Opportunités et défis

Linked (Open) Data*

Quand Tim Berners-Lee a commencé à développer le web en 1989, il avait déjà envisagé un lieu où les documents et les données pourraient être reliés entre eux. Ces idées ont été rendues plus concrètes dans l'article du Scientific American où le concept de « Web sémantique » a été introduit. L'activité web sémantique du W3C, le consortium du web, a travaillé plusieurs années pour fournir la technologie de base permettant cette liaison.

Le principal résultat de ce travail est le modèle RDF (Resource Description Framework) qui offre un moyen de faire des déclarations sur les ressources en utilisant un modèle générique de « triplet »: sujet-prédicat-objet ; par ex. « cet article » (sujet) – « sonTitre » (prédicat) – « Linked Data » (objet). Puis, en 2006, il a commencé à utiliser le terme « Linked Data » pour la prochaine génération du web où les données pourraient être liées de manière que les machines puissent donner un sens aux données.

Les quatre règles qui ont été formulées sont :

1 utiliser des URI comme noms des choses ;

2 utiliser des adresses HTTP URI afin que les gens puissent voir ces noms ;

3 quand quelqu'un regarde un URI, fournir des informations utiles en utilisant les normes (RDF, SPARQL) ;

4 inclure des liens vers d'autres URI afin qu'ils puissent découvrir plus de choses.

Un 5 étoiles

Tim Berners-Lee a défini un système à cinq étoiles pour les données liées.

* **Faire que vos ressources** soient disponibles sur le web (quel que soit leur format)

** **Les mettre à disposition** comme données structurées (c'est-à-dire excel plutôt qu'une image reproduisant une table)

*** **Dans un format non propriétaire** (c'est-à-dire CSV plutôt qu'excel)

**** **Utiliser des URL** pour nommer les choses afin que les gens puissent accéder à vos ressources

***** **Relier vos données** aux données des autres afin de contextualiser celles-ci

Le Linked Data se réfère principalement à l'Open Data mais peut également être utilisé pour des données d'entreprise.

Les deux utilisent les technologies du web sémantique (RDF) et visent à lier les informations. La différence est que le LOD (Linked Open Data) utilise des licences libres afin que les données soient disponibles pour utilisation par des tiers, tandis que le LED (Linked Enterprise Data) est utilisé dans des environnements fermés, contrôlés où les données ne sont généralement pas disponibles pour une utilisation faite par des tiers.

Nous entendons parler indifféremment du Linked Data et de l'Open Data. Mais ce n'est pas la même chose : le Linked Data est axé sur la technologie, en utilisant RDF et d'autres standards du web, tandis que l'Open Data a une orientation stratégique fondée sur l'idée que le partage est important et bénéfique à un large public ; la technologie est moins importante dans cette perspective.

Les développements récents

Divers groupes travaillent dans cet espace. Le W3C a plusieurs groupes de travail sur des thèmes spécifiques : le ***SWEO Community Project, the Library Link Data Incubator Group, the Government Linking Data Working Group, the Semantic Web Health Care and Life Sciences Interest Group.*** *Ces groupes se composent principalement de techniciens dont l'objectif est de définir les cas où il faut recourir à la technologie. Il y a aussi des groupes qui s'intéressent plus à des questions stratégiques, réunissant des chercheurs et des praticiens qui examinent les avantages escomptés et les questions stratégiques; l'****Open Knowledge Foundation*** *est l'un de ces groupes. Il gère aussi un registre des collections de données libres dont CKAN fait partie.*

Il existe plusieurs initiatives qui définissent le vocabulaire des prédicats, par exemple RDA, ***The bibliographic ontology*** *(bibo), le Dublin Core et bien d'autres qui peuvent être utilisés pour décrire différents types de ressources. D'autres définissent le vocabulaire d'objets tel que le* ***Virtual International Authority File*** *(VIAF), les autorités de la Bibliothèque du Congrés, AgroVoc, DBPedia qui peut être utilisé pour lier. Les données bibliographiques sont publiées sous forme de données liées, par exemple par LIBRIS en Suède, par la British Library ou par CrossRef. Dans le monde des médias, la BBC construit ses services sur le principe de données liées ; le New Tork Times est aussi un exemple pour la presse écrite. Et enfin, et non des moindres, de nombreuses organisations gouvernementales publient des données liées, par exemple les États-Unis, la France, la Finlande, le Royaume-Uni, la Catalogne, la Norvège, les Pays-Bas, l'Australie et bien d'autres.*

