



Wikipédia et les données

Ateliers de l'Information Grenoble, 22 mars 2016

Guillaume Allègre

Galaxie des projets Wikimedia



























De quelles données parle-t-on?

- Pas des contenus en général (documents, textes pédagogiques, photos...)
- Bases / banques de données
 - données structurées
 - données nombreuses, voire exhaustives
 - adaptées au traitement automatique recherche structurée ⇒ requête ⇒ langage de requête

Les données dans Wikipédia

Pourquoi des données ?

- Normalisation
- Comparaison
- Exhaustivité / lacunes
- Rapidité de lecture
- Facilité de mise à jour (données périssables)

Les listes structurées

- Des pages structurées comme des listes (ou des tables)
- ex. Liste des monuments historiques de l'Isère

Monument +	Commune \$	Adresse 4	Coordonnées	Notice +	Protection +	Date +	Illustration
Manoir de la Bâtie	Agnin		@ 45° 20′ 18″ Nord 4° 51′ 38″ Est	« РА00117114 » Ф	Inscrit	1979	Image manquante
Château de l'Alba	L'Albenc		45° 13′ 37″ Nord 5° 26′ 32″ Est	« PA00117115 » ₽	Inscrit	1978	Image manquante
<u>Château d'Anjou</u>	Anjou		<u>45° 21′ 00″ Nord</u> <u>4° 52′ 55″ Est</u>	« PA38000029 » ₽	Inscrit	2009	Image manquante
Éperon barré d'Annoisin- Chatelans	Annoisin- Chatelans	Le Devend La Larinaz Charbonnier	<u>45° 47′ 24″ Nord</u> <u>5° 18′ 00″ Est</u>	« PA00117116 » ₽	Inscrit	1983	Image manquante
Forges de Bonpertuis	Apprieu		45° 23′ 49″ Nord 5° 31′ 44″ Est	« PA38000013 » ₽	Inscrit	2003	2.5
<u>Château</u> <u>d'Amblérieu</u>	<u>La Balme-les-</u> <u>Grottes</u>		<u>45° 50′ 13″ Nord</u> <u>5° 19′ 40″ Est</u>	« PA00117117 » ₽	Inscrit	1977	Image manquante
<u>Château de</u> <u>Salette</u>	La Balme-les- Grottes		<u>45° 51′ 37″ Nord</u> <u>5° 19′ 27″ Est</u>	« PA38000001 » ₽	Inscrit	1996	Image manquante
Château du Fayet	Barraux		45° 25′ 03″ Nord 5° 58′ 20″ Est	« PA00117118 » @	Inscrit	1988	Image manquante
Fort Barraux	Barraux		45° 26′ 09″ Nord 5° 59′ 14″ Est	« PA00117119 »@	Classé	1990	A CONTRACTOR OF THE PARTY OF TH

Les infoboxes







Le précurseur **DBpedia**...

DBpedia (2007-) et le web sémantique (pour mémo)

- moissonner les données structurées de Wikipédia (dans chaque langue)
- construire et publier une base de données web requêtable "dataset"
- enrichir ou mettre à jour automatiquement le contenu de Wikipédia

Limites

- pas de mécanisme d'unification inter-langues
- pas de contrôle de validité (formel) des données
- pas d'appropriation par la communauté



Wikidata dans Wikipédia

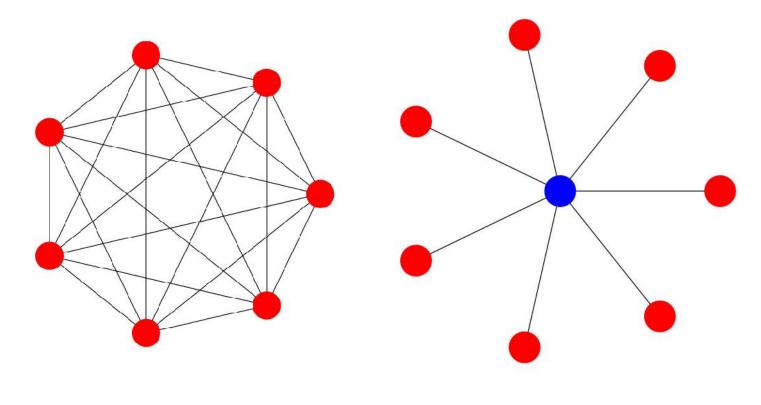


apprend le juken jutsu (combat à la baïonnette). Il participe à la guerre russo-japonaise comme caporal en Manchourie,

