Exploiter les données issues de Wikipédia

Webinaire M2Communs

Séance #2 : Apprentissage en ligne et production collective de connaissances

Université Paris Ouest Nanterre La Défense - Paris (France) - 25 février 2016

robert.viseur@cetic.be

Qui suis-je ?

- Je suis : Dr Ir Robert VISEUR.
- Ingénieur civil, Mastère en Management de l'Innovation, Docteur en sciences appliquées de la Faculté Polytechnique de l'UMONS (`www.umons.ac.be`).
- Affiliations :
  - Assistant dans le Service de Management de l'Innovation Technologique de la Faculté Polytechnique de l'UMONS(`www.umons.ac.be`).
  - Senior R&D Expert au CETIC (`www.cetic.be`).
- Expertises : management de l'innovation, co-création, open source (modèles d'affaires, licences,…), technologies de traitement de l'information, évaluation des technologies,...
- Photographe indépendant (`www.derriereleviseur.be`).

Qu'a-t-on fait avec Wikipédia ? (1/2)

- 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 Wikipedia » dans Google Scholar (`scholar.google.fr`).
- Contexte interne :
  - Demande d'une entreprise (2013) pour l'aider...
  - à créer une base de données biographiques depuis Wikipedia (personnalités belges).
- Recherche menée principalement au CETIC, avec le soutien de l'UMONS (FPMs).
- Plusieurs communications et publications scientifiques (voir « Références ») à la suite de cette recherche.

Qu'a-t-on fait avec Wikipédia ? (2/2)

- 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.

Comment pensait-on pouvoir procéder ?

- En exploitant les informations structurées.

Yves Leterme

Yves Leterme, en 2011.

Fonctions

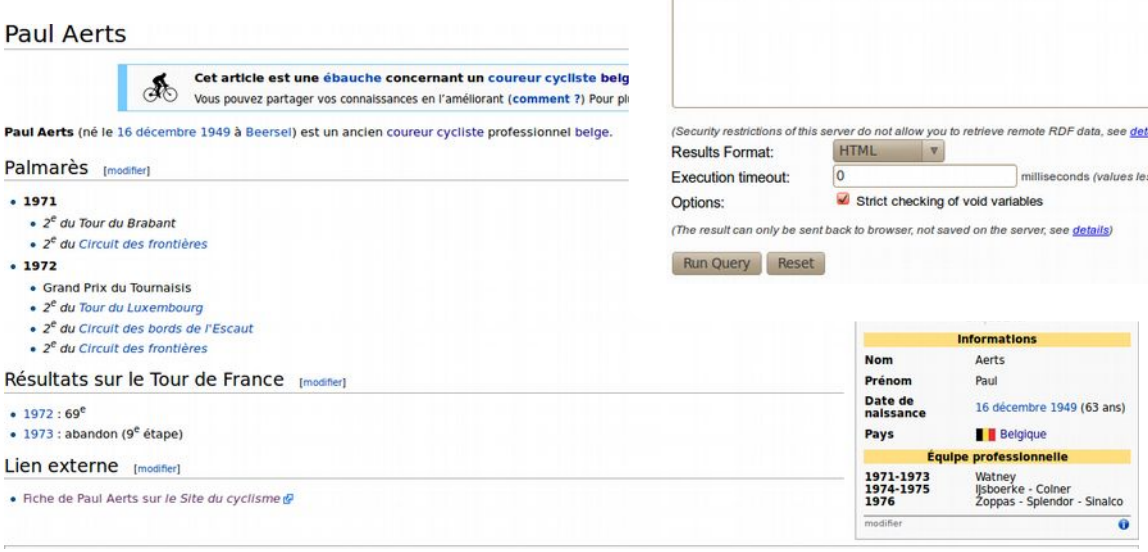
39e Premier ministre belge
(67e chef du gouvernement)
25 novembre 2009 - 6 décembre 2011
(2 ans, 0 mois et 13 jours)

Monarque	Albert II
Gouvernement	Leterme II
Législature	52e législature
Prédécesseur	Herman Van Rompuy
Successeur	Elio Di Rupo

Ministre fédéral des Affaires étrangères
17 juillet - 25 novembre 2009

Premier ministre	Herman Van Rompuy
Prédécesseur	Karel De Gucht
Successeur	Steven Vanackere

37e Premier ministre belge



Paul Aerts

Cet article est une ébauche concernant un coureur cycliste belg
Vous pouvez partager vos connaissances en l'améliorant (comment ?) Pour pl

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

Palmarès [modifier]

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

Résultats sur le Tour de France [modifier]

