Extraction de données biographiques depuis Wikipedia

Robert Viseur

CETIC
Rue des Frères Wright, 29/3
B-6041 Charleroi
robert.viseur@cetic.be

UMONS (FPMs)
Rue de Houdain, 9
B-7000 Mons
robert.viseur@umons.ac.be

RÉSUMÉ. *L'utilisation du contenu des articles de Wikipedia est fréquente dans les recherches académiques. Les modalités pratiques d'exploitation sont cependant rarement analysées. Notre recherche porte sur l'extraction de données biographiques relatives à des personnalités originaires de Belgique. Notre recherche sera organisée en trois sections. Une première section proposera un état de l'art en matière d'extraction de données dans l'encyclopédie Wikipedia. Une seconde section présentera le cas pratique de l'extraction de données biographiques de personnalités belges. Différentes solutions seront discutées et la solution retenue sera mise en œuvre. Dans une troisième section, la qualité de l'extraction sera discutée. Des recommandations pratiques à destination des chercheurs souhaitant exploiter Wikipedia seront en outre proposées sur la base de notre cas pratique.*

ABSTRACT. *Using the content of Wikipedia articles is common in academic research. However the practicalities are rarely analyzed. Our research focuses on extracting biographical information about personalities from Belgium. Our research is organized into three sections. The first section provides a state of the art for data retrieval in Wikipedia. A second section presents the case study about data mining for biographical Belgian personalities. Different solutions are discussed and the adopted solution is implemented. In the third section, the quality of the extraction is discussed. Practical recommendations for researchers wishing to use Wikipedia are also proposed on the basis of our case study.*

MOTS-CLÉS : *wikipedia, dbpedia, biographie, fouille de texte, open data.*

KEYWORDS: *wikipedia, dbpedia, biography, text mining, open data.*

1. Introduction

Wikipedia (wikipedia.org) est une encyclopédie multilingue et collaborative lancée en 2001. Le projet est depuis 2003 soutenu financièrement par la Wikimedia Foundation (wikimediafoundation.org). Le volume de l'encyclopédie n'a cessé de grandir depuis sa création. En janvier 2013, les plus grosses éditions de Wikipedia étaient les éditions anglophone (plus de quatre millions d'articles), germanophone (plus d'un million et demi d'articles), francophone (plus d'un million trois cent mille articles) et néerlandophone (plus d'un million cent mille articles).

Ces dernières années, les recherches académiques et les exemples pratiques d'utilisation du contenu de Wikipedia se sont multipliés. Hu *et al.* (2009) l'ont exploité pour améliorer les performances d'un système de clustering de documents. Kazama *et al.* (2007) ou Charton *et al.* (2010) l'ont exploité pour améliorer la reconnaissance d'entités nommées. Buscaldi et Rosso (2006) ont amélioré les performances d'un système de réponse à des questions (« *Question Answering technology* »). La BBC l'a utilisé pour permettre l'interconnexion des informations présentes dans ses bases de données internes et leur enrichissement par des sources de données externes (Kobilarov *et al.*, 2009). La requête « *Exploiting Wikipedia* » exécutée dans le moteur de recherche scientifique Google Scholar (scholar.google.fr) renvoie ainsi plus de 22 mille résultats !

Notre recherche concerne pour sa part l'extraction de données biographiques relatives à des personnalités originaires de Belgique. L'utilisation de Wikipedia pour l'alimentation d'une base de données biographiques apparaît appropriée, si l'on s'en réfère à la répartition par type de contenu au sein de l'encyclopédie. Les articles relatifs aux biographies et aux personnes représentaient en effet, en janvier 2008, 15% du contenu total, derrière les articles relatifs à la culture et aux arts (Kittur *et al.*, 2009).

Plusieurs questions se posent cependant.

- Les éditions francophone, germanophone et néerlandophone de Wikipedia sont intéressantes, car ces langues constituent les trois langues nationales belges. Il est cependant difficile, sur cette base, d'identifier la volumétrie du contenu concernant la Belgique, plutôt que la France, l'Allemagne ou les Pays-Bas.

