



# Exploiter les données issues de Wikipedia

Rencontres Mondiales  
du Logiciel Libre  
*Développement logiciel*

**Beauvais (France) – 09 juillet 2015**

*robert.viseur@cetic.be*

FEDER



UNION EUROPÉENNE



Wallonie



LE FONDS EUROPEEN DE DEVELOPPEMENT REGIONAL  
ET LA WALLONIE INVESTISSENT DANS VOTRE AVENIR.

- Je suis : Dr Ir Robert VISEUR.
- Ingénieur civil, Docteur en sciences appliquées de la Faculté Polytechnique de l'UMONS ([www.umons.ac.be](http://www.umons.ac.be)).
- Assistant dans le Service de Management de l'Innovation Technologique de la Faculté Polytechnique de l'UMONS([www.umons.ac.be](http://www.umons.ac.be)).
- Senior R&D Expert au CETIC ([www.cetic.be](http://www.cetic.be)).
- Photographe indépendant ([www.derrierelevisueur.be](http://www.derrierelevisueur.be)).

- Contexte général :
  - Usage de plus en plus fréquent du contenu de Wikipédia dans les domaines techniques et scientifiques (classification de documents, REN, création d'URI, etc.).
  - Plus de 22 mille résultats pour la requête « Exploiting Wikipédia » dans Google Scholar (`scholar.google.fr`).
- Contexte interne :
  - Demande d'une entreprise pour l'aider...
  - à créer d'une base de données biographique depuis Wikipedia (personnalités belges).
- Recherche menée principalement au CETIC, avec le soutien de l'UMONS (FPMs).

- Ce qui est présenté ici = travail d'évaluation.
- Cinq étapes principales :
  - Identification des articles pertinents.
  - Extraction des données depuis le texte.
  - Inventaire des difficultés rencontrées.
  - Évaluation de la qualité de l'extraction.
  - Évaluation de la fiabilité des données.

- En exploitant les informations structurées.

**Yves Leterme**



Yves Leterme, en 2011.

**Fonctions**

**39<sup>e</sup> Premier ministre belge**  
(67<sup>e</sup> chef du gouvernement)

**25 novembre 2009 - 6 décembre 2011**  
(2 ans, 0 mois et 13 jours)

<b>Monarque</b>	Albert II
<b>Gouvernement</b>	Leterme II
<b>Législature</b>	52 <sup>e</sup> législature
<b>Prédécesseur</b>	Herman Van Rompuy
<b>Successeur</b>	Elio Di Rupo

**Ministre fédéral des Affaires étrangères**  
17 juillet - 25 novembre 2009

<b>Premier ministre</b>	Herman Van Rompuy
<b>Prédécesseur</b>	Karel De Gucht
<b>Successeur</b>	Steven Vanackere

**37<sup>e</sup> Premier ministre belge**

## Paul Aerts



**Cet article est une ébauche concernant un coureur cycliste belge**

Vous pouvez partager vos connaissances en l’améliorant (**comment** ?) Pour plus d’informations, voir la page d’aide.

**Paul Aerts** (né le **16 décembre 1949** à **Beerse**) est un ancien **coureur cycliste** professionnel belge.

**Palmarès** [modifier]

- **1971**
  - 2<sup>e</sup> du *Tour du Brabant*
  - 2<sup>e</sup> du *Circuit des frontières*
- **1972**
  - Grand Prix du Tournaisis
  - 2<sup>e</sup> du *Tour du Luxembourg*
  - 2<sup>e</sup> du *Circuit des bords de l'Escaut*
  - 2<sup>e</sup> du *Circuit des frontières*

**Résultats sur le Tour de France** [modifier]

- 1972 : 69<sup>e</sup>
- 1973 : abandon (9<sup>e</sup> étape)

**Lien externe** [modifier]

- Fiche de Paul Aerts sur *le Site du cyclisme*

## Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

Query Text

```
select distinct * where {?s rdfs:label ?l} LIMIT 100
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see [details](#).)

Results Format: HTML

Execution timeout:  milliseconds (values less than 1000 are ignored)

Options:  Strict checking of void variables

(The result can only be sent back to browser, not saved on the server, see [details](#).)

Run Query Reset

**Informations**

<b>Nom</b>	Aerts
<b>Prénom</b>	Paul
<b>Date de naissance</b>	16 décembre 1949 (63 ans)
<b>Pays</b>	<span style="font-size: small;">Belgique</span>

**Équipe professionnelle**

<b>1971-1973</b>	Watney
<b>1974-1975</b>	Ijsboerke - Colner
<b>1976</b>	Zoppas - Splendor - Sinalco

[modifier](#)

- Comparaison :

- ~~Interrogation d'une copie de base de données Wikipédia (via les dumps publics).~~
- Accès par crawl des catégories (portail Belgique -> Personnalités belges) vs...
- Accès par requête SPARQL (exploitation de la propriété « `birthPlace` » dans DBpedia).

---

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

---

**Table 1.** Nombre d'articles trouvés par méthode


- Accès au texte des articles par URL du type `http://fr.wikipedia.org/w/index.php?action=raw&title=xxxxx`.
- Extraction du texte de l'article et de l'Infobox (si l'article en possède un).
- Extraction depuis le texte des dates de naissance et de décès, ainsi que des professions.


