

Journées Unitex, 2021-12-16 (Champs-sur-Marne)
Slides available under open license 2.0: <https://tinyurl.com/UnitexWikidata>



Creation of linguistic resources with Wikidata

Philippe Gambette

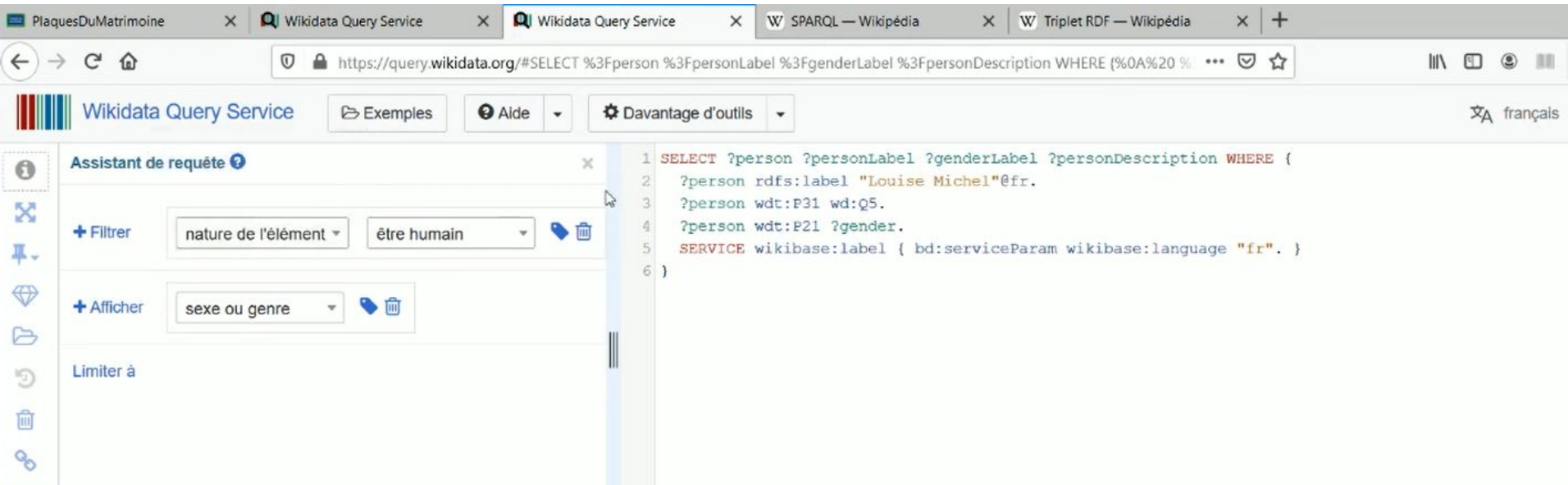


Outline

- Wikidata: a few video tutorials in French
- Querying Wikidata for entity linking: wikidataMultiSearch to make data FAIR
- Creating linguistic resources with Wikidata elements?
- Updating Unitex linguistic resources with Wikidata lexemes?

Wikidata: a few video tutorials in French

- general presentation of Wikidata, RDF triples and SPARQL queries
→ [Comment sont construites les requêtes Wikidata de l'outil PlaquesDuMatrimoine](#) (11'08)



The screenshot shows the Wikidata Query Service interface. The browser address bar displays the URL: `https://query.wikidata.org/#SELECT %3Fperson %3FpersonLabel %3FgenderLabel %3FpersonDescription WHERE (%0A%20 %...`. The interface includes a search bar, navigation buttons (Exemples, Aide, Davantage d'outils), and a language selector (français). The main area is divided into a left sidebar and a right main panel. The sidebar contains an "Assistant de requête" section with filters for "nature de l'élément" (être humain) and "sexes ou genres" (sexe ou genre). The main panel displays a SPARQL query:

```
1 SELECT ?person ?personLabel ?genderLabel ?personDescription WHERE (  
2   ?person rdfs:label "Louise Michel"@fr.  
3   ?person wdt:P31 wd:Q5.  
4   ?person wdt:P21 ?gender.  
5   SERVICE wikibase:label { bd:serviceParam wikibase:language "fr". }  
6 )
```

Wikidata: a few video tutorials in French

- creating a wikidata element after creating a new page on Wikipedia
→ [Créer une page Wikidata après avoir créé une nouvelle page Wikipédia](#) (16'47)



WIKIPÉDIA
L'encyclopédie libre

Rechercher dans Wikipédia

FreeCorp

[Accueil](#)
[Portails thématiques](#)
[Article au hasard](#)
[Contact](#)

[Basculer vers l'ancien habillage](#)

[Contribuer](#)
[Débuter sur Wikipédia](#)
[Aide](#)
[Communauté](#)
[Modifications récentes](#)
[Faire un don](#)

[Outils](#)
[Pages liées](#)
[Suivi des pages liées](#)
[Téléverser un fichier](#)
[Pages spéciales](#)
[Lien permanent](#)

Article [Discussion](#)

Lire [Modifier](#) [Modifier le code](#) [Voir l'historique](#) [Plus](#)

Lucie Quéméner

Lucie Quéméner, née à Paris en 1998, est une autrice française de bande dessinée.