Bosanski

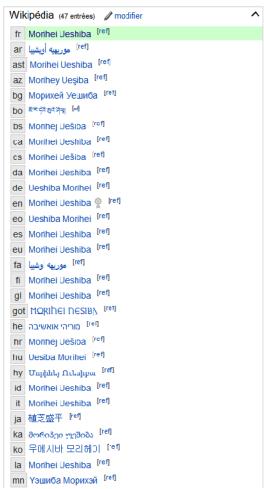
Omoto /

Gestion des interwikis



Avant Wikidata

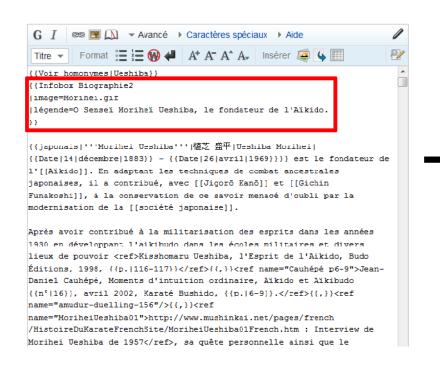
Avec Wikidata

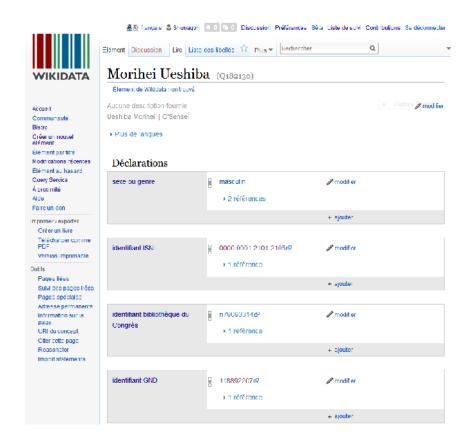


(élément Q182130)