```
1 [[Image:Andre Cools.jpg|thumb|André Cools]]
2 '''André H.P. Cools'''
   ⚙️ {{{date|1|août|1927}}} -
   ⚙️ {{{date|18|juillet|1991}}} est un
   ⚙️ [[Personnalité politique|homme politique]]
   ⚙️ [[Parti socialiste (Belgique)|socialiste]]
   ⚙️ [[Belgique|belge]] et un [[Mouvement
   ⚙️ wallon|militant wallon]] assassiné à
   ⚙️ [[Cointe]] ([[Liège]]).
3
4 ==Biographie==
5
6 Il fut ministre du Budget de [[1968]] à
   ⚙️ [[1971]], [[vice-Premier ministre]] de
   ⚙️ [[1969]] à [[1972]], président du [[Parti
   ⚙️ socialiste belge]] de [[1973]] à [[1978]],
   ⚙️ puis président du [[Parti socialiste
   ⚙️ (Belgique)|Parti socialiste]] de [[1978]]
   ⚙️ à [[1981]], et président du [[Parlement
   ⚙️ wallon]] de [[1982]] à [[1985]]. Il reçut
   ⚙️ aussi le titre honorifique de [[ministre
   ⚙️ d'État]] en [[1983]]. Il reçut le grade de
   ⚙️ Grand Officier de l'[[Ordre de Léopold]]
   ⚙️ et celui de Grand-Croix de l'[[Ordre de
   ⚙️ Léopold II]]. Il siégea pendant la [[47e
   ⚙️ législature de la Chambre des
   ⚙️ Représentants de Belgique]]. Au moment de
   ⚙️ sa mort, il était [[Bourgmestre
   ⚙️ (Belgique)|bourgmestre]] de [[Flémalle]],
   ⚙️ commune de la banlieue liégeoise, et
   ⚙️ ministre wallon des Pouvoirs locaux et
```

- L'analyse du texte se fait par la mise en œuvre d'un jeu d'expressions régulières exploitant des tournures de phrases typiques.
- Exemples : « né à ... », « naquit à ... », « est un ... », etc.
- Les outils standards d'extraction d'entités nommées ou d'étiquetage grammatical n'ont pas été utilisés.



- Une minorité d'articles dispose d'un Infobox.
- L'information est donc moins structurée qu'elle ne peut le sembler au départ.
- Les propriétés des Infobox ne sont elles-mêmes pas totalement standardisées.
- Exemple : les dates de naissance apparaissent avec différents labels (→ folksonomie).

Créer un compte  Connexion

Rechercher  


---

s.

---

iverses impliquant l'écurie

[Julien Lahaut](#) » alors qu'il est  
à l'on emprisonne, même



---

e

- L'extraction doit être mise en œuvre sur le texte par essais et erreurs en exploitant des tournures de phrases typiques.
- Le format de date est un bel exemple de l'hétérogénéité constatée dans le formatage de l'information au sein de l'encyclopédie.

---

```
[[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

- Volumétries suite au processus d'extraction :

---

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)

- Evaluation de la qualité de l'extraction par la comparaison entre données extraites dans le texte / extraites dans les Infobox.

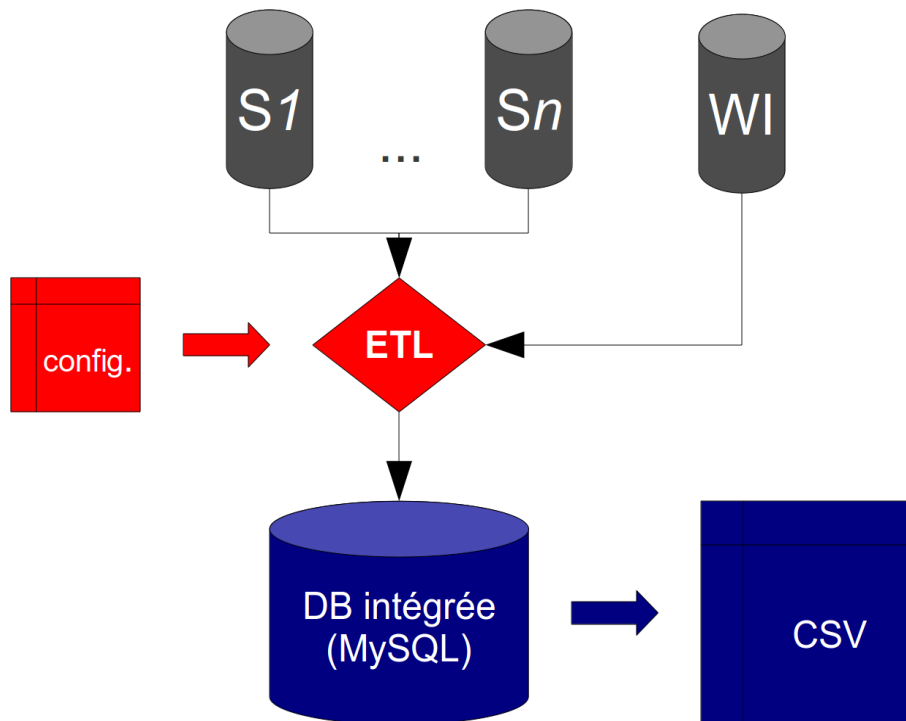
---

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)

- Comparaison des données extraites de Wikipédia avec des données de référence.