Sommaire [\[masquer\]](#)

- [Biographie](#)
- [Œuvre](#)
- [Notes et références](#)
- [Lien externe](#)

Biographie [\[modifier\]](#) [\[modifier le code\]](#)

Lucie Quéméner est diplômée de l'Académie Brassart-Delcourt, promotion Arthur De Pins¹.

Son père est breton et sa mère d'origine chinoise².

Lucie Quéméner

Une illustration sous licence libre serait la bienvenue

Biographie

Naissance [1998](#)
[Paris](#)
Nationalité [Française](#)
Activité [Auteure de bande dessinée](#)

[modifier](#) - [modifier le code](#) - [modifier Wikidata](#)

Wikidata: a few video tutorials in French

- extracting names of people from an event programme (with [SEM](#)) and checking if they have a photo in Wikidata
→ [Comment trouver les personnes présentes à un événement qui n'ont pas encore de photo dans Wikidata ? \(13'29\)](#)



Accueil
Bienvenue
Communauté
Bistro
Aide

Participer
Importer un fichier
Modifications récentes
Nouveaux fichiers
Un fichier au hasard
Contactez-nous

Outils
Pages liées
Suivi des pages liées
Pages spéciales
Lien permanent
Informations sur la page

français FreeCorp Discussion Préférences Bêta Liste de suivi Imports Contributions Se déconnecter

Catégorie Discussion

Lire Modifier Voir l'historique Plus Rechercher dans Wikimedia Commons

Category:Live des Jeux at the Trocadero - August 8th 2021

Une page de Wikimedia Commons, la médiathèque libre.

Média dans la catégorie « Live des Jeux at the Trocadero - August 8th 2021 »

Cette catégorie comprend 78 fichiers, dont les 78 ci-dessous.



Wikidata: a few video tutorials in French

- creating a webpage in HTML+Javascript listing nearby tourist places
→ [Comment utiliser Wikidata pour lister sur une page web les lieux touristiques à proximité \(27'25\)](#)

← → ↻ 🔒 query.wikidata.org/#%23defaultView%3AlmageGrid%0A%23Éléments%20de%20type%20musée%2C%... ☆



commons:Château de Champs-sur-Marne 01.jpg
2.051
Point(2.60416667 48.85388889)
<https://fr.wikipedia.org/wiki/Ch%C3%A2teau_de_Champs-sur-Ma...
🔍 château de Champs-sur-Marne



commons:Noisiel (77) Ancien château 02.JPG
2.693
Point(2.6184 48.8521)
🔍 château de Noisiel





commons:Le Parc départemental de la Haute-île - panoramio - ...
2.852
Point(2.56115 48.8582)
<https://fr.wikipedia.org/wiki/Parc_d%C3%A9partemental_de_la_H...
🔍 parc départemental de la Haute-île



Wikidata: a few video tutorials in French

- importing a scanned book on Wikisource and creating the wikidata element about the edition
→ [*Importer un livre dans Wikisource, en utilisant Wikimedia Commons et Wikidata*](#) (18'01)



https://www.wikidata.org/wiki/Q109663975

titre  Légendes du vieux Béziers (français)  modifier

- 0 référence

+ ajouter une référence




+ ajouter une valeur

auteur  Mathilde Bellaud-Dessalles  modifier

- 0 référence

+ ajouter une référence

+ ajouter une valeur

fac-similé    modifier

Bellaud-Dessalles - Légendes du vieux Béziers, 1923.pdf
1 237 × 1 752, 158 pages, 63,89 Mio

- 0 référence

+ ajouter une référence

+ ajouter une valeur

+ ajouter une déclaration

Querying Wikidata for entity linking

- wikidataMultiSearch: [demo](#)
- use cases (ongoing projects):
 - entity linking for places, people, pollutants, etc.:
 - <https://github.com/e-ditiones/Annotator> (TEI 2021)
 - Chuanming Dong's PhD project: [industrial pollution events](#)
 - mapping locations:
 - pictures of cities in the 16th and 17th centuries on [uMap](#)
 - [with a text already labeled in Wikisource](#) by project [Lessico Beni culturali](#)
 - making data FAIR (I = interoperable):
 - [CIDRE corpus](#) with Olga Seminck ([JOHD, 2021](#))

Creating linguistic resources with Wikidata elements

- using the property “instance of element”
 - 83 elements of instance “bone”: <https://w.wiki/4ZHN>
 - elements of instance “bone” (or subclasses) with their instance: <https://w.wiki/4ZHL>
- using identifiers in other databases:
 - [1032 women](#) with an [identifier in the *Dictionnaire des femmes de l'ancienne France*](#) by SIEFAR
 - > 100 000 [elements](#) with a [DSSTOX identifier](#) (Environment Protection Agency)

Updating Unisex list of first names?