Infobox avec Wikidata

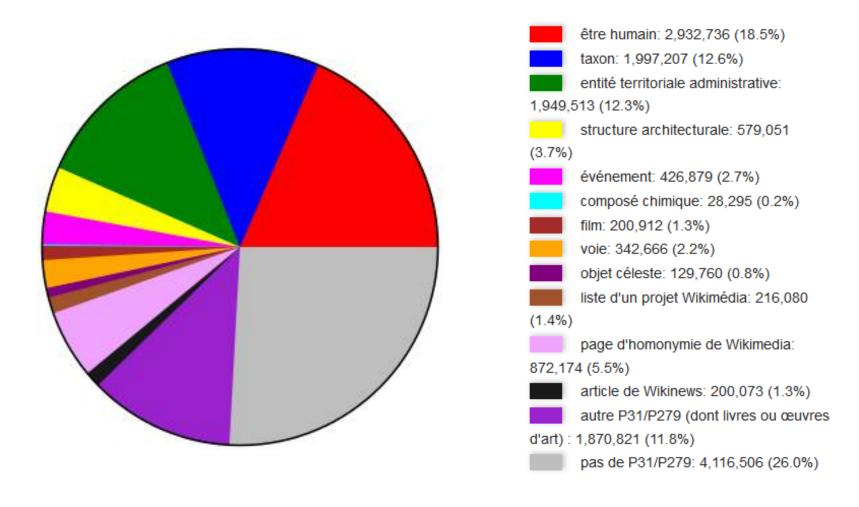








Domaines de la connaissance



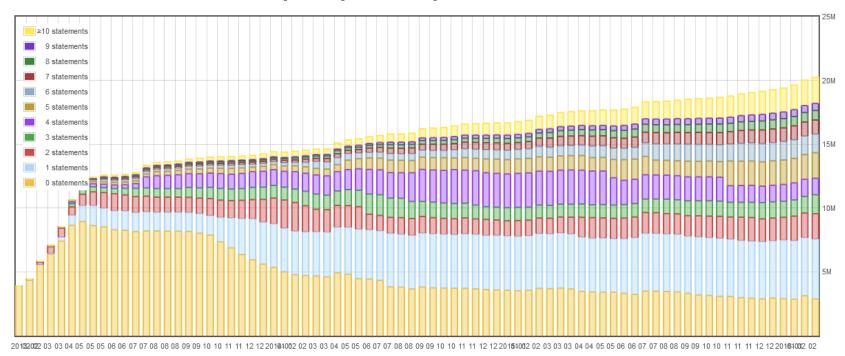
Élément Wikidata

- Un identifiant de la forme Qxxxxxxx
- Un libellé (label), une description et des alias par langue
- Des déclarations (statements) propriété/valeur
- Des liens vers des pages Wikimedia

[NB : élément = item]

Renseignement progressif

- Initialement les éléments sont vides
- Wikidata se remplit peu à peu





Q3605053

Évolution du nombre moyen de déclarations par élément (source : http://tools.wmflabs.org/wikidata-todo/stats.php)

Dénomination un élément



Déclaration (statement)

❖ La base : propriété / valeur



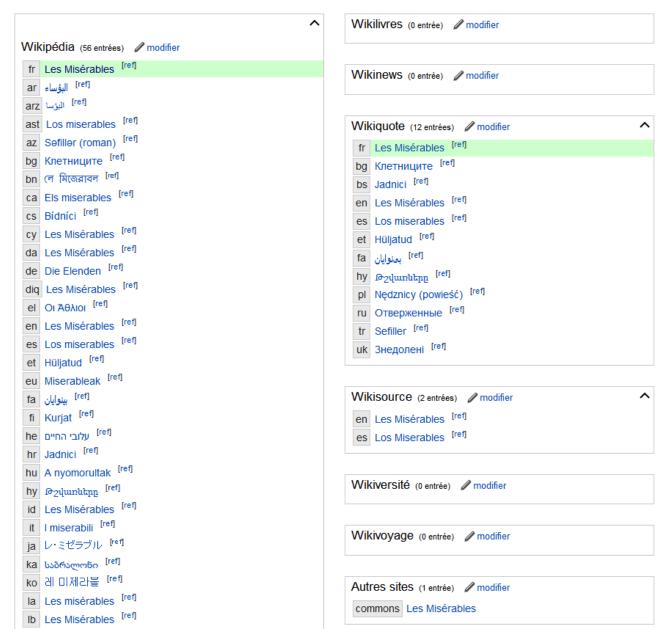
Les propriétés Wikidata

- ❖ Un identifiant de la forme Pxxxx
- ❖ Une page dédiée : https://www.wikidata.org/wiki/Property:P1639
- ❖ Un seul **type de valeur**, strict à l'édition
- *Des contraintes, souples à l'édition (voir la page de discussion de la propriété)
- ❖ Des propriétés qui peuvent évoluer à l'usage
- ❖ De nouvelles propriétés régulièrement créées (2637 aujourd'hui)

Types de données

- Élément Wikidata
- Chaîne de caractère
- ❖ Média commun (fichier sur <u>Wikimedia Commons</u>)
- ❖Date (proche ISO 8601)
- **❖**Coordonnées de géolocalisation
- ❖Quantité, avec unité et marge
- ❖Texte monolingue (chaîne de caractère avec langue)
- **∜**URL
- Identifiant externe
- Expression mathématique

