Étude de cas sur DBpédia en français

Jungyeul Park, Mouloud Kharoune, Arnaud Martin

Étude de cas sur DBpédia en français

Jungyeul Park, Mouloud Kharoune, Arnaud Martin

UMR 6074 IRISA, Université de Rennes 1, Lannion, France
{jungyeul.park, mouloud.kharoune, arnaud.martin}@univ-rennes1.fr
http://www.iut-lannion.fr

Résumé. Dans ce papier, nous présentons une étude de cas de DBpédia en français. DBpédia a été considéré comme un concentrateur des données interconnectées. L'état actuel de DBpédia en français est encore à ses débuts, à la deuxième étape de développement selon notre critère . La première étape du projet a été accomplie, c'est-à-dire qu'il est identifié par les URIs et qu'il est décrit par les RDFs qui témoignent des données interconnectées ainsi que des ressources lisibles par les humains. La deuxième étape est à 44,65% de son avancement de l'appariement des ontologies au mois d'octobre 2013. Pour finir, étant donné l'état d'avancement actuel, nous proposons différentes démarches afin d'améliorer le développement du DBpédia en français.

1 Introduction

Le Web sémantique est un concept qui commence à prendre conscience de l'hétérogénéité des données dans le Web et de la difficulté de l'intégration et l'accessibilité entre celles-ci (Berners-Lee et al., 2001). L'objectif principal d'un tel système est de permettre aux machines d'interpréter sémantiquement les données du Web. Le principe est de faciliter le traitement et l'interprétation des informations issues des ressources du Web en les présentant sous des formes compréhensibles par une machine. Pour cela, le W3C a publié un langage d'ontologique pour le Web (OWL : Web Ontology Language) dans lequel on peut publier et partager une ontologie pour la construction et la gestion d'une base de connaissances sous une forme plus avancée (Harman, 2008). De plus, OWL est construit selon le modèle RDF (Resource Description Framework) qui est un format de données de référence pour définir la sémantique et les relations dans les ressources du Web. Dans ce contexte, DBpédia [1] peut être vue comme une source de données interconnectées (*Linked Data*) sous la forme d'une base de connaissances qui utilise une ontologie et le RDF. DBpédia est le résultat de tentatives d'extraction d'informations structurées à partir de Wikipédia, avec un identifiant unique (URI). Par exemple, si l'on accède à la ressource DBpédia sur `http://dbpedia.org/page/François_Hollande`, l'information structurée sous le format RDF est retournée. DBpédia est ainsi en train de jouer le rôle d'un concentrateur (*hub*) pour les données intercon-

1. `http://dbpedia.org`

nectées et est actuellement interconnecté avec d'autres bases de connaissances à grande échelle telles que YAGO2s [2], Freebase [3] etc.

Dans DBpédia, l'extraction de l'information est possible pour 119 langues de Wikipédia. DBpédia est accessible par SPARQL endpoint qui est un service acceptant des requêtes, comparables à SQL, pour interroger une base de connaissances en RDF. Nous constatons, que nous ne pouvons accéder par SPARQL endpoint de DBpédia qu'à seulement une partie parmi elles. Cela signifie que seul ces langues peuvent être utilisées via DBpédia depuis l'extérieur. Ceci est certainement lié à différentes raisons : d'une part, il est difficile de développer une version localisée de DBpédia pour les différentes langues. D'autre part DBpédia a été développé pour l'anglais y compris le cadre d'extraction d'information (DIEF) [4]. Le DIEF qui est un processus essentiel pour le développement de DBpédia, constitue un module d'extraction brute (dump) de Wikipédia. En effet, dans le cas de la langue grecque, des problèmes sont apparus dans le développement dès le début en raison de l'encodage de l'URI avec des caractères non latins mais également en raison de l'encodage des chaînes pour la navigation des ressources (Kontokostas et al., 2012).

DBpédia peut jouer le rôle d'un concentrateur des données interconnectées pour chaque langue, et il peut également produire des données dans les nuages *via* l'interconnexion avec d'autres données. Dans ce papier, nous analysons l'état actuel de DBpédia français en étudiant différentes étapes de progrès que nous définissons, pour proposer ensuite différentes démarches afin d'améliorer le développement de DBpédia français.