- De nombreux articles exploitent le contenu de Wikipedia. Cependant, peu donnent des indications quant aux difficultés pratiques liées à l'extraction de données dans Wikipédia. Une extraction réussie suppose de savoir identifier les articles pertinents mais aussi d'être capable d'extraire les données souhaitées depuis le contenu des articles.

Notre recherche sera organisée en trois sections. Une première section proposera un état de l'art en matière d'extraction de données dans l'encyclopédie Wikipedia. Une seconde section présentera le cas pratique de l'extraction de données biographiques de personnalités belges. Différentes solutions seront discutées et la solution retenue sera mise en œuvre. Dans une troisième section, la qualité de l'extraction sera discutée. Des recommandations pratiques à destination des

chercheurs souhaitant exploiter Wikipedia seront en outre proposées sur la base de notre cas pratique.

2. État de l'art

L'extraction de biographie a déjà été réalisée par Biadsy *et al.* (2008). L'approche adoptée par les auteurs est cependant différente de la nôtre, puisqu'ils ont développé un système de résumé multi-document, construit sur un classificateur de phrases biographiques et d'un composant permettant l'ordonnancement des phrases jugées d'intérêt. Ils se sont basés sur les articles de biographies utilisant le template de Wikipédia dédié aux biographies, soit près de 17000 articles. Le traitement s'est opéré sur la copie XML de Wikipédia disponible en ligne.

En pratique, l'utilisation des copies XML n'est pas la seule manière pour manipuler le contenu de l'encyclopédie. D'une part, l'extraction des informations à l'aide d'outils de rétroingénierie directement sur les pages publiées en ligne reste possible. D'autre part, une version structurée de Wikipédia est proposée depuis 2007: DBpedia.

DBpedia (dbpedia.org) est un effort communautaire qui a démarré en 2007 (Auer *et al.*, 2007). Il vise à extraire des informations structurées de Wikipedia et à rendre ces informations disponibles sur Internet. Le processus d'extraction s'appuie sur les copies des bases de données Wikipedia (« *database dump* »). Les données sont actualisées grâce à l'utilisation du flux référençant les mises à jour de l'encyclopédie (Hellmann *et al.*, 2009). L'extracteur s'appuie sur le contenu des articles et, surtout, celui des Infobox associés aux articles. Les Infobox apparaissent sous forme de tableaux dans le coin supérieur droit de nombreux articles et présentent des informations factuelles.

Le contenu extrait depuis l'encyclopédie est converti dans le format RDF. Plusieurs mécanismes d'accès sont proposés pour explorer DBpedia: l'accès aux données RDF directement par URI (Universale Resource Identifier), l'utilisation d'agents Web (exemple: navigateurs pour le Web sémantique) et les points d'accès SPARQL permettant l'interrogation de DBpedia au moyen d'un langage évoquant le SQL utilisé pour les bases de données relationnelles.

DBpedia apparaît comme une solution partielle pour l'extraction de données dans les contenus de Wikipedia. Certes, la facilité d'interrogation permise par le langage SPARQL pour l'identification des articles utiles en fait un outil séduisant. Cependant, DBpedia présente plusieurs limitations.

Premièrement, la couverture linguistique de DBPedia est actuellement limitée à 13 langues (voir « *International DBpedia chapters* », dbpedia.org). Lors de sa création en 2007, DBpedia était uniquement proposé en langue anglaise. Un projet pour la langue française a été lancé fin 2012. Baptisé Sémanticpédia (www.semanticpedia.org), il associe depuis fin 2012 le Ministère français de la culture et de la communication, Wikimédia France et l'Inria, pour la production d'une version française de DBpedia (fr.dbpedia.org).

Deuxièmement, le processus d'extraction s'appuie principalement sur le contenu des Infobox (Auer *et al.*, 2007 ; Hellmann *et al.*, 2009). Or, un examen rapide d'articles Wikipedia permet de constater que toutes les pages de l'encyclopédie ne proposent pas d'Infobox, et que ces derniers ne sont pas toujours complets. Une part de l'information présente dans les articles échappe donc aux extracteurs. Néanmoins, à sa création, DBpedia revendiquait déjà près de 2 millions de références (Auer *et al.*, 2007).

