Vers les Données Liées : Conséquences Théoriques et Pratiques Pour les Sciences Humaines

Lyne Da Sylva
lyne.da.sylva@umontreal.ca
Université de Montréal, Canada

Introduction

Le Web ne sera peut-être jamais plus comme on le connaît maintenant. Ce vaste répertoire de connaissances et informations publiées et commentées par les internautes risque, si la tendance des travaux du W3C se maintient, de s'enrichir d'une couche de représentation supplémentaire. Le Web sémantique et les données liées ne visent pas à remplacer le Web existant, mais à s'y greffer. Ils représentent un nouveau paradigme de représentation de l'information sur le Web, non plus comme des documents cohérents (des pages Web lisibles par l'humain), mais plutôt comme des jeux de données (Linked Open Data) représentant des relations binaires entre un objet et une propriété de celui-ci : des « triplets », encodés selon des standards précis et permettant des traitements automatiques à grande échelle. À la place du Web existant, ou en complément de celui-ci, le Web de données ouvre de multiples possibilités de mise en rapport d'informations diverses. Il fait miroiter des possibilités de recherches d'information plus élaborées (parce que basées sur des ensembles d'inférences éventuellement complexes) et de traitements automatiques sophistiqués sur les données.

Tout ceci est sans doute vrai; plusieurs membres de la communauté scientifique et de diverses communautés professionnelles ont déjà contribué à construire le désormais gigantesque réseau Linked Open Data (Europeana, OCLC, Library of Congress, ACM, BBC Music,…).

L'objectif de cette communication n'est pas de remettre en question le processus, mais bien de poser un regard éclairé sur certaines des conséquences de ce passage des documents aux données liées pour les sciences humaines.

Brève description du Web sémantique et des données liées

Le Web sémantique repose de manière importante sur l'identification et la description d'entités : les entités dignes d'importance, dans les représentations du Web sémantique, incluent toute entité sur laquelle on peut vouloir exprimer des propriétés ou relations. Ces entités peuvent être des ressources présentes sur le Web (des sites web, des bases de données, etc.) ou non : des personnes en chair et en os, des livres imprimés, des sites géographiques, archéologiques, touristiques, etc. Ou toute autre entité abstraite comme une date, une couleur, un rite funéraire, etc.

Nous présenterons les éléments de base du Web de données : les identifiants (URI) qui représentent les entités, les relations de diverses natures qui permettent de relier les entités, les triplets RDF qui servent à encoder les relations entre entités, les vocabulaires et ontologies qui permettent des descriptions uniformisées, les triplestores qui emmagasinent les jeux de triplets RDF.

Nous soulignerons également les avantages de l'utilisation de données liées en sciences humaines, en illustrant à l'aide de quelques projets de recherche récents (en histoire : Michon, 2016; en littérature : RDA, 2015; en musique : Cannam et al, 2011)

Conséquences pratiques

Suite à cette brève présentation, nous examinerons d'abord les conséquences pratiques du passage aux données liées pour les sciences humaines. Celui-ci influe premièrement sur le focus de la recherche, qui se fait alors graduellement vers l'information et non le document (processus antinomique à certaines disciplines en sciences humaines). Deuxièmement, le travail nécessaire à l'extraction des données est considérable, contrairement à ce qui est de mise en sciences pures (où une partie de l'encodage des données liées est fait à partir de données discrètes obtenues par des appareils de mesure : sondes océanographiques, lectures géologiques, observations astronomiques). Troisièmement, nous verrons que cette transformation rappelle la notion de « redocumentarisation » déjà observée pour le document numérique (Pédauque, 2007). Le terme est utilisé pour décrire les bouleversements induits par l'apparition du document numérique; l'encodage des données dans les technologies du Web sémantique exige d'extraire l'information encodée dans des documents existants et de l'exprimer dans un nouveau format. La création de chaque triplet nécessite un travail d'analyse fine des documents et la

décomposition en ses unités élémentaires. On parlera ici de

l'atomisation de l'information contenue dans les documents, qui a également des conséquences théoriques (voir ci-dessous). Mais les conséquences pratiques sont déjà considérables.