2 Données interconnectées

DBpédia a été introduit en 2007 par l'interaction entre l'extraction de données de Wikipédia et une meilleur gestion des bases de connaissances (Auer et al., 2007). Il est fondé sur le concept de données interconnectées proposé par Tim Berners-Lee. Au début, DBpédia a tenté d'extraire et de modifier les données structurées de Wikipédia dans le format RDF, en considérant le titre d'un article de Wikipédia comme une entité. YAGO, apparaissant simultanément à DBpédia, a utilisé la hiérarchie de synset de WordNet pour reconstruire une base de connaissances ontologiques comme un thésaurus avec une taxonomie qui permet de fournir les données les plus propres de Wikipédia (Suchanek et al., 2007, 2008).

D'autre part, une version antérieure de DBpédia a souffert de l'ambiguïté des entités et de l'incohérence des catégories de Wikipédia car elle a utilisé les données de Wikipédia sans modification. Ceci a diminué la fiabilité des données, ce qui a permis la découverte de références incomplètes sur d'autres données ouvertes interconnectées (LOD : Linked Open Data). Pour résoudre ces problèmes, DBpédia a défini l'ontologie de DBpédia en 2009 (Bizer et al., 2009). Ces efforts ont ainsi permis de clarifier les entités nommées et leurs propriétés grâce à l'appariement des ontologies en considérant que le modèle Wikipédia était une ontologie. Fondées sur les idées de DBpédia et YAGO, les ressources linguistiques peuvent aussi être utilisées pour améliorer les ressources les plus utiles du Web (Chiarcos et al., 2011).

Le projet des LOD linguistique (LLOD) a pour objectif principal l'ouverture de ressources linguistiques. Il promeut la recherche grâce au libre accès des ressources linguistiques, permet-

2. http://www.mpi-inf.mpg.de/yago-naga/yago
3. http://www.freebase.com
4. https://github.com/dbpedia/extraction-framework/wiki

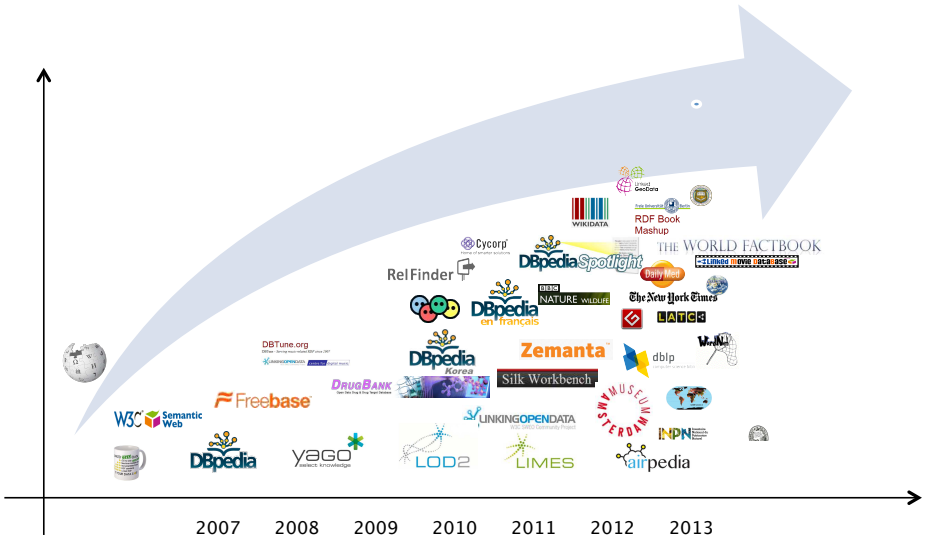

FIG. 1: DBpédia et les ressources interconnectées

tant également d'interconnecter les autres ressources existantes. Pour réaliser cet objectif, les ressources linguistiques ont été converties en RDF. En particulier, dans les meilleures pratiques pour le Web sémantique le groupe de travail WordNet [5] a fait des efforts pour convertir WordNet (`http://wordnet.princeton.edu`) en RDF et plusieurs travaux se sont engagés à interconnecter les DBpédia avec la version localisée de WordNet pour chaque langue (van Assem et al., 2006; Huang et Zhou, 2007; Koide et al., tted; Lim et al., 2013). DBpédia a constamment effectué l'interconnexion entre ses ressources et les LLOD. Dans ce contexte, les travaux de découvertes de références qui continuent depuis 2011 tels que SILK (Volz et al., 2009) et LIMES (Ngomo et Auer, 2011), se focalisent sur l'interconnexion automatique des ressources.