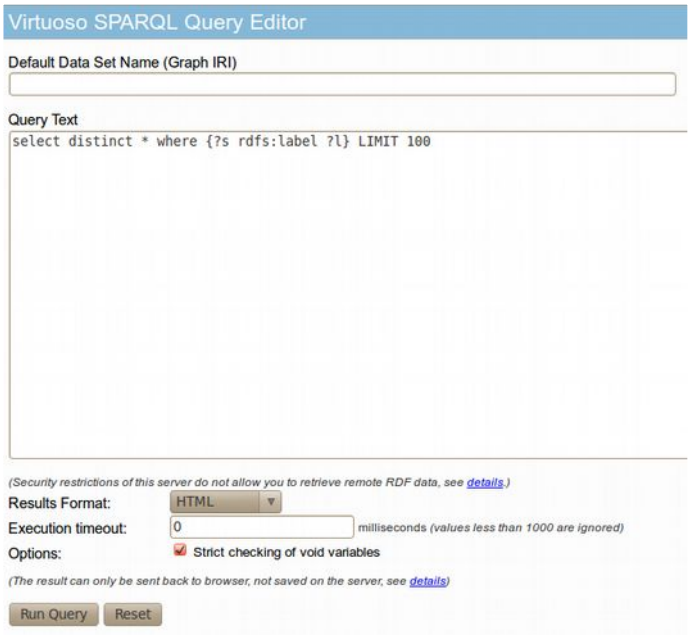
- 1972 : 69e
- 1973 : abandon (9e étape)

Lien externe [modifier]

- Fiche de Paul Aerts sur le Site du cyclisme

Informations	
Nom	Aerts
Prénom	Paul
Date de naissance	16 décembre 1949 (63 ans)
Pays	Belgique
Équipe professionnelle	
1971-1973	Watney
1974-1975	Ijsboerke - Colner
1976	Zoppas - Splendor - Sinalco

modifier



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: 0 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

Étape 1 : identification des articles

- 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

Étape 2 : extraction des données depuis le texte (1/2)

- 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.

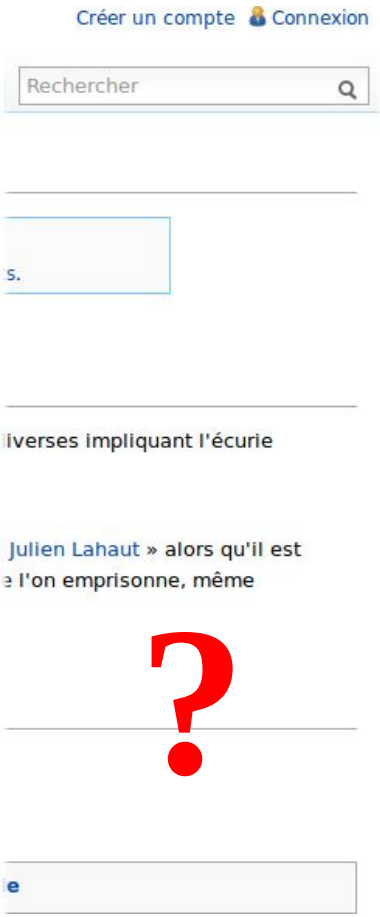
```
[[Image:Andre Cools.jpg|thumb|André Cools]]
'''André H.P. Cools'''
({{date|1|août|1927}} -
{{date|18|juillet|1991}}) est un
[[Personnalité politique|homme politique]]
[[Parti socialiste (Belgique)|socialiste]]
[[Belgique|belge]] et un [[Mouvement
wallon|militant wallon]] assassiné à
[[Cointe]] ([[Liège]]).

==Biographie==

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
```

Étape 2 : extraction des données depuis le texte (2/2)

- 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.

Étape 3 : inventaire des difficultés rencontrées (1/2)

- 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).



Étape 3 : inventaire des difficultés rencontrées (2/2)

- 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

Étape 4 : évaluation de la qualité de l'extraction (1/2)

- 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)

Étape 4 : évaluation de la qualité de l'extraction (2/2)

