Méthode pour enrichir sémantiquement les données en utilisant l'UML annoté

Sarra Ouelhadj[*,***], Pierre-Antoine Champin[*,**], Stéphanie Jean-Daubias[*]
Jérémy Gaillard[***]

[*] Univ Lyon, UCBL, CNRS, INSA Lyon, Centrale Lyon, Univ Lyon 2, LIRIS, UMR5205, F-69622 Villeurbanne, France
[**] Université Côte d'Azur, Inria, CNRS, I3S (UMR 7271), France
[***] Direction de l'Innovation Numérique et Systèmes d'Information, Métropole de Lyon, 20 rue du Lac, CS 33569, 69505 Lyon CEDEX 3, France

1 Introduction

Les données publiées sur le web posent des problèmes d'interopérabilité car elles sont syntaxiquement et sémantiquement hétérogènes, et leurs sémantique est implicite. L'intervention humaine est nécessaire pour la capturer, mais elle s'avère fastidieuse à l'échelle du web. Le domaine du Web Sémantique (WS) introduit RDF (Schreiber et Raimond, 2014) comme modèle de données standard fournissant une structure de liaison et une sémantique explicite des données pour les humains et les machines. Plusieurs outils et langages de mapping ont été proposés pour convertir les données de formats courants (ex. CSV) en RDF, mais la majorité requiert un certain degré d'expertise en WS, ce qui ralenti leur adoption. Nous proposons une méthode destinées à des experts métiers, pour convertir en RDF les données ouvertes de divers formats courants. Nous présentons les premiers retours des experts métiers vis-à-vis de cette méthode.

2 Approche d'enrichissement sémantique

Notre méthode permet aux experts métiers de produire des données RDF en maximisant la réutilisation des ontologies existantes. La Figure 1 illustre cette approche.

L'expert métier produit un diagramme de classes UML reflétant la sémantique du jeu de données en entrée. Ce diagramme est retranscrit dans notre *modèle sémantique*, un tableur structuré en 5 feuilles de calcul inspirées de la terminologie UML (Classes, Attributs, Énumérations, Valeurs d'énumération et Associations).

Ensuite, l'expert métier sélectionne dans des ontologies existantes les termes (IRI [1]) qui correspondent à la sémantique implicite de chaque élément UML, et inclut ces IRIs dans le modèle sémantique. En cas d'absence de terme adapté, de nouveaux IRIs sont forgés automatiquement dont l'expert métier doit alors décrire la sémantique dans le champs 'définition' du

1. Internationalized Resource Identifier

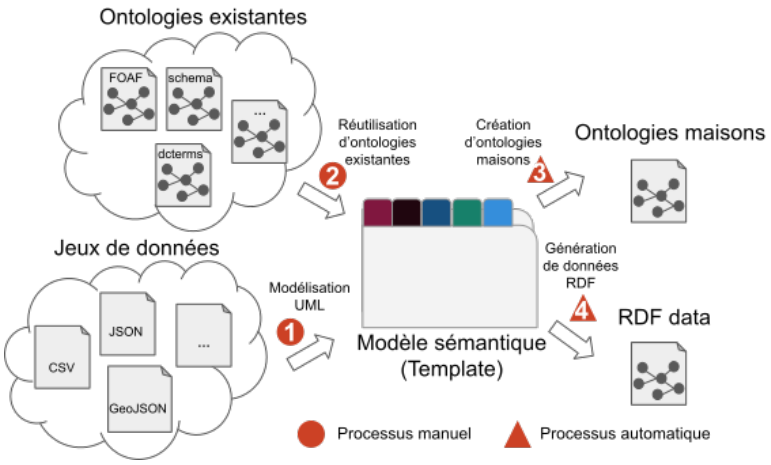

FIG. 1 – *Aperçu de l'approche d'enrichissement sémantique des données.*

modèle sémantique. Une ontologie maison est ainsi créée avec une approche bottom-up (Gandon, 2002) sujette à évoluer. Les détails d'implémentation sont disponibles dans le dépôt [2]. Enfin, les données peuvent être converties en données RDF utilisant les ontologies, existantes et maison, correspondantes.

3 Retours des experts métiers

Un atelier a été organisé auprès de 7 experts métiers de 4 départements distincts de la Métropole de Lyon. Les principales conclusions sont que l'approche de modélisation est à la portée des experts métiers, mais qu'ils ont besoin d'être assistés pour la sélection de termes dans les ontologies existantes. Pour cela, nous prévoyons de définir un processus de sélection et de maintenance d'un ensemble de vocabulaires partagés pertinents pour les experts métiers et de fournir certains outils pour y accéder (par exemple, une instance spécifique de LOV [3]).

Une autre remarque concerne la charge de travail importante que constitue la modélisation d'un jeu de de données conséquent. Nous comptons y répondre en insistant à l'avenir sur la possibilité de faire évoluer le modèle sémantique de manière incrémentale.

Références

Gandon, F. (2002). *Ontology Engineering : a Survey and a Return on Experience*. report, INRIA. https://hal.inria.fr/inria-00072192.

Schreiber, G. et Y. Raimond (2014). RDF 1.1 Primer. W3C Working Group Note, W3C. https://www.w3.org/TR/rdf11-primer/.

2. https://github.com/Sarra-Ouelhadj/YKWIM
3. https://lov.linkeddata.es/