

SemanticPedia, le DBpedia francophone.

Alexandre Monnin

INRIA (Wimmics)

Institut de Recherche et d'Innovation du Centre Pompidou

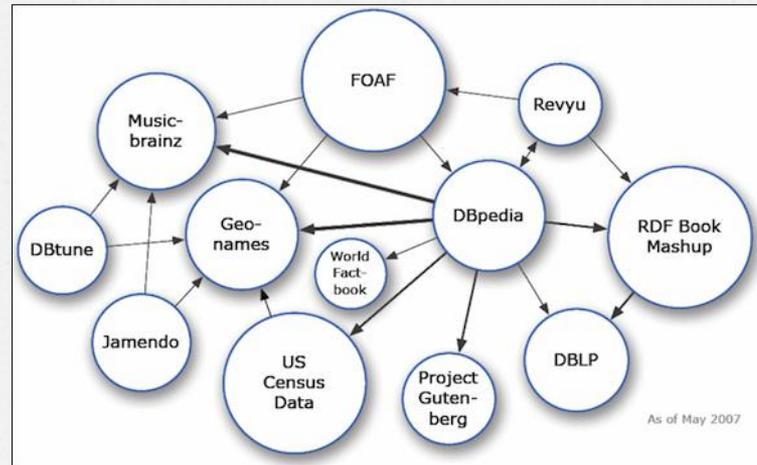
Université Paris 1 Panthéon-Sorbonne (PhiCo, EXeCo)

CNAM (DICEN)

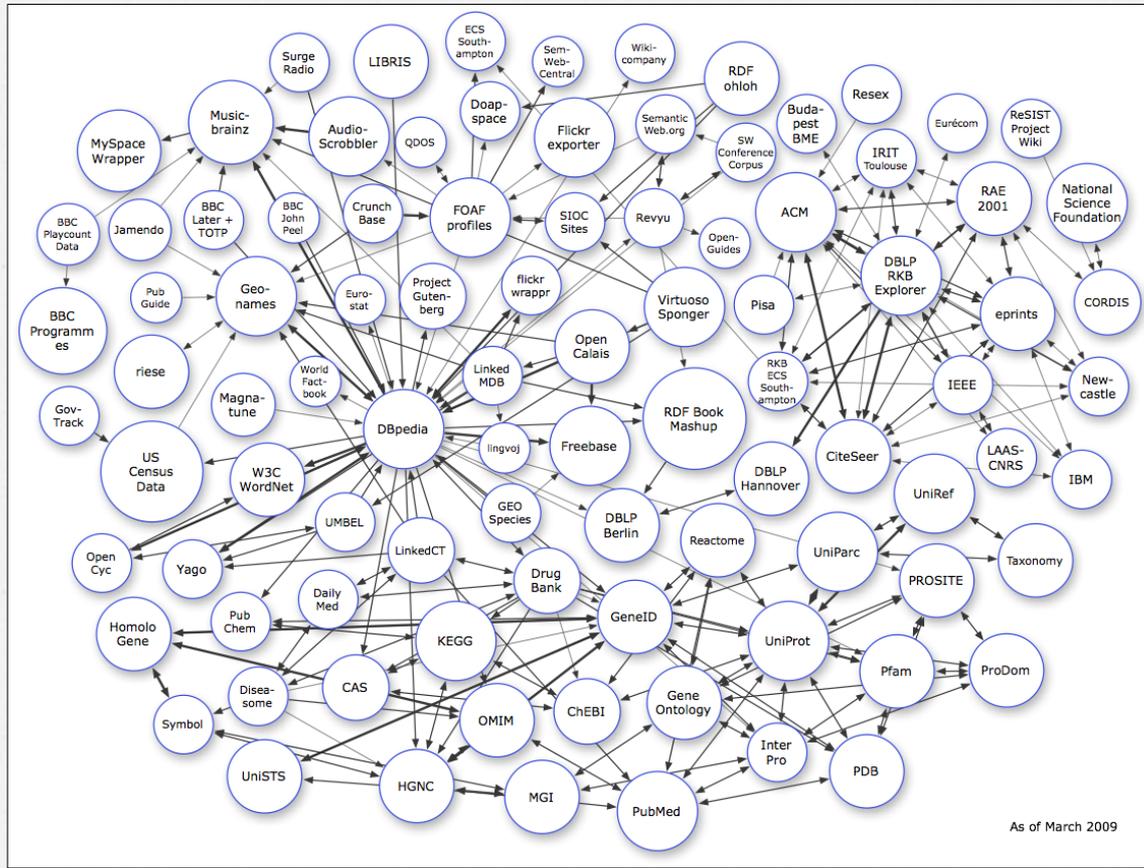
Dbpedia.org :
principale application du Web de
données



