Web de données, big data, open data, quels rôles pour les documentalistes ?

Gildas Illien, Odile Hologne, Stéphane Pouyllau, Gilles Alfonsi, Jean-Pierre Troeira, Jean Delahousse, Sylvie Dalbin, Ursula von Rekowski, Christophe Aubry, Charles Huot

Web de données, big data, open data, quels rôles pour les documentalistes ?

Stéphane Pouyllau

Avec l'apparition du web, il y a un peu plus de vingt ans, les pratiques des chercheurs en sciences humaines et sociales concernant la réutilisation des documents et données, devenus massivement numériques, ont changé. Le mouvement de l'open data touche aussi le monde de la recherche mais de façon différente suivant que l'on parle de publications, de données ou encore de notes de terrain. Dans cet article, nous utiliserons le terme « document » afin de désigner les publications (articles, mémoires, thèses), impliquant donc un acte éditorial et nous proposons de réserver le terme « données » aux images, fichiers informatiques issus de capteurs numériques (caméras, scanner 3D, enquêtes). Le terme « information » devra être entendu au sens atomique du terme, c'est à dire qu'il désignera des faits contenus dans des données ou décrits dans des documents. Ce propos liminaire est important car la définition de ces termes est différente suivant les communautés scientifiques.

Les documents et données utilisés par les scientifiques pour « faire de la recherche » sont devenus numériques et les échanges de ces derniers se sont accélérés depuis l'invention du web par Tim Berners-Lee, il y a un peu plus de 20 ans. Echanger, diffuser et partager via le web est aujourd'hui au cœur des pratiques des scientifiques et des étudiants. Dans un premier temps, le web a été le support et le vecteur pour la diffusion des documents et des données qui restaient - pour un temps - au fond de bases de données documentaires, dans des systèmes plus ou moins fermés. Il s'agit de la « web-ification » des bases de données dont l'accès peut se faire par formulaires pour les humains et via des systèmes d'interopérabilité/API (dont OAI-PMH) pour les machines. Plus récemment, l'utilisation du web lui-même comme une grande base de données mondiale a été rendue possible par la généralisation et l'utilisation des protocoles, langages, normalisations du web lui-même ; on parle alors de données « dans » le web (web de données). Associé au protocole HTTP, l'utilisation des URI[1] et du modèle RDF transforme le web lui-même en un gisement d'informations qu'il est possible de requêter directement à l'aide du langage SPARQL[2]. Les informations y sont reliées entre elles par des relations de type URL (qui sont aussi des URI) composant ainsi des données inter-dépendantes les unes des autres. Cette « interopérabilité » au niveau des informations est le linked data.

Là ou il fallait maîtriser plusieurs systèmes informatiques et documentaires fondées sur des API[3] différentes entre elles, la promesse faite par le web de données et le *linked data* est de pouvoir décloisonner les bases de données et d'avoir une méthodologie mondiale pour diffuser, interroger, réutiliser les données et les informations dans l'idée de construire des documents plus riches. Dans le monde scientifique cela permet, par exemple, de donner accès de façon normalisé - donc ré-exploitable par d'autres - aux données ayant été utilisées pour infirmer ou affirmer une hypothèse. Dans le monde scientifique, l'open data est fortement lié à l'administration de la preuve scientifique et nous assistons aujourd'hui, avec les techniques du web de données, à un profond changement dans les méthodologies de recherche. De plus en plus de chercheurs seront intéressés à refaire les démonstrations et donc à réutiliser les données. A ce jour le « partage » des données scientifiques s'est majoritairement fait au travers d'interrogation de bases de données « sur » le web (et non pas « dans » le web) d'ou le recours le plus souvent à de multiples API, plus ou moins ouvertes, mais toujours fortement lié à une vision purement informatique des choses et de ce fait peu accessibles pour les métiers de l'information et de la documentation. De plus, les API évoluent dans le temps, leurs spécifications changent et impliquent donc un suivi important

[1] Pour Uniform Resource Identifier.
[2] SPARQL (SPARQL Protocol and RDF Query Language) est le langage d'interrogation des données RDF dans le web de données
[3] Les API (Application Programming Interface) sont des programmes informatiques permettant de faire de l'interopérabilité entre logiciels.

au niveau de son système d'information documentaire pour continuer à « dialoguer » avec l'API sur laquelle est fondée le service que l'on propose (c'est particulièrement vrai des API servant à géo-localiser des données). La multiplication des API entraine une quasi impossibilité de maintenir un outil qui exploiterait plusieurs dizaines de bases de données différentes. Dans les laboratoires de recherche en sciences humaines et sociales, la rareté des informaticiens reporte parfois sur les documentalistes et bibliothécaires le soin de proposer de tels outils. Mais il n'est pas forcément aisé de comprendre et suivre des spécifications techniques du domaine informatique. Les API de logiciels documentaires sont de bons vecteurs pour le partage des données à partir du moment où elles sont conçues avec une vision documentaire et pas uniquement informatique.