Liens vers les projets Wikimedia



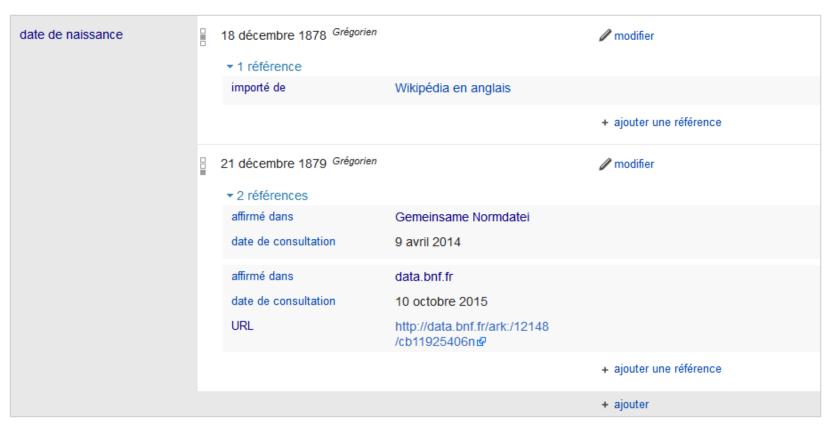
Des identifiants externes

Propriété +	Quantité d'éléments +		
identifiant GBIF (P846)	1807414		
identifiant Encyclopedia of Life (P830)	1370738		
identifiant Freebase (P646)	1157531		
identifiant GeoNames (P1566)	1028243		
identifiant VIAF (P214)	944113		
code de division administrative de la Chine (P442)	742245		
code postal (P281)	457181		
identifiant IMDb (P345)	405730		
identifiant GND (P227)	402571		
identifiant d'un gène Entrez (P351)	398324		
identifiant BnF (P268)	346821		
identifiant ITIS (P815)	336309		
identifiant The Plant List (P1070)	332560		
identifiant Tropicos (P960)	323686		
identifiant IdRef (P269)	320867		
identifiant bibliothèque du Congrès (P244)	303690		
identifiant NCBI (P685)	290366		
identifiant National Center for Biotechnology Information (P2393)	280300		
identifiant IPNI (P961)	278848		
identifiant ISNI (P213)	277273		
identifiant WoRMS (9850)	254245		

(source: https://www.wikidata.org/wiki/Wikidata:Database_reports/List_of_properties/Top100)

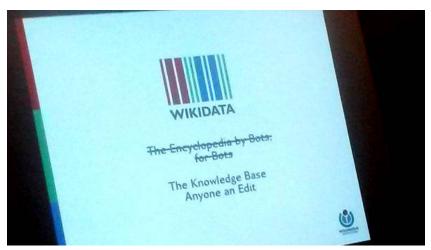
Une base de données secondaire

❖Wikidata ne présente pas la vérité mais des déclarations référençables



Une base de données que tout le monde peut modifier

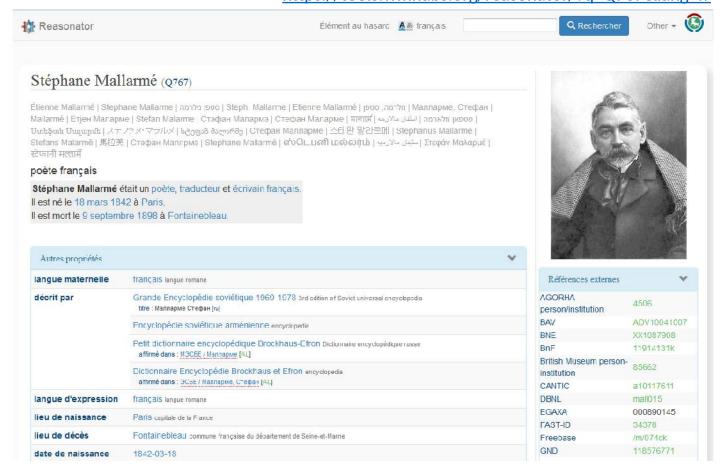
- Les éléments sont éditables
- Gadgets d'édition (Wikidata useful, ...)
- Outils d'édition semi-automatiques Quickstatements http://tools.wmflabs.org/wikidata-todo/quick statements.php Autolist http://tools.wmflabs.org/autolist/index.php
- Édition avec programmes informatiques (bots)
- Collaboratif et ouvert



