Une approche de conception collaborative et d'exploitation des modèles ontologiques des données, facilement extensibles et compatibles avec le Web des Données Ouvertes (LOD) pour les Humanités Numériques (DH)

Hammou Fadili
fadili@msh-paris.fr
Fondation Maison des Sciences de l'Homme, France

Ahcène Ouguenoun
aouguenoune@adbi.fr
Accelerator Data & Business Intelligence, France

Résumé

Le but du présent article est de présenter une approche dont l'objectif est de mettre en place une plateforme générique, permettant la conception collaborative et l'exploitation de modèles ontologiques des données particuliers. Ils ont la particularité d'être facilement extensibles et compatibles avec le Web des Données Ouvertes *(Linked Open Data ou LOD),* destinée à être utilisée dans le domaine des humanités numériques *(Digital Humanities ou DH).* La démarche a été appliquée dans un premier temps à un instrument particulier : conception ontologique d'un wiktionnaire sémantique multilingue, multiculturel et multidisciplinaire des sciences humaines et sociales (SHS) afin d'une part de vérifier sur un exemple concret les fonctionnalités de la plateforme, et d'autre part de l'améliorer afin d'en faciliter la déclinaison à d'autres outils particuliers. En somme, le projet veut concevoir une fabrique de données intelligentes pour les humanités numériques *(Smart data factory for digital humanities)*; où la création des données suit un processus « cyclique », en deux étapes qui consistent (a) à créer directement dans la plateforme, par les experts du domaine, des données respectant toutes les normes exigées ; (b) à exploiter les données créées dans (a), en tant que données « expertes » validées, pour produire intelligemment et automatiquement, à partir de l'open data, de nouvelles données compatibles.

Introduction & motivation

L'objectif de ce travail vise la mise en place d'une plateforme centralisée d'aide à la conception collaborative de modèles ontologiques extensibles des données, facilitant la création et l'intégration des données interprétées et non ambiguës, dites données intelligentes *(Smart data)* au service des humanités numériques. Les contenus doivent être crées et générés sous forme de données structurées, sémantiquement annotées et liées, suivant des schémas de description bien adaptés. Dans notre cas, cela consiste à mettre en place un méta-modèle permettant de générer des modèles d'ontologies, des ontologies multilingues, multiculturelles et multidisciplinaires du domaine des SHS et une base de connaissances partagée et reconnue par des communautés de chercheurs.

Notre travail a été motivé par la fait que :

- Il n'existe pas suffisamment de données intelligentes, automatiquement exploitables, en SHS, à l'échelle internationale reflétant l'état de la coopération scientifique et culturelle entre la France et d'autres pays dont les concepts et lexiques pourraient évoluer de manières indépendantes
- Il y a un grand déséquilibre, d'un point de vue de la disponibilité des ressources numériques, entre le Français et les langues d'autres pays notamment celles des pays du sud
- Les dictionnaires multilingues, peu nombreux, sont souvent les résultats d'élaborations unilatérales
- Les traducteurs existants, également peu nombreux, ne prennent pas en compte tous les aspects liés aux contextes des définitions
- Les corpus multilingues pouvant constituer des sources de données sont également rares
- Les travaux sur les nouvelles technologies et la normalisation des données des langues de certains pays sont encore à leurs débuts

Dans cet article, la présentation de notre approche, se fera à partir de la description d'une contribution à la construction ontologique d'un Wiktionnaire sémantique multilingue, multiculturel et multidisciplinaire des SHS. Cette dernière est basée, entre autres, sur une adaptation et sur une extension de la plateforme collaborative « Mediawiki sémantique » existante afin qu'elle puisse prendre en compte nos modèles ontologiques des données. Sa construction devrait permettre aux chercheurs

d'échanger et de partager des connaissances dans le domaine des SHS et cela quelques soient leurs spécialités, leurs langues et leurs lieux géographiques de travail et/ou de résidence. Sa réalisation a intégré en plus, des normes et des protocoles bien spécifiques, en vue de son intégration dans le Web de Données Ouvertes (LOD).

Notre approche

Modèles des données

La conception de la structure de l'ontologie et donc du Wiktionnaire repose sur des correspondances entre les éléments de départ dans leurs contextes pour une langue source et les éléments d'arrivée dans leurs contextes pour une/des langue(s) cible(s) selon un sous-ensemble du schéma de la norme ISO1951. Pour simplifier, on va considérer les langues par paires. Donc, pour définir le modèle, on doit prendre en compte le fait qu'une entrée A_k dans une langue source peut avoir plusieurs sens et donc plusieurs traductions B_1, ..., B_j, ...B_m dans une langue cible. Cette même entrée A_k peut être définie avec plusieurs éléments A_1, ..., A_i, ...A_n du schéma du dictionnaire (cf. FIG. 2) qui peuvent êtres à leurs tour des entrées dans la même langue source et par conséquent, peuvent avoir plusieurs sens dans cette même langue source et plusieurs traductions dans la langue cible (FIG. 1). Notons que selon le sens de la traduction une langue source peut devenir cible et réciproquement.