3. Cas pratique: extraction de données biographiques de personnalités belges

3.1. *Identification des articles pertinents*

Nous avons tout d'abord comparé deux approches: d'une part, l'interrogation de DBpedia depuis les points d'accès anglophones et francophones et, d'autre part, l'identification des articles pertinents à l'aide de techniques de crawl sur le site de l'encyclopédie.

L'interrogation du DBpedia anglophone et du DBpedia francophone a été réalisée avec une requête en langage SPARQL, grâce à la propriété `birthPlace` (égale à « Belgique » pour la langue française et « Belgium » pour la langue anglaise).

L'identification des biographies des personnalités belges se fait en deux étapes. La première étape prend comme point de départ la page Wikipedia relatives aux personnalités belges (`http://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Personnalit%C3%A9_belge`), pointée depuis le portail belge de Wikipedia (`http://fr.wikipedia.org/wiki/Portail:Belgique`). Elle consiste en un crawl récursif de cette page et des pages de catégories suivantes de manière à identifier les pages de catégories contenant des personnalités belges. Ce mécanisme permet de trouver plus de 700 catégories pertinentes. Les URLs de ces catégories sont sauvegardées. La seconde étape explore ensuite les pages de catégories et identifie les articles Wikipédia dédiés aux personnalités belges. Les URLs de ces fiches sont sauvegardées dans un fichier. Plus de 10.000 articles sont collectés grâce à cette méthode (voir Table 1).

	Nombre de résultats
DBpedia (en)	899
DBpedia (fr)	200
Wikipedia (fr)	10.884

Table 1. Nombre d'articles trouvés par méthode

La volumétrie de la méthode classique par crawl dans Wikipédia plutôt que par interrogation de DBpedia se révèle donc nettement plus intéressante.

3.2. *Extraction des données dans le texte*

3.2.1. *Processus d'extraction*

Une copie des articles est sauvegardée en local. En pratique, nous travaillerons sur la version des articles dans le format propre à Wikipedia. Cette version est accessible via des URLs de la forme `http://fr.wikipedia.org/w/index.php?action=raw&title=xxxxx` et permet d'obtenir un texte brut (texte + syntaxe Mediawiki) exempt de balises HTML et débutant par un Infobox lorsque ce dernier existe.

L'analyse du texte brut passe par deux opérations. La première consiste à extraire l'Infobox, lorsque ce dernier existe. Le seconde consiste à identifier les phrases de la biographie susceptibles de contenir des informations biographiques importantes telles que la date de naissance, la date de mort et la profession. En pratique, la première phrase de l'article est systématiquement retenue, car elle contient conventionnellement les informations les plus importantes sur la personne. Elle est éventuellement complétée par une seconde phrase, en cas de correspondance avec un ensemble de mots-clefs déclencheurs. Ce traitement aboutit à une biographie condensée, qui est sauvegardée pour chaque article. Ces biographies condensées passent ensuite par un jeu d'expressions régulières pour en extraire la date de naissance, la date de mort (si la personne est morte) et la profession de la personne. Ces données structurées sont stockées dans un fichier CSV.

Ce fichier contient 10610 entrées, avec les champs suivants: nom, date de naissance, date de mort, profession, URL de la catégorie et URL de l'article au format HTML (voir Table 2). Sur un total initial de 10884 articles, 57,6% autorisent une extraction de la date de naissance, 26,9%, de la date de décès et 56,3%, des professions. Pour ces dernières, la catégorie de l'article apporte généralement une information de substitution en cas d'extraction ratée (les catégories indiquent souvent une profession ou une fonction dans la société). Seules 27,4% des fiches possèdent un Infobox.

Nombre d'articles:	10884	100,0%
Nombre d'Infobox:	2980	27,4%
Nombre de biographies condensées:	10610	97,5%
Nombre d'extractions réussies		
Dates de naissance:	6269	57,6%
Dates de mort:	2936	26,9%
Métiers:	6129	56,3%

Table 2. Volumétries (processus d'extraction)

3.2.2. *Difficultés rencontrées*

Nous avons principalement rencontré quatre difficultés.

Premièrement, les articles ne sont accompagnés d'un Infobox que dans moins d'un cas sur trois. Il en résulte une obligation de passer par des techniques d'analyse de texte pour parvenir à extraire les dates (naissance, décès) et professions. L'extraction des dates est particulièrement délicate, car les articles comprennent souvent d'autres dates (dates liées à des événements importants dans la vie des personnes). L'extraction fait appel à un ensemble d'expressions régulières, dont l'écriture est délicate pour le non-spécialiste.