Visualiser un élément avec Reasonator

Wikidata in pretty

https://tools.wmflabs.org/reasonator/?q=Q767&lang=fr

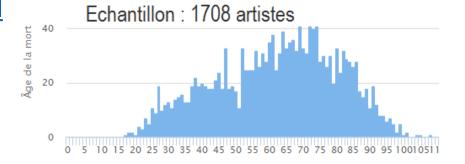


Réutiliser Wikidata



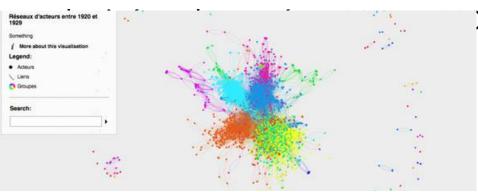
Les artistes ont-ils vraiment plus de risque de mourir à 27 ans ?

http://bit.ly/1pflghl



Modéliser le réseau social des act avec Wikidata

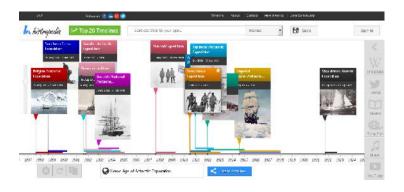
http://scoms.hypotheses.org/498



Réutiliser Wikidata

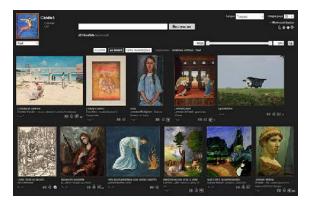
Histropedia

http://histropedia.com





http://zone47.com/crotos



Accéder aux données de Wikidata

- Les dumps
- ❖ API de MediaWiki
- WikidataQuery http://tools.wmflabs.org/autolist/index.php
- Endpoint SparQL https://query.wikidata.org/
 - ex. http://tinyurl.com/jq3mbuu : art public dans Paris
- Page dédiée https://www.wikidata.org/wiki/Wikidata:Data_access/fr

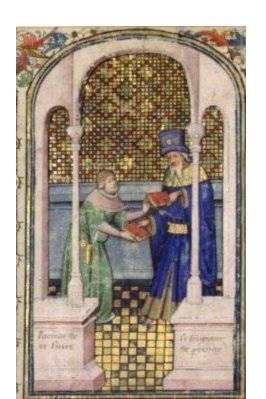
En conclusion, Wikidata est

- un pivot de liens entre les projets Wikimédia
- une base de données
- ❖... multilingue
- ❖ ... collaborative
- ❖ ... secondaire
- ❖ ... libre
- ❖ ... requêtable
- … en développement continu

Ressources

- ❖ Introduction à Wikidata, par Coyau, mis à jour par Ash Crow & Harmonia Amanda https://commons.wikimedia.org/wiki/File:Wikidata_pour_la_science.pdf
- A categorical imperative?, par Denny Vrandečić http://blog.wikimedia.de/2013/09/12/a-categorical-imperative/
- Wikidata: Knowledge from different points of view, par Lydia Pintscher https://en.wikipedia.org/wiki/Wikipedia:Wikipedia Signpost/2015-12-09/Op-ed
- Vidéo: Créer et renseigner un élément œuvre d'art sur Wikidata https://www.youtube.com/watch?v=-PiS-A3w3AM
- ❖ Voyage de DBpedia en Wikidata à bord d'un bot http://zone47.com/dozo/voyage-de-dbpedia-en-wikidata-a-bord-dun-bot
- ❖ Wikidata Enpoint SparQL et les peintures de Goya http://zone47.com/dozo/wikidata-enpoint-sparql-et-les-peintures-de-goya
- Mais combien y a-t-il de femmes scientifiques sur Wikipédia ?, par Ash Crow http://blog.ash.bzh/mais-combien-y-a-t-il-de-femmes-scientifiques-sur-wikipedia-this-time-with-sparql/
- Wikidata, SPARQL et chiens de traîneaux, par Ash Crow http://blog.ash.bzh/wikidata-sparql-et-chiens-de-traineaux/
- Wikidata, SPARQL et dynasties électives, par Ash Crow http://blog.ash.bzh/wikidata-sparql-et-dynasties-electives/

Merci



Document sous licence CC0 Transfert dans le domaine public



La personne qui a associé une œuvre à cet acte a dédié l'œuvre au domaine public en renonçant dans le monde entier à ses droits sur l'œuvre selon les lois sur le droit d'auteur, droit voisin et connexes, dans la mesure permise par la loi.

