fréquemment cité.
La problématique de la constitution de **bases d'œuvres** (issues des notices de titres, mais qu'il faut étendre à tous les types de documents pour le bon fonctionnement du regroupement FRBR) a également été largement évoquée, à travers des expériences aussi diversifiées que celles de la National Diet Library (Japon) ou d'Electre. Electre a en effet présenté des cas d'usage auxquels son modèle œuvre doit répondre, et qui est conçu pour satisfaire des besoins propres aux acteurs du monde de l'édition : par exemple, répercuter rapidement les informations relatives aux prix littéraires sur un ensemble d'éditions, ou les évolutions du prix des livres dans le contexte de la multiplication de leurs éditions (physiques et surtout électroniques, avec le développement des e-books).
Dans le cas des personnes ou des œuvres, tous les intervenants ont insisté sur l'importance de disposer **d'identifiants fiables, pérennes, normalisés et actionnables** indispensables à la mise en relation par alignement des référentiels utiles. Pour les auteurs et les collectivités, il a été beaucoup question de l'identifiant ISNI, qui gagne progressivement en notoriété et en confiance dans les communautés du web sémantique.

Une des nouveautés de cette journée a concerné l'**exploitation des vocabulaires contrôlés**, autrement dit les référentiels de mots-matières et autre langages d'indexation documentaire s'appuyant sur des listes de sujets et des thesauri. Plusieurs intervenants particulièrement concernés par les

problématiques de multilinguisme et de traduction (Europeana, National Diet Library…) ont ainsi manifesté leur intérêt d'exploiter davantage de liens automatiques et d'alignements entre des référentiels de ce type afin, notamment, d'enrichir leurs services de fonctions de recherche et d'affichage multilingues. Certains participants ont appelé à la création d'un « VIAF des sujets » et ont interrogé le devenir du projet MACS (projet assez ancien porté par la Bibliothèque nationale suisse, qui met en relation le langage RAMEAU, les LCSH (Library of Congress Subject Headings) et le fichier mots matières de la Deutsche National Bibliothek. Cette problématique de l'interopérabilité des langages d'indexation sujet, la plus complexe selon les experts, a donc été largement débattue au cours de la journée et constitue un point de recherche et d'attention particulier pour l'avenir.

- **A la recherche de hubs de données de confiance : un écosystème en quête de repères**

Dans ce contexte, une des préoccupations actuelles des artisans du *Linked Data* reste d'identifier des entrepôts de données d'autorité fiables et robustes afin de construire des services maintenables sur la durée. De l'avis général, **la situation est encore extrêmement instable**, car toutes les organisations sont encore en recherche de solutions et font rapidement évoluer leurs services qui ont un statut généralement expérimental.

Pour atteindre le niveau de maturité attendu pour des services performants et capables de gérer des mises à jour fréquentes, l'identification de puits de données de confiance devient un enjeu majeur. C'est pour cette raison par exemple que Electre a indiqué préférer utiliser les données de Wikidata à celles de DBpedia (données mises à jour plus fréquemment et de manière centralisée). C'est pour la même raison que dans la plupart des projets, on retrouve un peu toujours les mêmes « data hubs » dont les utilisateurs considèrent qu'ils remplissent des critères de confiance et de qualité suffisants : VIAF, Geonames, Wikidata, id.loc.gov (mais aussi, de plus en plus souvent cité, data.bnf.fr, dont Electre a indiqué qu'elle réutilisait depuis peu les données d'autorité relatives aux auteurs).

- **Conclusions des ateliers consacrés au web sémantique : les recettes du succès**

Une des conclusions de l'atelier consacré au web sémantique pour débutants (animé par Teodore Fons, OCLC) a été que pour **réussir un projet *Linked Data***, une organisation avait **besoin de quatre choses : de bonnes données, des personnes compétentes, des outils adaptés, et une bonne capacité à penser les produits et les services dans la totalité de leur cycle de vie**. Les bibliothèques disposent généralement de bonnes données. Elles ont souvent besoin de faire appel à des sociétés extérieures (qui ne viennent pas du monde des bibliothèques et des SIGB, mais plutôt des industries et de la R&D du web) pour disposer des bons outils. Le recrutement, la formation et la montée en compétence des agents constitue pour beaucoup un point d'attention prioritaire, comme en témoigne le grand nombre de sollicitations adressées à l'équipe data.bnf.fr de la BnF pour dispenser des formations (en France et à l'étranger) ou pour accueillir des stagiaires. Enfin, la question du cycle de vie des produits et des services (qui n'est pas propre aux technologies du *Linked Data* mais rejoint la problématique plus globale des cycles d'innovation : **on commence par un projet, on doit finir avec un produit**) a été soulignée comme un autre point d'attention à l'heure où beaucoup d'initiatives atteignent le seuil de leur attractivité en tant qu'expérimentations et doivent désormais être organisées comme des services, en intégrant des problématiques de robustesse et d'exploitation sur plusieurs années.