Deuxièmement, même lorsqu'un Infobox est présent, le nom des champs des Infobox n'est pas homogène. La date de naissance pourra ainsi être annoncée par `date_naissance`, `date naissance`, `date de naissance`, `date_de_naissance` ou encore `naissance`. Un regroupement préalable est donc nécessaire. Cela ne présente pas de grande difficulté technique.

```
([[Bree]], [[12 avril]] [[1876]] – [[Ixelles]], [[14 septembre]] [[1953]])
([[Pétange]], {{Date de naissance|12|juillet|1817}} - Pétange, {{Date de décès|14|mai|1898}}])
né le [[12 janvier]] [[1597]] à [[Bruxelles]] ([[Belgique]]) et mort le [[12 juillet]] [[1643]] à [[Livourne]] ([[Italie]])
'''Ellen Petri''' (née le 25 mai [[1982]], [[Merksem]] ([[Anvers]]))
'''Paul Deschanel''', né le {{date|13|février|1855}} à [[Schaerbeek]] ([[Bruxelles]]) et décédé le {{date|28|avril|1922}} à [[Paris]]
'''Robert Gruslin''' né à [[Rochefort (Belgique)|Rochefort]] le [[18 mars]] [[1901]], décédé à [[Profondeville]] le {{1er juin}} [[1985]]
```

Table 3. Hétérogénéité des formats de dates

Troisièmement, les formats de dates ne sont pas homogènes, que ce soit dans le texte ou dans les Infobox (voir Table 3). Les dates peuvent être écrites avec des chiffres uniquement, avec le mois écrit en lettres ou encore être complétées par d'autres informations comme le lieu de naissance ou le type d'activité pour laquelle la personne s'est faite remarquer.

Quatrièmement, la sélection préalable de phrases candidates pour l'extraction de données nécessite une mise en œuvre plus fine que la technique mise en œuvre ici. Un classificateur tel que mis en œuvre par Biadsy *et al.* (2008) mériterait un investissement pour améliorer les performances globales de l'extraction.

3.2.3. *Taux d'erreur*

L'évaluation a été réalisée sur un ensemble de 2980 entrées (il s'agit d'entrées comprenant un Infobox). Les dates de naissance extraites dans le texte des articles ont été comparées aux dates de naissances fournies dans les Infobox. Le contenu des Infobox est structuré. L'extraction est donc sensiblement simplifiée, et les données extraites pourront être considérées comme exemptes d'erreurs d'extraction.

Nombre total d'éléments	2980	100,0%	
Pas de comparaison possible	1336	44,8%	
Nombre d'Infobox sans date	743	24,9%	
Comparaison possible	1644	55,2%	100,0%
Dates identiques	1486		90,4%
Dates différentes	158		9,6%
Information partielle	126		7,7%
Erreur d'extraction	32		1,9%

Table 4. Taux d'erreur d'extraction (date de naissance)

Une comparaison a été faite entre les données extraites dans le texte des articles Wikipédia et les données extraites dans les Infobox (voir Table 4). Le test a été réalisé sur 2980 dates de naissance (100%). La comparaison a pu être réalisée sur 1644 dates pour lesquelles la donnée était présente dans l'Infobox et dans le résultat de l'extraction depuis le texte de l'article. Des dates différentes sont constatées dans 9,6% des cas. Cependant, 7,7% des dates sont correctes mais l'information est incomplète. Typiquement, l'année de naissance a été extraite, mais pas la date complète (exemple: `mai 1988` vs `1988`). L'information extraite du texte peut être plus complète que celle extraite de l'Infobox. L'information peut être présente dans le texte et pas dans l'Infobox.

La présence de l'information dans l'Infobox et pas dans le texte est par contre due à des erreurs d'extraction. En pratique, l'information donnée dans l'Infobox semble toujours présente dans le texte. Ce constat permet d'obtenir une borne inférieure du taux d'extractions ratées dans le texte, soit 19,9%.