- Création d'une liste fusionnée (938 lignes)

Albert Bruylants	0	1915	0	0	1915	0	0	0	0	0	1
Thomas Buffel	0	1981	0	0	0	0	0	1981	0	0	1
Auguste Buisseret	0	1888	0	0	1888	0	0	0	0	0	1
Charles Buls	0	1837	0	0	1837	0	0	0	0	0	1
Ernest Bumelle	0	1908	0	0	1908	0	0	0	0	0	1
Jan Burssens	0	1925	0	0	1925	0	0	0	0	0	1
Max Buset	0	1896	0	0	1896	0	0	0	0	0	1
Yoni Buvens	0	1988	0	0	0	0	0	1988	0	0	1

- Différences de valeurs sur 14,4% des lignes.
  - → Problème des homonymies...
    - → Vérification manuelle...

- Taux d'erreur :
  - Taux d'erreur dans Wikipedia = 0,75%.
  - Taux d'erreur dans les sources de référence = 0,21%.

---

Erreur dans Wikipedia	0,75%
Erreur d'extraction	1,71%
Erreur dans la source de référence	0,21%
Indeterminé	0,75%

---

**Table 5.** Taux d'erreur (date de naissance)

- Evaluation par comparaison à des sources de référence (sites de musées, de fondations,...).

- Envisageable : automatiser la détection des données (potentiellement) erronées.
- Moyen : utiliser les critères de qualité des articles dans Wikipédia.
- Exemples : nombre de mots, nombre d'éditeurs distincts, nombre d'éditions, etc.
- Voir (Blumenstock, 2008), (Chevalier *et al.*, 2010), (Stvilia *et al.*, 2005), (Wilkinson et Huberman, 2007), etc.



- Le projet Dbpedia, version sémantique de Wikipédia, donne une image de structuration et d'exhaustivité. Cette image est partiellement trompeuse.
- Wikipédia est un projet basé sur les contributions des utilisateurs, et souffre d'un manque de structuration et d'homogénéisation pour en faciliter l'exploitation.



- L'exploitation du texte des articles peut heureusement être abordée avec des techniques simples (jeu d'expressions régulières) grâce à la structure typique des articles et des phrases.
- Résultat obtenu :
  - Précision : ~90%.
  - Rappel : ~80%.

- La fiabilité des données paraît fort satisfaisante (> 99%).
  - Limitation : test réalisé sur des personnalités encodées dans plusieurs bases de données, donc probablement populaires ( → davantage de révision par les pairs ?).
- Pas de sureprésentation de personnalités contemporaines (comparé aux sources de référence).
  - Moyenne (date de naissance) : 1880.
    - Sources de références : 1878.
  - Ecart-type (date de naissance) : 156.
    - Sources de références : 66.

- Evaluer l'intérêt d'utiliser des outils spécialisés pour l'extraction d'entités nommées.
  - Plus d'infos sur les outils open source d'extraction de terminologie dans Viseur (2013b, 2014a).
- Evaluer les possibilités liées à l'utilisation de Wikidata ([www.wikidata.org](http://www.wikidata.org)).

- Robert Viseur (2015), « Utiliser Wikipédia pour la création d'une base de données biographiques : mise en œuvre et étude des limitations », « Wikipédia, objet scientifique non identifié », Presses universitaires de Paris Ouest, 978-2-84016-205-6.
- Robert Viseur (2014b), « Reliability of User-Generated Data: the Case of Biographical Data in Wikipedia », « OpenSym », Berlin, Germany.
- Robert Viseur (2014a), « Automating the Shaping of Metadata Extracted from a Company Website with Open Source Tools », « International Journal of Advanced Computer Science and Applications ».
- Robert Viseur (2013c), « Extraction of Biographical Data from Wikipedia » in « Data », Reykjavik, Islande.
- Robert Viseur (2013b), « Presentation of OpenNLP », « Rencontres Mondiales du Logiciel Libre (RMLL) », Université Libre de Bruxelles, Bruxelles, juillet 2013.
- Robert Viseur (2013a), « Extraction de données biographiques depuis Wikipedia », « InforSID », Paris, France.

# Thanks

## Contact :

*robert.viseur@cetic.be*

## Plus d'information :

*www.roberviseur.be*

*twitter.com/roberviseur*

*www.linkedin.com/in/roberviseur*



Aéropôle de Charleroi-Gosselies  
Rue des Frères Wright, 29/3  
B-6041 Gosselies  
**info@cetic.be**

**[www.cetic.be](http://www.cetic.be)**