1-05-2007



27-03-2009



Trois acteurs



Pourquoi un nouveau nom ?

Aucune volonté de faire original à tout prix.

- Les outils de DBpedia sont réutilisés,
- L'ingénieur en charge du projet, Julien Cojan, est le représentant francophone de DBpedia.org,
- Les données de SemanticPedia sont disponibles sur fr.dbpedia.org

Mais alors, pourquoi ???

On note également des différences :

- Wikimedia France a ouvert un « Lab » consacré à la sémantisation de ses outils. Wikipédia n'est que le premier de la liste,
- Le fait est qu'une association comme Wikimedia France, dont le but premier est le soutien à Wikipédia et aux contributeurs de Wikipédia, met désormais à disposition des outils qui vont être utiles à Wikipédia, aux contributeurs, et plus globalement qui aident à la diffusion et réutilisation de l'encyclopédie.
- Le projet comprend enfin une part d'innovation propre

Techniquement



Extracteurs réutilisés

- ◊ LabelExtractor
- ◊ GeoExtractor
- ◊ InfoboxExtractor
- ◊ PageLinksExtractor
- ◊ WikiPageExtractor
- ◊ MappingExtractor
- ◊ HomepageExtractor

Extracteurs employés sans modifier le code-source

- o ArticleCategoriesExtractor
- o CategoryLabelExtractor
- o ExternalLinksExtractor
- o PageIdExtractor
- o RedirectExtractor
- o RevisionIdExtractor
- o SkosCategoriesExtractor

Extracteurs pour lesquels de nouveaux paramètres ont été ajoutés

- DisambiguationExtractor
- ImageExtractor
- InterLanguageLinksExtractor
- PndExtractor
- PersondataExtractor

Innovation, retours d'expériences et améliorations



Un projet mûr

Depuis 2006 les nombreux retours d'expériences permettent d'envisager de nouvelles fonctionnalités.

- o A titre d'exemple, l'amélioration de la pérennité des URIs : d'après Yves Raimond* (BBC) lors de chaque mise à jour du dump de DBpedia, 20% des liens deviennent caducs. Il est possible de faire mieux comme en témoigne Dbpedia lite (et le projet HDA de l'IRI).
- o L'historique et les pages de discussions font également partie du périmètre du projet pour mieux refléter le caractère contributif de Wikipédia.

dbpedia lite

[dbpedia lite](#) takes some of the structured data in [Wikipedia](#) and presents it as [Linked Data](#). It contains a small subset of the data that [dbpedia](#) contains; it does not attempt to extract data from the Wikipedia infoboxes. Data is fetched live from the [Wikipedia API](#).

Unlike dbpedia it uses stable Wikipedia pageids in its URIs to attempt to mitigate the problems of article titles changing over time. If the title of a Wikipedia page changes, the dbpedia lite URI will stay the same. This makes it safer to store dbpedia lite identifiers in your own database.

Search

Bookmarklet

Drag this [dbpedialite flipr](#) link to your bookmarks bar in your browser. Now, when you're on Wikipedia page click on the bookmarklet to switch to the corresponding dbpedia lite page. When you're on a dbpedia lite page, click on the bookmarklet to take you back to the respective Wikipedia page.