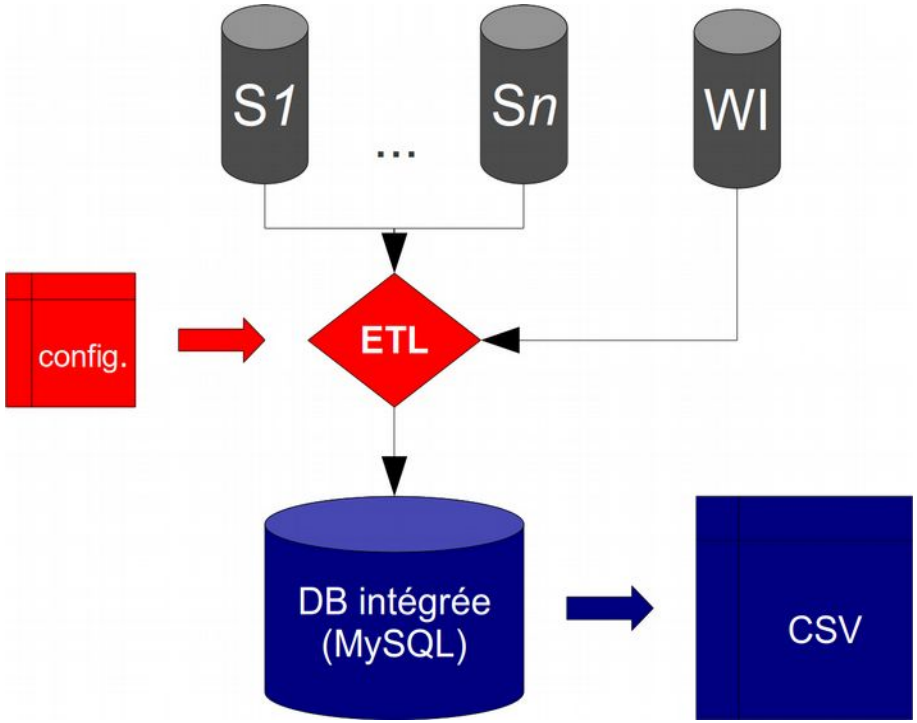
- Evaluation de la qualité de l'extraction par la comparaison entre données extraites dans le texte / extraites dans les Infobox.

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)

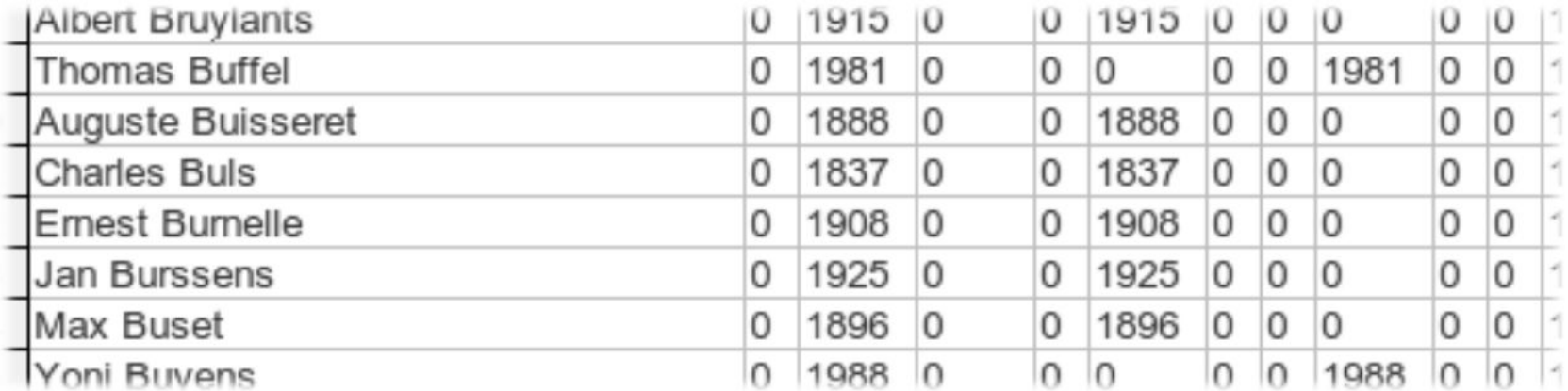
Étape 5 : évaluation de la fiabilité des données (1/4)

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



Étape 5 : évaluation de la fiabilité des données (2/4)

- Création d'une liste fusionnée (938 lignes)



Albert Bruyiants	0	1915	0	0	1915	0	0	0	0	0
Thomas Buffel	0	1981	0	0	0	0	0	1981	0	0
Auguste Buisseret	0	1888	0	0	1888	0	0	0	0	0
Charles Buls	0	1837	0	0	1837	0	0	0	0	0
Ernest Burnelle	0	1908	0	0	1908	0	0	0	0	0
Jan Burssens	0	1925	0	0	1925	0	0	0	0	0
Max Buset	0	1896	0	0	1896	0	0	0	0	0
Yoni Buyens	0	1988	0	0	0	0	0	1988	0	0

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

Étape 5 : évaluation de la fiabilité des données (3/4)

- 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,…).

