Humanités Numériques et Web Sémantique : du langage naturel à une représentation computationnelle structurée et sémantique des données

Pascaline Tchienehom
pkenfack@u-paris10.fr
Université de Paris 10, France

Résumé

ModRef est un projet du laboratoire Labex "Les passés dans le présent" qui accompagne divers projets sur des problématiques relatives aux humanités numériques (Oldman et al., 2014). Le projet ModRef s'intéresse spécifiquement au web sémantique (Berners-Lee et al., 2001) et aux données ouvertes et liées. Le but de ce projet est de réaliser une migration de données hétérogènes vers des triplestores encore appelés entrepôts ou collections de fichiers RDF afin d'améliorer le partage, l'échange et la découverte de nouvelles connaissances. Pour ce faire, la norme CIDOC-CRM (Boeuf et al., 2015) a été choisie car elle est actuellement la norme de référence pour la description sémantique de l'information muséographique ou d'héritage culturel (Hooland et Verborgh, 2014). Cette norme permet de décrire les caractéristiques globales des objets (identifiant, type, titre, matériau, dimension, note) mais également leur historique au travers d'évènements ou d'activités (origine, transfert de garde -localisations anciennes, localisation actuelle-, conservation) ainsi que les relations qui existent entre objets ou parties d'objets (bibliographie, composition, similarité, autre représentation -photo, dessin, tableau-, inscription). Par ailleurs, trois sous projets pilotes de ModRef ont été sélectionnés pour réaliser la migration des données : un conservatoire numérique de l'ensemble des documents rédigés en écriture cunéiforme, un corpus numérique d'objets archéologiques à iconographie mythologique et une bibliothèque numérique sur l'histoire de France du 20ième siècle. Les données de ces différents projets sont initialement décrites dans des bases de données ou dans des fichiers XML-EAD (Encoded Archival Description). Pour réaliser la preuve conceptuelle du projet ModRef, une architecture générale a été définie; une modélisation sémantique CIDOC-CRM et un alignement des données des différents sous projets pilotes ont été proposés; une migration des données vers des triplestores a également été effectuée. Une application web a été développée et déployée à l'adresse "http://triplestore.modyco.fr". Cette application web permet de décrire le projet ModRef mais aussi de consulter et d'interroger les triplestores créés. Les triplestores posent deux principaux défis scientifiques et techniques. Le premier est la migration de données souvent décrites initialement en langage naturel vers une représentation computationnelle structurée, puis sémantique de ces dernières. L'autre défi est l'exploitation des triplestores via des Endpoint Sparql (interface de saisie et d'éxécution de requêtes Sparql) ou via des interfaces sous forme de formulaires généraux d'interrogation.

Migration de données vers des triplestores

Une migration efficace et cohérente de données fait appel à différentes compétences. Pour assurer la pérénisation de cette procédure, une architecture générale et rigoureuse du workflow des différents types de données à manipuler doit être définie. Cette architecture explicite la démarche globale de tout projet qui souhaite faire migrer ses données vers des triplestores. Cette démarche se subdivise en différentes étapes bien identifiées : préparation des données (étude et description structurelle), modélisation sémantique et alignement des données structurées avec le modèle sémantique et enfin création et exposition des triplestores qui vont alors pouvoir être consultés et intérrogés. Notons que initialement les données sont souvent non structurées ou semi-structurées (notes, rapports, livres, html) et qu'il faut dans un premier temps en extraire une représentation structurée (tableurs, base de données, fichiers XML) pour pouvoir ensuite construire leur représentation sémantique plus facilement. Ce continuum d'étapes fait intervenir des compétences diverses et nécessite parfois d'adjoindre des profils intermédiaires entre deux étapes pour assurer le passage d'un format de représentation de données à un autre : (1) données non structurées ou semi structurées vers données structurées, et (2) données structurées vers données sémantiques. Par ailleurs, l'élément clé de l'architecture de la migration de données vers des triplestores est la modélisation et l'alignement des données avec le modèle de graphe

sémantique choisi. Un graphe sémantique est un ensemble de noeuds et d'arcs orientés qui obéissent à un certain nombre de contraintes et règles (raccourci, héritage, inverse, symétrie, transitivité). Ce sont ces règles et contraintes qui définissent la cohérence et la validité d'un modèle. Nous avons utilisé la version 6.2 de mai 2015 du CIDOC-CRM qui définit 94 classes et 168 propriétés ainsi que son implémentation par l'Université d'Erlangen-Nuremberg. Afin de réaliser la migration, il a fallu procéder à un alignement des données avec certains noeuds du graphe sémantique à partir des informations extraites de bases de données ou de collections de fichiers XML-EAD. Les noeuds remplis par des valeurs sont des noeuds terminaux et les noeuds intermédiaires sont remplis avec des URIs qui définissent ainsi des chemins vers les noeuds terminaux. Notons qu'une rigueur particulière doit être apportée à la construction des URIs, à la fois pour leur lisibilité mais également pour la cohérence des chemins dans le graphe afin d'éviter des conflits de chemins et garantir ainsi l'unicité d'un chemin donné par rapport à un autre. Nous avons identifié les classes utiles (menant vers au moins une valeur non vide) pour modéliser les données des projets pilotes. Ainsi, la modélisation et l'alignement effectués représentent des extraits de graphes relatifs aux quatre thèmes suivants : (1) caractéristiques générales (identifiant, type, titre, matériau, dimension, note), bibliographie, composition et similarité d'objets; (2) évènements de début d'existence (origine) et de fin d'existence; (3) activités diverses (transfert de garde, conservation, mesure); (4) inscriptions et autres représentations (photo, dessin, tableau). De façon générale, ces extraits sont assez stables pour tout projet car, dans le CIDOC-CRM, il est possible d'identifier les chemins possibles menant à une information donnée sur un objet. L'alignement n'est pas une tâche programmatique mais fait appel à des détails de structure logique propre au modèle de description de données choisi par chaque sous projet. C'est une tâche à mi-chemin entre la modélisation et l'implémentation qu'elle permet d'entrevoir un peu plus clairement. L'alignement définit ce à quoi correspond chaque noeud de notre graphe et il ne reste plus qu'à générer les fichiers CIDOC-CRM correspondants tout en respectant la syntaxe de la norme RDF. Les triplestores créés vont ensuite être exposés pour consultation (sous trois formes : *rdf, triplets et résumé attribut-valeur*) et interrogation (*formulaires généraux et Endpoint Sparql*) via notre application web. L'exploitation des triplestores via l'interrogation et l'exploration de ces derniers et les bénéfices que l'on peut en tirer est l'autre aspect majeur autour de la question de ces nouveaux entrepôts de documents RDF que sont les triplestores.