Vous pouvez copier, modifier, distribuer et représenter l'œuvre, même à des fins commerciales, sans avoir besoin de demander l'autorisation. Voir d'autres informations ci-dessous.

https://creativecommons.org/publicdomain/zero/1.0/deed.fr

Présentation d'origine (Wikidata) : Benoît Deshayes

Préambule Wikipédia : Guillaume Allègre

Wikidata en français (ou autre langue que l'anglais)

❖ Ajouter ?uselang=fr à l'URL

Exemple: https://www.wikidata.org/wiki/Q182130?uselang=fr

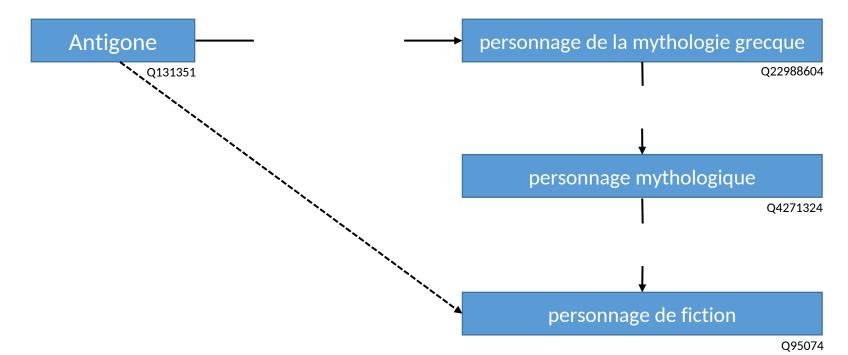
Avoir un compte et paramétrer sa ou ses langues dans sa page utilisateur {{#babel:fr-N}}

ou {{#babel:fr-N|en-3|de-1|en-1|it-1|es-0|nl-0|ru-0}}

exemple: https://www.wikidata.org/w/index.php?title=User:Shonagon&action=edit

P31 et P279 (Gné?!)

- Propriétés de base pour chaque élément (théoriquement exclusives)
- ❖ P31 : nature de l'élément (= est un, rdf:type, instance de)
- ❖ P279 : sous-classe (= est un type de, rdfs:subClassOf)



De data.bnf en Wikidata

❖Alignements

- À partir de l'<u>Endpoint SparQL de data.bnf</u> (<u>exemple</u>)
- Propriétés renseignées : BnF, mais aussi IdRef, ISNI, INSEE code commune

Bilan des identifiants ajoutés :

```
105 000 identifiants BnF (+70 000 via VIAF)
100 000 IdRef (+110 000 via VIAF)
```

*****Enrichissements

- Sourçage de déclarations existantes
- Ajout de nouvelles informations si lacune
- Propriétés (libellé, description, nature, dates de vie, genre, langue)

Bilan des enrichissements:

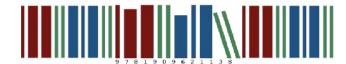
40 000 libellés et descriptions 400 000 références sur des données existantes 100 000 langues d'expression (propriété P1412)

De data.bnf en Wikidata

Des données déjà utilisables

- Images d'écrivaines du 18e siècle reliées au catalogue BnF : http://tinyurl.com/Ecrivaines18e
- Écrivains d'expression bretonne : http://tinyurl.com/EcrivainsBr

Les livres dans Wikidata



- Wikidata:WikiProject Books
- Nombre de livres référencés : <u>tinyurl.com/LivresCatWD</u>
- Livres disponibles : <u>tinyurl.com/LivresWD</u>