La façade du Corum
Crédit photo : Lilian Martorell
Les Journées ABES 2011
ont eu lieu à Montpellier, au Corum

Comparaison des objectifs

Si l'on considère ses objectifs stratégiques, le Linked Data vise à assurer l'interopérabilité mondiale avec un minimum de coordination afin de regrouper le savoir humain en soutien à la démocratie, à la transparence et à la nécessité de rendre des comptes. Il essaie d'améliorer et d'enrichir les données et donne l'espoir de créer un environnement où des applications créées et gérées par les citoyens apparaîtront. Par ailleurs, les bibliothèques ont toujours organisé l'information à l'usage d'utilisateurs spécifiques pour des objectifs spécifiques, assuré et maintenu la qualité de services durables. Une attention particulière a été accordée à la préservation de l'information à long terme et à la fourniture de services de confiance.

Sur le plan fonctionnel, le Linked Data vise à permettre la recherche dans les collections publiées et la navigation instinctive : le saut depuis un élément d'information vers un autre en suivant les liens entre eux. La responsabilité des déclarations sur les ressources est considérée comme étant répartie entre de nombreux fournisseurs de données, en laissant l'usager décider à qui et à quoi faire confiance. Les développements de produits et services sont laissés à un marché ouvert pouvant inclure des fournisseurs de services commerciaux mais aussi des programmeurs individuels qui utilisent les données pour construire toutes sortes d'applications. Les bibliothèques visent à décrire l'information par des professionnels, rassembler des ensembles d'information ainsi gérés. La bibliothèque sélectionne l'information pouvant être pertinente pour l'utilisateur et mélange ressources imprimées et ressources numériques. Techniquement parlant, le Linked Data s'occupe de la publication et de l'utilisation d'instructions lisibles par les machines (« les données qui parlent à elles mêmes »), exprimées en RDF. Cela devrait permettre des corrélations pour de grands ensembles de données publiées, la création de connaissances à partir des informations contenues dans les liens, mais il faut dire que les solutions pour le moissonnage, la mise à jour en cache et en temps réel ne sont pas entièrement résolues. Dans le domaine des bibliothèques on met moins l'accent sur la technologie en soi. Le plus important est que les solutions soient basées sur une technologie éprouvée qui permette des services de haute qualité, garantissant des performances, la disponibilité et la cohérence entre les données. En comparant les approches de développement, une spécificité du Linked Data est que dans cet univers les choses bougent vite, les développements se font plus ou moins en tâtonnant, par essais et échecs successifs, tandis que dans le domaine des bibliothèques il y a une base installée avec les données existantes qui doivent être entretenues et gérées. Il en résulte que les projets de développement sont généralement encadrés plus fortement. Une différence importante est que les initiatives Linked Data, comme le terme l'indique, sont très centrées sur les données, parfois même appelées « données brutes », où la disponibilité et la quantité sont plus importantes que la qualité et où on s'attend à ce que, tant que les données sont disponibles, les développeurs viennent les utiliser ; alors que dans le domaine des bibliothèques l'objectif principal est de fournir des services aux lecteurs où la qualité est essentielle et où les données et la technologie sont utilisées à l'appui de ces services.

Enfin, sur les aspects économiques, il y a aussi une différence. Le Linked Data est un environnement où chaque jour apporte son lot de nouveaux défis et d'idées ; les développeurs peuvent oublier les outils d'hier assez rapidement ; il y a apparemment peu d'intérêt pour la fracture numérique : les gens qui n'ont pas les compétences ou les moyens de manipuler les ressources numériques ou même pour programmer ne sont pas pris en compte. Le monde des bibliothèques est, bien sûr, davantage concerné par le fait de proposer des services continus à la communauté dans son ensemble et doit veiller à ce que les budgets de ces services reposent sur une base stable.

... / ...