Exploitation des triplestores

L'intérêt des triplestores est qu'on a un modèle connu public et publié de représentation de l'information ce qui permet d'interroger les triplestores indifféremment avec des procédures identiques. Nous avons défini deux procédures d'exploitation de nos triplestores : des interfaces sous forme de *formulaires généraux* et des *Endpoint Sparql*. Les formulaires généraux sont un moyen simple et assez intuitif, car très proche du langage naturel, pour formuler des requêtes vers nos triplestores. Il suffit de remplir les rubriques du formulaire qui nous intéressent et de lancer la recherche. Une requête Sparql est automatiquement construite à partir des valeurs des champs renseignés du formulaire et c'est cette requête qui est utilisée pour interroger le triplestore. Au terme de l'exécution de la requête, une liste d'objets séléctionnés est renvoyée en résultat à l'usager. Par ailleurs, on peut aussi interroger nos triplestores via des Endpoint Sparql. Ce deuxième mode d'interrogation nécessite la connaissance du langage Sparql qui est aujourd'hui le langage de référence pour l'interrogation de documents RDF. Sparql est un langage assez simple mais pas toujours à la portée de tous. Ainsi, les formulaires généraux peuvent être vus comme un premier point d'entrée pour l'interrogation des triplestores tandis que les Endpoint Sparql assurent une exploitation plus large de ces triplestores via une formulation libre de requêtes de type "Select". Notons que la notion d'exploitation de triplestores fait appel aux notions d'interrogation et d'exploration de graphe. Ainsi, l'interrogation de triplestores consiste à formuler une requête Sparql pré-formatée (formulaires généraux) ou libre (Endpoint Sparql) tandis que l'exploration de triplestores est une forme d'interrogation uniquement possible via des Endpoint Sparql qui permet aussi de découvrir différents chemins dans un graphe sémantique vers des données précises. En effet, plusieurs chemins peuvent permettre d'obtenir une même information dans un graphe (usage de diverses notions : raccourci, héritage, inverse, raffinement), sachant que ces chemins ne sont pas toujours tous renseignés. On peut donc écrire des requêtes Sparql pour découvrir si différents chemins vers une donnée précise existent ou pour connaitre les noeuds terminaux. L'exploration est donc importante pour s'approprier un triplestore spécifique. L'exploration

permet aussi la comparaison de différents triplestores qui décrivent des données similaires (objets d'une même période historique, objets de même type, objets identiques) dans un contexte de LOD (Linked Open Data), par exemple. Ainsi, la comparaison de chemins assure une meilleure découverte des connaissances et augmente la correction ou l'enrichissement mutuel des connaissances des différents acteurs du LOD. Notre application web fournit un LOD pour ModRef ainsi qu'une liste de modèles de requêtes Sparql pour interroger, explorer et valider nos triplestores séparément ou ensemble. A plus long terme, l'objectif est d'intégrer d'autres LOD sur internet (Beek et al., 2016) (Daga et al., 2016) pour un partage, un échange et une découverte de nouvelles connaissances à plus grande échelle. Ainsi, le LOD doit améliorer la découverte de nouvelles connaissances, du fait de l'usage de formalismes, de langages de métadonnées, de thésaurus publiés, standardisés voire normalisés.

Remerciements

L'auteur remercie le laboratoire Labex "Les passés dans le présent" de l'Université de Paris 10 et le projet ANR ModRef de référence ANR-11-LABX-0026-01.

Bibliographie

Beek, W., Rietveld, L., Schlobach, S., et van Harmelen, F. (2016). Lod laundromat: Why the semantic web needs centralization (even if we don't like it). *IEEE Internet Computing* 20(2), 78–81.

Berners-Lee, T., Hendler,, J., et Lassil, O. (2001). The semantic web. *Scientific American*.

Boeuf, P. L., Doerr, M., Ore, C.E., et Stead, S. (2015). Definition of the cidoc conceptual reference model, version 6.2. *Produced by the ICOM/CIDOC Documentation Standards Group, Continued by the CIDOC CRM Special Interest Group*.

Daga, E., d'Aquin, M., Adamou, A., et Brown, S., (2016). The open university linked data data.open.ac.uk. semantic web. *Semantic Web* 7(2), 183–191.

Hooland, S.V., et Verborgh, R. (2014). *Linked Data for Libraries, Archives and Museums. How to Clean, Link and Publish Your Matadata*. ISBN 978-0-8389-1251-5.

Oldman, D., Doerr, M, , de Jong, G., Norton, B., et Wikman, T., (2014). Realizing lessons of the last 20 years : A manifesto for data provisioning and aggregation services for the digital humanities. *D-Lib Magazine* 20(7/8).