Les participants à l'atelier consacré au web sémantique pour les décideurs (co-animé par Emmanuelle Bermès et Gildas Illien) ont complété ces recommandations de quelques autres. Ils ont d'abord rappelé l'importance de disposer d'un cadre juridique facilitant **l'ouverture des données**, cette ouverture juridique (licences appliquées à la réutilisation des métadonnées) étant un préalable à leur ouverture technique. A aussi été soulignée la nécessité d'être en mesure de construire un discours et de proposer **une vision** (une « destination ») pour accompagner la révolution ou, plus humblement, le changement *Linked Data* au niveau opérationnel comme au niveau politique. Si les équipes s'enferment dans un discours technique sans emporter l'adhésion des personnels (approche « *bottom-up* », comme à la BnF) ni celle de ses décideurs (approche « *top-down* », comme au Centre Pompidou), le portage du projet sur la durée et son insertion dans une stratégie durable sera beaucoup plus difficile. Il faut pouvoir expliquer et partager les objectifs et les bénéfices du *Linked Data* à long terme à partir de réalisations concrètes. Même des représentants de sociétés commerciales (comme Ex Libris) témoignent de leur difficulté à faire partager une vision stratégique au sein de leurs équipes car cela implique de grands changements dans les mentalités comme dans les processus.

Enfin, comme dans tout projet, un **bon management des ressources et des personnes** est crucial : optimiser la combinaison des différents types de ressources (données, agents, outils) est essentiel au bon pilotage. De ce point de vue, un point d'attention particulier évoqué durant la discussion a porté sur **l'organisation des relations de travail**. Toute innovation révèle ou ravive les dysfonctionnements ou des cloisonnements au sein de l'organisation de travail existante. C'est particulièrement le cas des projets web sémantique qui obligent à réfléchir à la convergence d'environnements, voire de « territoires » de production de données historiquement organisés en silos qui coexistent en parallèle au sein de la même organisation. Se pose également la question de gérer les tensions entre les agents chargés de la modélisation des données (travail plus conceptuel) et ceux chargés du traitement des données (travail plus opérationnel et confronté aux contraintes des « tuyaux » existants) : quelles organisations peuvent favoriser de meilleures itérations entre théorie et pratique ? Et quelle serait la fiche de poste idéale d'un agent en capacité de conduire un projet *Linked Data* ? Ces questions de métier et d'identité professionnelle ont largement fait écho aux réflexions en cours sur les nouveaux métiers de « *data specialist* », de « *data architect* » ou de « *data manager* »).

- **L'exemple de la BnF**

La présentation faite en duo par la BnF (Agnès Simon) et son prestataire Logilab (Adrien Di Mascio) de l'organisation du projet data.bnf.fr a été saluée par le public. Dans ses présentations officielles et même sa documentation professionnelle, OCLC cite désormais systématiquement l'exemple de data.bnf.fr comme l'initiative la plus avancée dans le domaine du web sémantique en bibliothèque, considérant que les statistiques de fréquentation de data.bnf.fr (en particulier le taux élevé des visites en provenance des moteurs de recherche) apportent une preuve tangible du succès de cette approche pour améliorer la visibilité des métadonnées de bibliothèque auprès des internautes.

Parmi les points forts du projet, l'équipe data.bnf.fr a insisté sur l'utilisation systématique et la bonne gouvernance interne des identifiants ARK par la BnF, la qualité des fichiers d'autorités à la source de data.bnf.fr, l'environnement logiciel retenu (CubicWeb), mais aussi la méthode de développement (méthode agile Scrum) qui a permis au service de se construire en intégrant pas à pas les attentes des utilisateurs tout en prenant en compte les évolutions rapides d'un environnement très compétitif (évolution des stratégies de moissonnage des moteurs de recherche, par exemple). Parmi les difficultés du projet, l'équipe a rappelé des points souvent abordés par ailleurs au cours de la journée : les épreuves du changement d'échelle, l'ajustement nécessaire dans les méthodes de travail comme dans l'architecture informatique pour passer de la preuve de concept au service en production, la dépendance à des sources de données externes dans un contexte très mouvant.

La BnF a été interpellée au cours de la journée sur la possibilité de mettre à disposition le code et les briques logicielles de data.bnf.fr auprès de tiers, afin de permettre à d'autres bibliothèques ou d'autres organismes de le récupérer pour construire d'autres services du même type. Le code de départ (CubicWeb) est libre, il n'y a donc pas d'obstacle à cette ouverture du point de vue du prestataire Logilab. Gildas Illien a répondu qu'une telle décision n'était pas à exclure sur le principe mais qu'elle serait prématurée. En effet, mettre à disposition un logiciel c'est aussi bien souvent s'engager à aider des tiers à l'installer, à se l'approprier, à le maintenir et garantir enfin une certaine stabilité et une documentation à jour du produit. L'application data.bnf.fr n'est pas encore assez mûre et les équipes qui la développent pas assez disponibles pour faire une telle promesse, qu'elles seraient dans l'incapacité de tenir sur la durée.

Cette demande rejoint des échanges qui ont eu lieu durant la journée sur les leviers disponibles pour les petites et moyennes organisations qui souhaiteraient se lancer aujourd'hui dans le *Linked Data*. Pour ces organisations aux ressources limitées, hormis le rattachement à une solution commerciale ou institutionnelle plus globale, il existe aujourd'hui peu de possibilités de développement. Si la discussion entre les grandes bibliothèques nationales comme la BnF ou la DNB donne l'impression de progrès notables dans ce domaine, il a été noté qu'on « entend toujours parler des mêmes projets » portés par de grosses institutions et qu'il y a « peu de place et de solutions » pour les opérateurs moins nantis.
C'est un point d'attention, autant qu'un signe d'intérêt et de maturité du *Linked Data* dans la communauté professionnelle. Les expérimentations et collaborations engagées par la BnF en prolongement du projet OpenCat constituent de ce point de vue une voie d'exploration de nouvelles collaborations en réseau qui mérite d'être poursuivie.