Makx Dekkers, consultant indépendant
en gestion d'information et projets internationaux
http://www.makxdekkers.com/index.html
mail@makxdekkers.com

Risques, défis et récompenses

Il y a plusieurs aspects intéressants dans la manière dont est menée l'évolution du Linked Data. Parmi ceux-ci, la tentative de parvenir à une plateforme technique commune pour les données lisibles par machine qui va au-delà des capacités de base du XML. Il y a beaucoup d'enthousiasme dans le domaine technique qui se frotte certainement au niveau politique en raison de la promesse d'interopérabilité mondiale. De nombreuses collectivités sont impliquées, tels que les chercheurs, les communautés d'utilisateurs, les pirates et les fournisseurs de données professionnels comme les organismes gouvernementaux et les universités. Mais il y a aussi des risques. Les initiatives sont souvent motivées par la technique et non par des exigences. En ce sens, on pourrait dire qu'il s'agit d'une « solution en quête d'un problème ». La technologie n'est pas (encore) stable avec une deuxième version de RDF en cours de développement tandis que les aspects opérationnels, par exemple performance, fiabilité, qualité, sécurité et confiance, n'ont pas encore été résolus. Il est peut être également naïf de penser qu'un accord mondial, au-delà des domaines et des frontières, sur une seule technologie pourra être atteint. En fin de compte, il y a un risque que si le Linked Data ne tient pas sa promesse de fournir une interopérabilité mondiale, le résultat puisse être la déception et la perte d'intérêt de la part de ses premiers supporters.
Si l'on regarde le domaine des bibliothèques, les points forts sont que les bibliothèques ont une longue expérience opérationnelle dans la gestion de l'information et de le faire en utilisant des modèles professionnels durables (quoique avec des budgets en diminution constante). Les bibliothèques ont toujours joué un rôle d'intermédiaire entre les utilisateurs et les besoins d'information et ont une vision à long terme : le passé (les données existantes) aussi bien que l'avenir (la conservation).

Pour les points négatifs, on pourrait mentionner l'évolution rapide de la technologie, qui peut être difficile à suivre et qui nécessite de développer de nouvelles compétences devant être propagées dans l'organisation. Le monde extérieur peut aussi considérer que les bibliothèques sont une chose du passé (« le musée du livre ») et que les compétences de traitement de l'information ne sont pas si importantes. En outre, comme les bibliothèques elles-mêmes s'en rendent compte, la quantité d'informations est telle qu'il est difficile d'appliquer les traditionnelles méthodes du manuel. Cependant, il n'est pas impossible pour ces deux mondes de se rencontrer. Un bon exemple est Europeana, qui compte quatre domaines (bibliothèques, archives, musées, archives audiovisuelles) et a d'abord travaillé avec une approche « traditionnelle » en définissant des mappages de métadonnées dans ce qu'on appelle les « Éléments sémantiques » d'Europeana. Au vu de la réussite d'Europeana aujourd'hui, on ne peut pas nier que cette approche a fonctionné. Toutefois, le niveau de détail que les différents domaines offrent dans leurs données a été perdu à la suite du « nivellement par le bas » des données. Le développement à venir d'Europeana est maintenant basé sur une approche Linked Data qui préserve les spécificités des domaines et permet de généraliser le soutien aux services communs. L'objectif de cette approche est de parvenir à une meilleure interopérabilité entre les domaines, soutenue par une précoordination.

Opportunités

Les bibliothèques ont certainement un rôle important à jouer dans l'espace des données liées. Tout d'abord, les bibliothèques devraient adopter la technologie Linked Data comme une étape possible vers la connexion des services ; être impliquées permet aux bibliothèques de rester informées des possibilités. Deuxièmement, les bibliothèques ont beaucoup à offrir : leurs compétences en gestion de l'information bénéficieraient grandement aux approches plus informelles utilisées dans le domaine technologique. Enfin, les bibliothèques gèrent de grandes quantités de données de valeur qui permettent de créer ce qu'on pourrait appeler des « pôles de qualité » à la fois comme données primaires ainsi que comme source pour les liens entre les ressources.

Lorsque les bibliothèques s'engagent dans le mouvement Linked Data, elles peuvent apporter stabilité et durabilité aux espaces de données liées, les aider grâce à leurs compétences dans la gestion des collections et fournir au web des données liées de haute qualité. Lorsque cela sera combiné avec la prochaine génération de systèmes et d'outils que les techniciens développent, le nouveau web sera plus utile à tous.

Makx Dekkers

*Extraits de la conférence inaugurale des Journées ABES 2011 traduits de l'anglais par Fabien Bénistant