Using the property “instance of element” for first names: [20622 results](#)

```
SELECT ?firstName ?firstNameLabel WHERE {  
  SERVICE wikibase:label { bd:serviceParam  
    wikibase:language "[AUTO_LANGUAGE],en,fr,es". }  
  ?firstName wdt:P31 wd:Q202444.  
}  
LIMIT 100
```

With gender: [male](#), [female](#), [unisex](#)

Wikidata lexemes

Développement liés aux [données lexicographiques dans Wikidata](#) depuis 2016

[FAQ Wikidata/Wiktionnaire](#)

[Proposition de propriétés lexicographiques](#)



- Accueil
- Communauté
- Bistro
- Créer un nouvel élément
- Modifications récentes
- Élément au hasard
- Query Service
- À proximité
- Aide
- Faire un don

- Données lexicographiques
- Créer un nouveau lexème
- Modifications récentes
- Lexème au hasard

- Outils
- Pages liées
- Suivi des pages liées
- Pages spéciales
- Lien permanent
- Informations sur la page
- URI du concept
- Citer cette page

Lexeme Discussion

(L579430) **poutine** modifier

fr

Langue français
Nature nom

Déclarations

genre grammatical	féminin	modifier
	0 référence	+ ajouter une référence
		+ ajouter une valeur
		+ ajouter une déclaration

Sens

L579430-S1 français mets de restauration rapide québécois fait de pommes de terre frites, de fromage en grains et d'une sauce chaude modifier

Déclarations concernant L579430-S1

décrit par	Le grand dictionnaire terminologique	modifier
	1 référence	+ ajouter une valeur

élément pour ce sens	poutine	modifier
	0 référence	+ ajouter une référence
		+ ajouter une valeur

Updating Unitex linguistic resources?

- Linguistic resources in French:

<https://github.com/UnitexGramLab/unitex-lingua/tree/master/fr>

- Uncompressing compressed linguistic resources (.bin files)

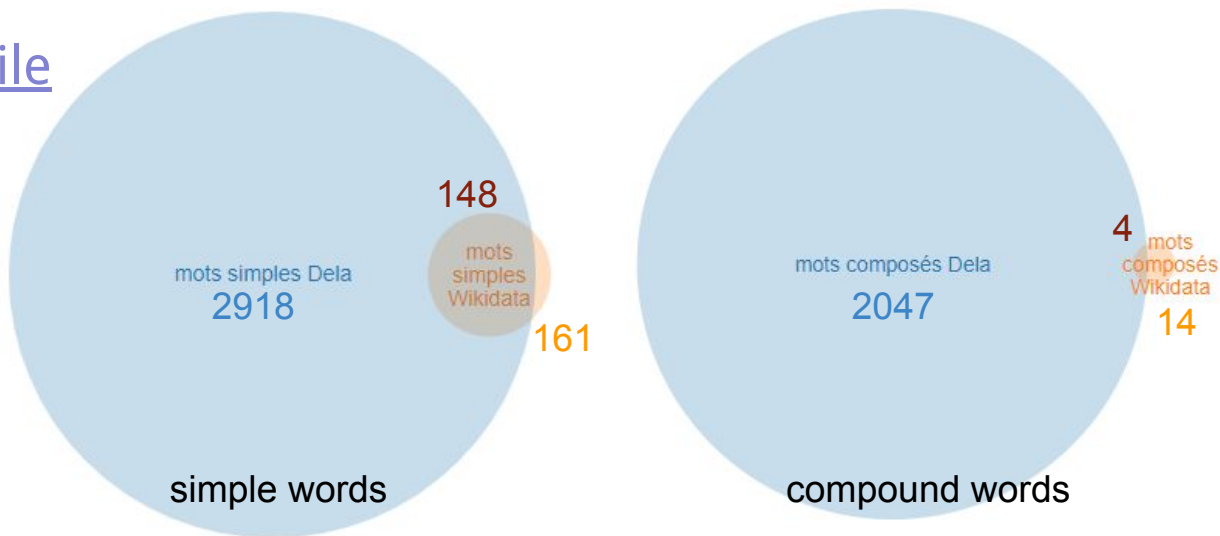
```
C:\Users\...\AppData\Local\Unitex-GramLab\App\UnitexToolLogger.exe  
Uncompress C:\Users\...\AppData\Local\Unitex-GramLab\French\Dela\  
Dela_fr.bin -o dela.txt
```

```
(base) C:\Users\...\AppData\Local\Unitex-GramLab\App>UnitexToolLogger.exe Uncompress  
..\French\Dela\Dela_fr.bin -o dela.txt  
Uncompressing ..\French\Dela\Dela_fr.bin...  
Done.
```

Updating Unitex linguistic resources?