Conséquences théoriques

Les conséquences théoriques du passage au Web de données touchent d'abord la nature des objets de l'étude. En premier lieu, nous aborderons la réification des éléments d'information : tout énoncé (triplet RDF) destiné au Web de données requiert (i) que l'entité à décrire soit définie ontologiquement dans l'univers de référence (le vocabulaire ou l'ontologie); (ii) que l'entité qui lui est reliée reçoive le même traitement (même lorsqu'il s'agit de concepts abstraits comme la couleur d'un objet); et (iii) que la relation entre les deux entités soit elle aussi réifiée. La réification des deux entités ne choquera pas le chercheur, habitué à scruter l'essence des concepts qu'il étudie. Mais la réification de la relation – l'élévation au rang de concept de tout type de relation comme « est l'auteur de » ou « a correspondu avec » – changera sans doute de manière considérable la vision que le chercheur aura de cette relation.

Une deuxième conséquence théorique du passage au Web de données est le fait que la distinction entre données et métadonnées devient floue, voire inutile. Par le biais de l'extraction de données à partir de documents primaires, tout ce qui est encodé devient métadonnée. Mais si tout est métadonnée, de quelles données sont-elles les métadonnées? Les données disparaissent-elles? ou est-ce plutôt le concept de métadonnées qui disparaît?

Nous discuterons également de l'uniformisation exigée par les vocabulaires (ou référentiels) partagés. Si les vocabulaires sont davantage stables dans les sciences dites exactes, la terminologie fait moins consensus en sciences humaines et sociales (ceci a été largement documenté dans la construction de thésaurus documentaires). On peut s'attendre à un plus grand nombre de référentiels concurrents. Il est possible que le travail sur les ontologies ait un effet important sur la définition des concepts de base des disciplines et les tentatives de rapprochement entre disciplines. Les travaux en terminologie et en conception de thésaurus pourraient être mis à contribution dans l'entreprise. D'un autre côté, le fait que les données soient liées permet des interconnexions (et un potentiel de normalisation) qui n 'est pas requis lorsque les recherches se font davantage en parallèle.

Enfin, nous présenterons la notion de l'atomisation des objets de recherche. L'apport intellectuel des chercheurs en sciences humains est le résultat d'un travail d'analyse, de mise en rapport, d'abstraction à partir de données disséminées (documents historiques, phénomènes linguistiques, observations sur le terrain ...). Or ces documents sont des créations, des synthèses, exprimant des idées réunies par l'auteur en propositions, en phrases, en paragraphes cohérents. À partir du moment où ces informations sont atomisées pour le Web de données, ce travail d'analyse est (potentiellement) perdu, disséminé dans les simples triplets. On peut avancer qu'au contraire, les relations identifiées par les chercheurs, et exprimées par les documents résultants, peuvent faire partie des relations exprimées par les triplets RDF. Mais on doit bien prendre conscience du fait que chaque énoncé reste, dans le formalisme, isolé des autres. Reste aux chercheurs futurs la tâche de développer des moyens de faire des abstractions additionnelles (à un niveau plus macro) à partir des simples triplets RDF.

Conclusion

Le passage au Web de données peut être extrêmement utile à la recherche en sciences humaines. Nous présenterons un certain nombre de conséquences pratiques et théoriques de cette mutation. Le passage aux données implique l'atomisation du sujet de recherche et la redocumentarisation de sa matière première, qui exige un travail important d'extraction d'informations et de réencodage. Il a comme conséquence la réification de chaque élément d'information inclus dans la nouvelle description. Les représentations résultantes brouillent la frontière entre métadonnées et données. Enfin, le passage aux données liées entraînera un partage des référentiels, un potentiel de plus grande interopérabilité entre les descriptions et, peut-être, une injection de travaux sur les concepts fondamentaux des disciplines, de pair avec des travaux en terminologie et en vocabulaires contrôlés. Il est important pour les chercheurs en sciences humaines de mesurer l'apport potentiel du Web de données et les transformations qu'il pourra apporter au développement de leur science ; la présente communication veut amorcer la réflexion sur le sujet.

Bibliography

Cannam, C., Sandler, M., Jewell, M.O., Rhodes, C., d'Inverno, M. (2011). Linked Data and You: Bringing Music Research Software into the Semantic Web. *Journal*

of New Music Research. Volume 39, no 4: Music Informatics and the OMRAS2 Project.

Linked Open Data. http://linkeddata.org/

Michon, P. (2016). Données liées historiques : De la nécessité d'un partenariat entre l'archivistique et l'histoire. http://congres.archivistes.qc.ca/wp-content/uploads/2016/08/DonneeHistoriques.pdf

Pédauque, R.T. (collectif), (2007), *La Redocumentarisation du Monde*, Paris : Éditions Cépadues.

RDF, (2015). *Jane-athon*. http://rballs.info/topics/p/jane/janeathon.html