Récemment les outils d'annotation qui utilisent les ressources et l'ontologie de DBpédia ont été réalisés (Mendes et al., 2011). De plus les méthodes de l'appariement des ontologies sont aussi abordées par (Aprosio et al., 2013a,b). Enfin, la dernière version de DBpédia a montré que beaucoup d'efforts ont été apportés pour améliorer la qualité des données *via* l'extension de l'ontologie (Lehmann et al., tted). Ainsi la FIG. 1 montre DBpédia et la croissance de ses ressources interconnectées [6].

5. 'Semantic Web Best Practices : WordNet Task Force', `http://www.w3.org/2001/sw/BestPractices/WNET/tf`

6. `http://wiki.dbpedia.org/Interlinking`

3 DBpédia en français

3.1 État actuel

DBpédia en français a commencé à mettre des données en ligne [7] depuis 2009 (la version 3.2). L'équipe française a donc achevé une extraction des données de Wikipédia en RDF pour le développement de DBpédia en français, qui permet à présent sa consultation sur le site de DBpédia. Depuis 2011, DBpédia en français a effectué l'appariement des ontologies entre DBpédia et Wikipédia. Depuis lors, il est achevé avec la mise en place le SPARQL endpoint et l'interconnexion vers DBpédia en anglais, avec des pages lisibles par les humains. DBpédia devient particulièrement approprié pour les machines et on le considère comme une base de connaissances interconnectées. Pourtant, ces données ne sont pas lisibles par les humains, ce qui est contraire au principe des connaissances interconnectées. C'est pour cela qu'il est important que DBpédia fournisse les informations utiles pour les humains.

Au mois d'octobre 2013, DBpédia en français montre un taux de 10% entre les classes de DBpédia et les modèles de Wikipédia, qui représente 44.65% pour les documents. Tandis que le taux d'appariement des ontologies en français est assez haut, et se place même à la 8ème position parmi les autres langues, le taux d'appariement des modèles y compris leurs occurrences est relativement bas [8]. Une comparaison avec DBpédia en italien dans le tableau 1 le montre bien. Wikipédia en français et en italien contiennent un nombre similaire de documents. DBpédia en français montre effectivement un taux assez haut dans l'appariement des ontologies (10,08%) et un nombre important d'appariements (195). Toutefois, le taux d'appariement des occurrences des modèles n'est que de 44,65%. En ce qui concerne DBpédia en italien, ces taux sont, respectivement, de 5,86% avec 55 appariements, dont 79,41% de documents appariés. Les utilisateurs de Wikipédia en français ont tendance à créer leurs propres modèles au lieu d'utiliser ceux qui sont déjà fournis. Le nombre de modèles dans Wikipédia en français est deux fois plus important que celui en italien (1 935 vs. 939). En outre les modèles les plus utilisés dans Wikipédia en français ne sont pas appariés, comme par exemple `Modèle:Autres projets`, qui est utilisé pour plus de 200 000 de documents [9].

	l'appariement des modèles	l'appariement des documents
français	10,08% (195 de 1 935)	44,65% (579 700 de 1 298 362)
italien	5,86% (55 de 939)	79,41% (905 259 de 1 139 956)

TAB. 1: Comparaison entre DBpédia en français et en italien : taux de l'appariement des modèles et des documents

7. `http://fr.dbpedia.org`
8. `http://mappings.dbpedia.org/server/statistics/fr`
9. `http://fr.wikipedia.org/wiki/Modèle:Autres_projets`

3.2 Interconnexion au-delà de DBpédia en français

Pour améliorer DBpédia en français afin de l'introduire dans l'étape de l'interconnexion au-delà de DBpédia, nous pourrions proposer les points suivants :

— l'addition de l'appariement des ontologies entre Wikipédia et DBpédia en français.