Étape 5 : évaluation de la fiabilité des données (4/4)

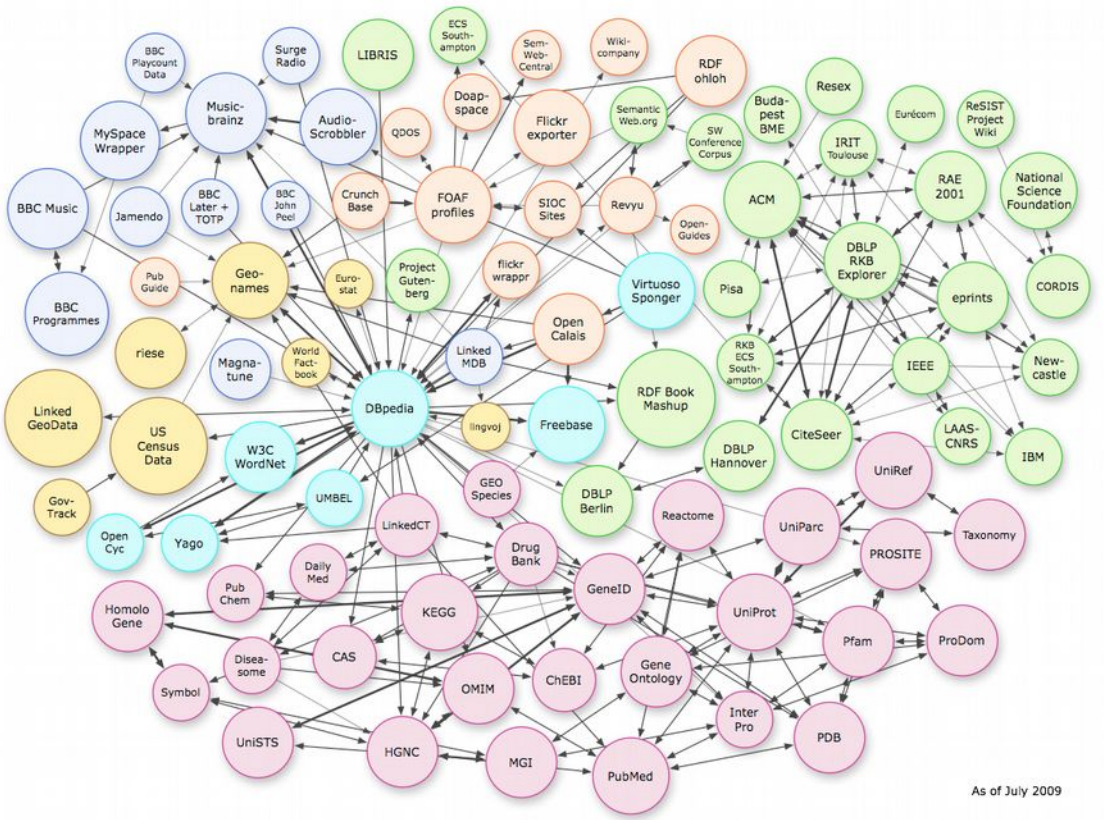
- Envisageable : automatiser la détection des données (potentiellement) erronées.
- Moyen : utiliser les critères de qualité des articles dans Wikipédia.
- Exemples : nombre de mots, nombre d'éditeurs distincts, nombre d'éditions, etc.
- Voir (Blumenstock, 2008), (Chevalier *et al.*, 2010), (Stvilia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

Que conclure ? (1/4)

- 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.

Que conclure ? (2/4)

- Dbpedia reflète cette caractéristique. Dbpedia reste cependant une excellente base pour des opérations de « linked data ».



(source : `data.culture.fr`)

Que conclure ? (3/4)

- 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%.

Que conclure ? (4/4)

- 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 surepresentation 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.

Quelles sont les perspectives ?

- 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`).
  - Projet d'externalisation de l'acquisition de données (« *Wikipedia for data* »), sous licence CC0. Lancé en octobre 2012, devenu un des projets les plus actifs de la Wikimedia Foundation.
  - Traitements automatiques (« *bots* ») et humains (éditions).
  - Enjeu important : réconcilier les 287 éditions linguistiques de Wikipédia.
  - Pluralité de valeurs possibles par propriété !
  - Aussi partie du Web des données (URI).
  - Voir (Vrandečić et Krötzsch, 2014) pour plus d'informations.

Références

- 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.

Merci

Contact :

robert.viseur@cetic.be

Plus d'information :

www.robertviseur.be
twitter.com/robertviseur
www.linkedin.com/in/robertviseur

cetic
Your Connection to ICT Research

Aéropôle de Charleroi-Gosselies
Avenue Jean Mermoz, 28
B-6041 Gosselies
info@cetic.be

www.cetic.be