L'environnement technologique de l'information scientifique et technique (IST) a donc « muté » depuis l'invention du web. Là ou le seul signalement était suffisant, le web a permis l'échanges des documents et les chercheurs ont inventés de nouvelles manières de partager les résultats de la recherche : les archives ouvertes[4] en sont un exemple. Ayant accès aux articles, puis à des données au travers du web et enfin à des moteurs de recherche performant et mondiaux, les pratiques de recherche d'information des chercheurs ont changé. Le volume des données augmentant, la puissance des ordinateurs avec, une certaine « autonomie » s'est installée chez les scientifiques délaissant parfois certains savoir-faire des professionnels de l'IST : en particulier sur les aspects de structuration de l'information et dans l'utilisation de vocabulaires documentaires adaptés aux différents besoins, en particulier dans le domaine de l'édition électronique. Ainsi, même si la définition de « big data » ne peut pas être comparable, en volume, entre des données de physique des particules et des données d'histoire (quoique), les volumes des données accessibles via le web sont de plus en plus important. Hélas, ces volumes ne sont pas toujours immédiatement compréhensibles du fait de métadonnées peu ou pas assez qualifiées pour des réutilisations ; ou encore par l'absence de référentiels réellement utilisables dans le monde du web par manque de mise en relation les uns avec les autres. C'est plus largement tout le problème du mouvement de l'open data : un déluge de données « tableurs » mais un silence important quant au contexte de production, de validité des données, d'utilisations antérieures, d'intégrité, etc. Sans parler encore de la problématique des formats ouverts. Or le web de données repose sur la fiabilité de l'information et donc sur une relation de confiance entre producteurs de données et utilisateurs ; ce que j'appelle l'inter-dépendance entre les données. Il y a là, un formidable enjeux pour les métiers de l'IST qui doivent ré-investir leurs savoir-faire dans cette évolution du web.

La diffusion et la publication de documents et données structurées et la gestion de référentiels structurés (thésaurus, listes d'autorités, ontologies) dans le web de données sont deux enjeux majeurs pour les documentalistes et bibliothécaires dans les années qui viennent. Maintenir un thésaurus ou un référentiel en RDF/SKOS qui sera utilisé et ré-utilisé en ligne doit faire partie des savoir-faire que ces métiers doivent proposer dans les laboratoires ; afin justement de garantir l'inter-dépendance entre les données d'archives et les publications par exemple. Il s'agit de valoriser les données structurées en proposant leur valeur scientifique dans les usages documentaires multiples. Il s'agit là de construire des responsabilités et garanties d'accès sur le long terme. Qui mieux que les documentalistes et bibliothécaires pourraient être les responsables de la pérennité de ces relations tissées entre les informations contenues dans les documents et les données ?
C'est la proposition faite par le projet Isidore[5], première plateforme d'enrichissement et d'accès aux données et documents ouverts de la recherche en sciences humaines et sociales. Conçue en 2010 par le très grand équipement Adonis (aujourd'hui Huma-Num[6]) avec les savoir-faire d'un

[4] BOSC, H., « Archives Ouvertes : quinze ans d'histoire », In Les Archives Ouvertes : enjeux et pratiques. Guide à l'usage des professionnels de l'information, 2005. http://hdl.handle.net/10670/1.p8378e
[5] Voir http://www.rechercheisidore.fr
[6] Voir http://www.huma-num.fr

consortium d'industriels (Antidot, Sword et Mondéca), Isidore permet aux professionnels de l'information et de la documentation de valoriser thésaurus et référentiels en réalisant des enrichissements sémantiques sur les métadonnées qui sont moissonnées par Isidore. Il s'agit sans doute d'une première expérience dans ce sens, qui pourrait se décliner par exemple dans les collectivités territoriales qui ont des masses de données importantes et de nombreux référentiels à ouvrir et à partager sur le web. L'une des spécificités de ce projet et qu'il utilise strictement les méthodes et techniques du web sémantique et du linked data pour proposer – dans une base de données RDF ou triple store – le résultat de ces métadonnées catégorisées et « augmentées » d'annotations. Isidore privilégie ainsi les informations structurées : tant pour les métadonnées et données qu'il moissonne, que pour les référentiels qui permettent les enrichissements. Ce qui montre que les professionnels de l'info/doc doivent accompagner en profondeur les enjeux de l'accès aux données (big et open data). Il s'agit là de l'indispensable garantie de succès afin que ces masses de données soient réellement réutilisables et ce par apport d'une contextualisation des données, par l'ajout de métadonnées, par la production d'enrichissement qui permettront de relier les informations entre elles et non simplement les mettre à disposition. Car aujourd'hui, l'open data ressemble le plus souvent à dresser des listes de catalogues de données comme on le faisait aux premières heures du web de vitrine 1.0.