— l'interconnexion avec les autres ressources en *français*.

Dans Wikipédia en français, les quatre modèles parmi les plus utilisés tels que `Modèle:Autres projets`, `Modèle:Ouvrage`, `Modèle:ÉluDébut`, et `Modèle:Lien` ne sont pas encore appariés. Ces modèles pourraient aider à augmenter le taux jusqu'à 13% de documents de Wikipédia.

L'*interconnexion* est en effet un processus essentiel des données interconnectées. Les ressources décrites en français doivent être interconnectées par DBpédia en français pour devenir le vrai centre des données interconnectées. Le projet DBpédia en français et Sémanticpédia [10] pour Wikipédia francophone ont déjà commencé à constituer une avancée majeure pour la politique du ministère de la Culture et de la Communication en faveur de l'accessibilité aux données culturelles [11]. Ils offrent ainsi aux musées, aux bibliothèques et même aux opérateurs culturels, des perspectives de diffusion et de partage des ressources extraits de Wikipédia en français.

L'autre axe de l'interconnexion pour DBpédia en français à développer est celui des ressources linguistiques pour la langue française. Les ressources linguistiques se distinguent par rapport à leur importance pour la communauté du Web sémantique. La fourniture de ces ressources en RDF et en données interconnectées est aujourd'hui une pratique bien établie. Elle pourrait ouvrir une voie des données interconnectées linguistiques pour DBpédia en français. Nous citons d'abord EuroWordNet [12] et WOLF [13] pour WordNet en français. EuroWordNet est un système de réseaux sémantiques pour les langues européennes, fondé sur WordNet. Chaque langue développe son propre WordNet, et elles sont interconnectées avec des liens interlingue dans l'Interlingual Index (ILI). Le WOLF est aussi une ressource lexicale sémantique, et il a été construit à partir de diverses ressources multilingues (Sagot et Fišer, 2008). Le L*efff* (Lexique des Formes Fléchies du Français) est un lexique morphologique et syntaxique à large couverture (Sagot, 2010) et il peut être couplé avec le WOLF [14]. Les données linguistiques du LADL (`http://infolingu.univ-mlv.fr`) sont aussi une ressource non négligeable pour la langue française : le système DELA qui comporte un lexique de mots simples (DELAS), un lexique associé de transcriptions phonétiques (DELAP), et un lexique de noms composés (DELAC). Il contient également le lexique-grammaire des phrases élémentaires du français ainsi que des grammaires locales représentant des phrases dans des domaines spécifiques.

4 Conclusion

Depuis la naissance des données interconnectées, un nombre important de personnes et d'organisations a adopté des données interconnectées comme un moyen de publier leurs propres

10. `http://www.semanticpedia.org`
11. `http://www.culturecommunication.gouv.fr/Actualites/A-la-une/Lancement-de-DBpedia-et-de-Semanticpedia`
12. `http://www.illc.uva.nl/EuroWordNet`
13. `http://alpage.inria.fr/~sagot/wolf.html`
14. `http://alpage.inria.fr/~sagot/lefff.html`

données. Ils placent ces données interconnectées sur le Web, mais les utilisent aussi dans le Web [15]. Dans ce papier, nous nous sommes intéressés à DBpédia qui a été considéré comme un concentrateur des données interconnectées. Nous avons analysé l'état actuel de DBpédia en français. Il en résulte que pour DBpédia en français, il est important que les ressources décrites en français doivent être interconnectées. Pour cela, nous nous sommes concentrés sur les ressources culturelles et linguistiques. Mais il serait également intéressant d'interconnecter les données gouvernementales de la France qui existent depuis 2011 [16]. Pour améliorer DBpédia y compris en français, l'interconnexion est essentielle car elle pourrait permettre à DBpédia de se développer en tant que base de connaissances au-delà de Wikipédia.

Références

Aprosio, A. P., C. Giuliano, et A. Lavelli (2013a). Automatic Expansion of DBpedia Exploiting Wikipedia Cross-Language Information. *The Semantic Web : Semantics and Big Data, Lecture Notes in Computer Science 7882*, 397–411.