Près des deux tiers des personnes sont nés après 1900 (63,1% des dates de naissance données dans les Infobox). Le faible nombre de dates de décès serait donc dû à l'âge moyen des personnes fichées plutôt qu'à d'éventuelles erreurs d'extractions dans le texte des articles Wikipedia.

Cette méthode pose deux difficultés. D'une part, les formats de dates peuvent différer entre données extraites du texte et données extraites des Infobox (exemple: `8 mars 1965` vs `8 03 1965`). Une méthode de conversion des dates est donc nécessaire pour en homogénéiser le format. Des tags Mediawiki ainsi que des informations complémentaires peuvent par ailleurs accompagner la date de naissance (exemple: `date_de_naissance = [[28 juillet en sport|28 juillet]] [[1982 en football|1982]]`). D'autre part, la structure des Infobox n'est pas normalisée et le nom des champs peut varier d'une fiche à l'autre.

4. Discussion et perspectives

Ce travail d'extraction avait été initié avec l'a priori qu'un usage de DBpedia permettrait d'obtenir facilement, via le langage de requête SPARQL, les données biographiques souhaitées. Un premier test a permis de constater que la volumétrie disponible dans DBpedia était sensiblement moindre que ce que l'on pouvait obtenir à partir de techniques classiques de crawl et d'extraction dans Wikipedia. Le projet DBpedia reste intéressant, voire indispensable, pour le chercheur participant à des

projets de Linked Data ou souhaitant disposer d'un vocabulaire contrôlé. Cependant, il montre actuellement ses limites en matière d'exhaustivité sur des thématiques précises.

L'existence d'un projet comme DBpedia ainsi que la visibilité des données structurées au travers des Infobox peut donner l'impression que Wikipedia se prête facilement à l'extraction de données. Il ressort cependant de notre expérimentation que, d'une part, les Infobox sont loin d'être systématiques (moins de 30% des articles considérés en possèdent) et que, d'autre part, la structure des Infobox n'est pas totalement homogène. L'existence d'un ensemble de conventions, sous la forme de balisages ou de tournures de phrases, en matière de dates ou de professions, rend cependant l'extraction réalisable à partir du contenu des articles, sans nécessiter l'usage de techniques sophistiquées.

6. Références

Auer S., Bizer C., Kobilarov G., Lehmann J., Cyganiak R., Ives Z., « DBpedia: A Nucleus for a Web of Open Data », *Lecture Notes in Computer Science*, Vol. 4825, 2007, pp 722-735.

Bekavac B., Tadić M., « A Generic Method for Multi Word Extraction from Wikipedia », *Proceedings of the Int. Conf. on Information Technology Interfaces*, June 23-26, 2008.

Biadsy F., Hirschberg J., Filatova E., « An Unsupervised Approach to Biography Production using Wikipedia », *Proceedings of ACL-08: HLT*, 2008, pp. 807–815.

Buscaldi D., Rosso P., « Mining Knowledge from Wikipedia for the Question Answering task », *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, 2006.

Charton E. Gagnon M., Ozell B., « Extension d'un système d'étiquetage d'entités nommées en étiqueteur sémantique », *TALN 2010*, 19–23 juillet 2010.

Hellmann S., Stadler C., Lehmann L., Auer S., « DBpedia Live Extraction », *Lecture Notes in Computer Science*, Vol. 5871, 2009, pp 1209-1223.

Hu X., Zhang X., Lu C., Park, E. K., Zhou, X., « Exploiting Wikipedia as external knowledge for document clustering », *KDD '09 Proceedings of the 15th international conference on Knowledge discovery and data mining*, 2009.

Kazama J., Torisawa K., « Exploiting Wikipedia as External Knowledge for Named Entity Recognition », *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, June 2007, pp. 698–707.

Kittur A., Chi E.H., Suh B., « What's in Wikipedia?: Mapping Topics and Conflict using Socially Annotated Category Structure », *Proceedings of the 27th international Conference on Human Factors in Computing Systems*, April 04-09, 2009.

Kobilarov G., Scott T., Raimond Y., Oliver S., Sizemore C., Smethurst M., Bizer C., Lee R., « Media meets Semantic Web - How the BBC uses DBpedia and Linked Data to make Connection », *ESWC 2009*, 2009, pp. 723-737.