Example Things

- [U2](#)
- [U2 \(band\)](#)
- [Florence + the Machine](#)
- [Stoat](#)
- [Ermine](#)
- [Ceres, Fife](#)
- [Berlin](#)

Example Categories

- [Towns in Fife](#)
- [James Bond films](#)

Data Dumps

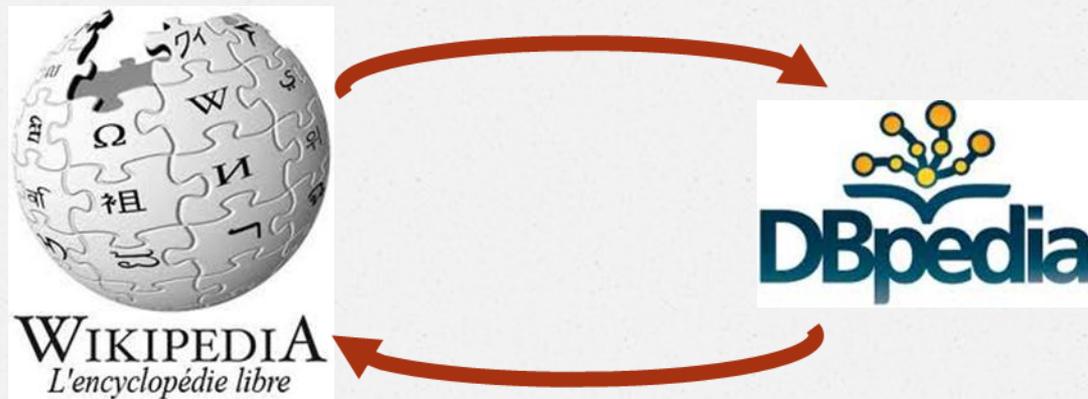
Although there is currently no data dump of the whole of dbpedia lite available, there is an [owl:sameAs](#) mapping to dbpedia URIs available at [downloads.dbpedialite.org](#). This dump was created from an [Wikipedia XML article stubs](#) export. The mapping has been loaded into [sameas.org](#). See [Tim Berners-Lee](#) as an example.

Powered by: [heroku](#), [Sinatra](#), [RDF.rb](#) and [more...](#)
Source code: <http://github.com/njh/dbpedialite>



Dernièrement, un alignement entre les données de DBpedia lite, DBpedia, Freebase a été entrepris sur DBpedia lite. Cela fait également parti des possibilités envisagées.

Quels liens entre Wikimédia France et SemanticPedia ?



L'enrichissement va dans les deux sens

- o Le traitement des *templates*, *infobox* et autres licences sur Wikipédia est complexe. L'aide de la communauté s'avère indispensable.
- o A l'inverse, les premiers dumps de SemanticPedia ont permis de mesurer le nombre d'entrées en français ne disposant pas de liens interlangues avec d'autres versions de Wikipédia en langues étrangères : plus de 400 000 sur un corpus de 1 200 000 articles (soit 1/3 du total !).

Créer un écosystème viable



- o Parmi les retours d'expériences glanés certains l'ont été durant le séminaire muséologie organisé par l'IRI et financé par le Ministère de la Culture,
- o Le projet HDA-BO & Lab entend sensibiliser les institutions culturelles aux apports du tagging sémantique grâce à SemanticPedia/Dbpedia tout en proposant un enrichissement du corpus de Wikipédia
- o L'IRI et Alcatel Bell Labs lancent un projet pour le mois de septembre autour de la sémantisation des données « sociale » de Wikipédia avec pour objectif de mettre au point un système de suivi des articles de Wikipédia et de notification afin de s'assurer de la pertinence d'une URI de DBpedia dans le temps.

Conclusion



- o De plus en plus d'institutions culturelles travaillent avec Wikimedia France, il y a fort à parier que le mouvement se poursuivra de plus belle grâce SemanticPedia.
- o L'apport des entreprises est attendu pour enrichir l'écosystème décrit tantôt, par exemple en connectant SemanticPedia à d'autres corpus et ontologies, plus spécialisés.

P.S. : un vœu



Espérons que l'enrichissement demeurera mutuel, tant que faire se peut !

Merci !*

**Tout particulièrement à Bertrand
Sajus, Fabien Gandon, Julien Cojan,
Adrienne Alix et Thibault Grouas.*