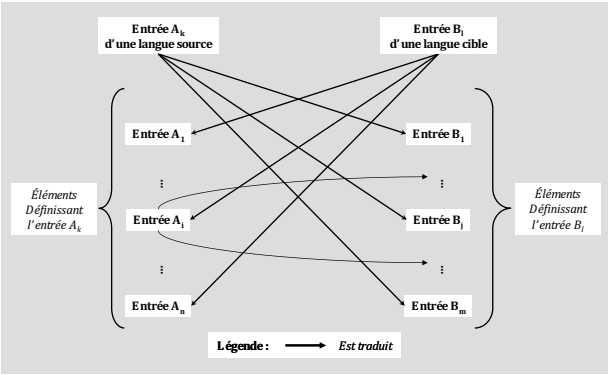


Figure 1. Extrait du schéma du dictionnaire des SHS

Ce schéma, récapitulant ce qui a été décrit précédemment, montre la complexité des renvois entre une langue source et une langue cible, ayant des spécificités différentes. Ce qui nécessite la définition d'un schéma spécifique dans chaque langue ainsi que la mise en place d'un système de gestion des correspondances d'une manière automatique. Nous pouvons procéder tout d'abord par une première simplification du problème de départ, qui consiste à associer à une entrée source (mot, locution, etc.) un ou plusieurs sens (définitions) qui renvoient à une ou plusieurs entrées cibles ; puis revenir du terme traduit, pris cette fois-ci comme entrée source. Ce procédé a été pris en charge par la mise en place d'un système de guidage d'aide à la génération et à la définition des correspondances entre les entrées, leurs définitions et leurs traductions dans les différentes langues. L'utilisateur peut modifier ou valider les suggestions du système pour compléter les fiches des entrées suivant les critères suivants (figure 2) :

- La définition d'une entrée se fait par la description d'une fiche suivant un format structuré déterminé par le schéma : contextes, définitions, relations sémantiques, traductions, indications grammaticales, parlers, etc.
- la signification attribuée à une entrée dépend d'un contexte de définition. Ce dernier est décrit par un ensemble fini et connu de paramètres contextuels des aspects : temporels, géographiques, disciplinaires, culturels, linguistiques, etc.
- Les relations entre les termes, se fait par le biais de relations sémantiques telles que : la synonymie, l'antonymie, l'hyperonymie, l'hyponymie, l'isonymie, la conversion, etc.

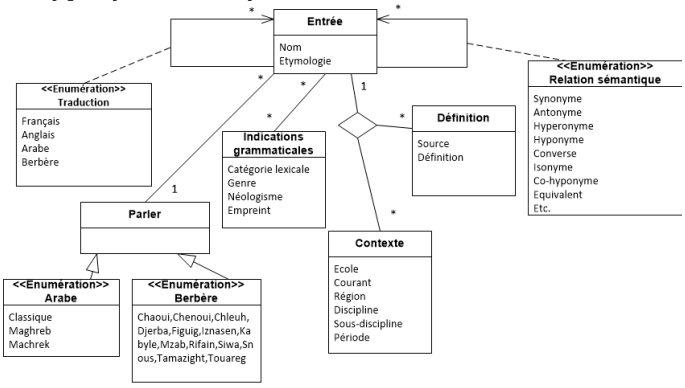
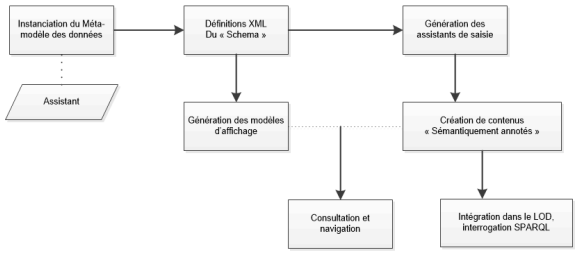


Figure 2.- Modèle du schéma du Wiktionnaire

Plateforme

L'architecture mise en place pour la réalisation repose sur une conception générique, capable de prendre en charge plusieurs modèles de données « *schémas de données*» et plusieurs langues ; et reste flexible et facilement extensible à d'autres schémas et généralisable à d'autres langues. L'implémentation a été réalisée, suivant 4 modules fondamentaux, assurant la gestion de la totalité du workflow. (figure 3).



Modules du processus de l'application Wiktionnaire sémantique des SHS

Figure 3. Processus général

Le processus consiste en l'instanciation du méta-modèle des données et la génération des différents schémas d'utilisation. Cela se fait par des définitions XML (qu'on peut générer grâce à des assistants dédiés) permettant la génération des modèles des données, des assistants de saisies et d'annotations ainsi que des modèles d'affichage. Le processus intègre également des modules d'exploitation et d'intégration des données dans le LOD (import/export des triplets RDF), ainsi qu'module de consultation et de navigation (exposition d'un point de terminaison SPARQL).

Exploitation

Après les phases de tests et de validation des modèles implémentés, l'application est actuellement en production. Elle est alimentée par un réseau d'enseignants chercheurs répartis suivant les disciplines et les langues de leurs spécialités. Elle mobilise de plus en plus de chercheurs et comporte actuellement plusieurs entrées dans plusieurs langues et disciplines. Il est également important de noter que certains éléments du wiktionnaire ont été enrichis, grâce à des requêtes SPARQL bien paramétrées sur du LOD (comme DBpedia).

Conclusion

La plateforme permet, d'une part, de créer des contenus « Intelligents » directement dans la plateforme, d'autre part, d'utiliser les données créées pour les confronter avec des sources de données externes de l'Open Data et du Web de données (Linked open Data), pour générer de nouvelles données Intelligentes. La solution développée étant générique, son extension et son adaptation à d'autres domaines et à d'autres langues est une tâche facile. Une des perspectives de ce travail consiste à encourager la dynamique actuelle, afin que la plateforme puisse devenir une référence dans le domaine des ressources ontologiques et dictionnairiques au service des Humanités Numériques.

Bibliographie

Berner-Lee, T. (2010). "Open, Linked Data for a Global Community." presented at the Gov 2.0 Expo 2010, May 26.

Harris, K. (2012). "Explaining Digital Humanities in Promotion Documents." Journal of Digital Humanities 1, no. 4

Grupo Tragsa. (2013) Smart Open Data. http://www.smartopendata.eu/

Linked Data Group (2009). Linked Data: Connect Distributed Data Across the Web. http://linkeddata.org/