Aprosio, A. P., C. Giuliano, et A. Lavelli (2013b). Automatic Mapping of Wikipedia Templates for Fast Deployment of Localized DBpedia Datasets. In *Proceedings of (i-KNOW 2013)*, Graz, Austria.

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, et Z. Ives (2007). DBpedia : A Nucleus for a Web of Open Data. *The Semantic Web, Lecture Notes in Computer Science 4825*, 722–735.

Berners-Lee, T., J. Hendler, et O. Lassila (2001). The Semantic Web. *Scientific American 284*(5), 28–37.

Bizer, C., J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, et S. Hellmann (2009). DBpedia - A Crystallization Point for the Web of Data. *Web Semantics : Science, Services and Agents on the World Wide Web 7*(3), 154–165.

Chiarcos, C., S. Hellmann, et S. Nordhoff (2011). Towards a Linguistic Linked Open Data cloud : The Open LinguisticsWorking Group. *Traitement Automatique des Langues (TAL) 52*(3), 245–275.

Harman, G. (2008). DeLanda's Ontology : Assemblage and Realism. *Continental Philosophy Review 41*(3), 367–383.

Huang, X.-x. et C.-l. Zhou (2007). An OWL-based WordNet lexical ontology. *Journal of Zhejiang University SCIENCE A 8*(6), 864–870.

Koide, S., H. Takeda, F. Kato, I. Ohmukai, F. Bond, H. Isahara, et T. Kuribayashi (Submitted). DBpedia and Wordnet in Japanese. *Semantic Web Journal*.

Kontokostas, D., C. Bratsas, S. Auer, S. Hellmann, I. Antoniou, et G. Metakides (2012). Internationalization of linked data. the case of the greek dbpedia edition. *Web Semantics : Science, Services and Agents on the World Wide Web 15*, 51–61.

Lehmann, J., R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, et C. Bizer (Submitted). DBpedia - A Large-scale, Multi-

15. `http://www.w3.org/2001/tag/doc/selfDescribingDocuments.html`
16. `http://www.data.gouv.fr`

lingual Knowledge Base Extracted from Wikipedia. *Semantic Web - Interoperability, Usability, Applicability*.

Lim, K., Y. Hahm, M. Rezk, J. Park, Y. Yongun, et K.-S. Choi (2013). Enrichment of DBpedia by Linking Korean WordNet and Improving Web Resource Accessibility. *Journal of KIISE : Computing Practices and Letters 19*(9), 474–478.

Mendes, P. N., M. Jakob, A. García-Silva, et C. Bizer (2011). DBpedia Spotlight : Shedding Light on the Web of Documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, New York, NY, USA, pp. 1–8. ACM.

Ngomo, A.-C. N. et S. Auer (2011). LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Barcelona, Catalonia, pp. 2312–2317.

Sagot, B. (2010). The Le*fff*, a Freely Available and Large-coverage Morphological and Syntactic Lexicon for French. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, et D. Tapias (Eds.), *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Sagot, B. et D. Fišer (2008). Building a Free French WordNet from Multilingual Resources. In *Proceedings of OntoLex 2008*, Marrakech, Morocco.

Suchanek, F. M., G. Kasneci, et G. Weikum (2007). YAGO : A Core of Semantic Knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, New York, NY, USA, pp. 697–706. ACM.

Suchanek, F. M., G. Kasneci, et G. Weikum (2008). YAGO : A Large Ontology from Wikipedia and WordNet. *Web Semantics : Science, Services and Agents on the World Wide Web 6*(3), 203–217.

van Assem, M., A. Gangemi, et G. Schreiber (2006). Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, pp. 237–242.

Volz, J., C. Bizer, M. Gaedke, et G. Kobilaro (2009). Silk – A Link Discovery Framework for the Web of Data. In *Proceedings of WWW2009 workshop : Linked Data on the Web (LDOW2009)*, Madrid, Spain.

Summary

We present a case study on French DBpedia based on its progress on Linked Data. DBpedia has been considered as the hub for Linked Data. We analyze the current state of French DBpedia where it can be uniquely identified by URIs and be described in RDF, which bespeaks a web of data as well as human readable resources. It has also achieved 44.6% of all template occurrences on October, 2013 for ontology mapping. Based on the current state of progress, we suggest some guidelines of how to improve French DBpedia.