An example, comparing lemmas starting with “po”:

- [Wikidata query](#) derived from the example “[Lexemes in English that match an expression](#)” ([Lexèmes en anglais correspondant à une expression](#))
- [Google Sheets data file](#)



Updating Unitex linguistic resources?

An example, comparing lemmas starting with “po”

Simple words in Wikidata (lexemes) but not in Dela :

podcast

pointille

polychèteologue

polymédication

poméranche

portecle

portecrayon

portevoix

postgrade

posttraitement

potpourri

poussepousse

poutine

Updating Uniflex linguistic resources?

An example, comparing lemmas starting with “po”

Compound words in Wikidata (lexemes) but not in Dela :

point de basculement dans le système climatique

politique climatique

pollueur-payeur

porte-assiettes

porte-bouquet

porte-parapluie

post mortem

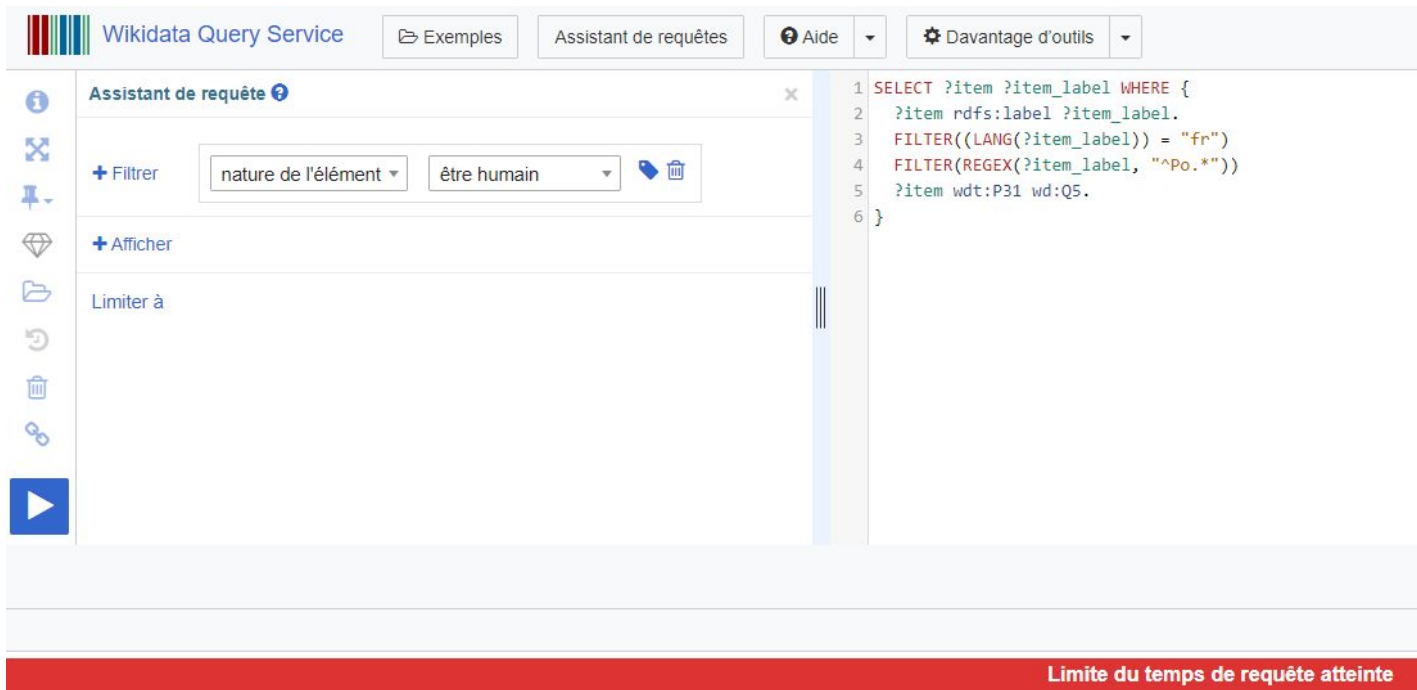
pot-au-feu

potentiel de réchauffement global

pourvu que

Updating Unitex linguistic resources?

Proper nouns?



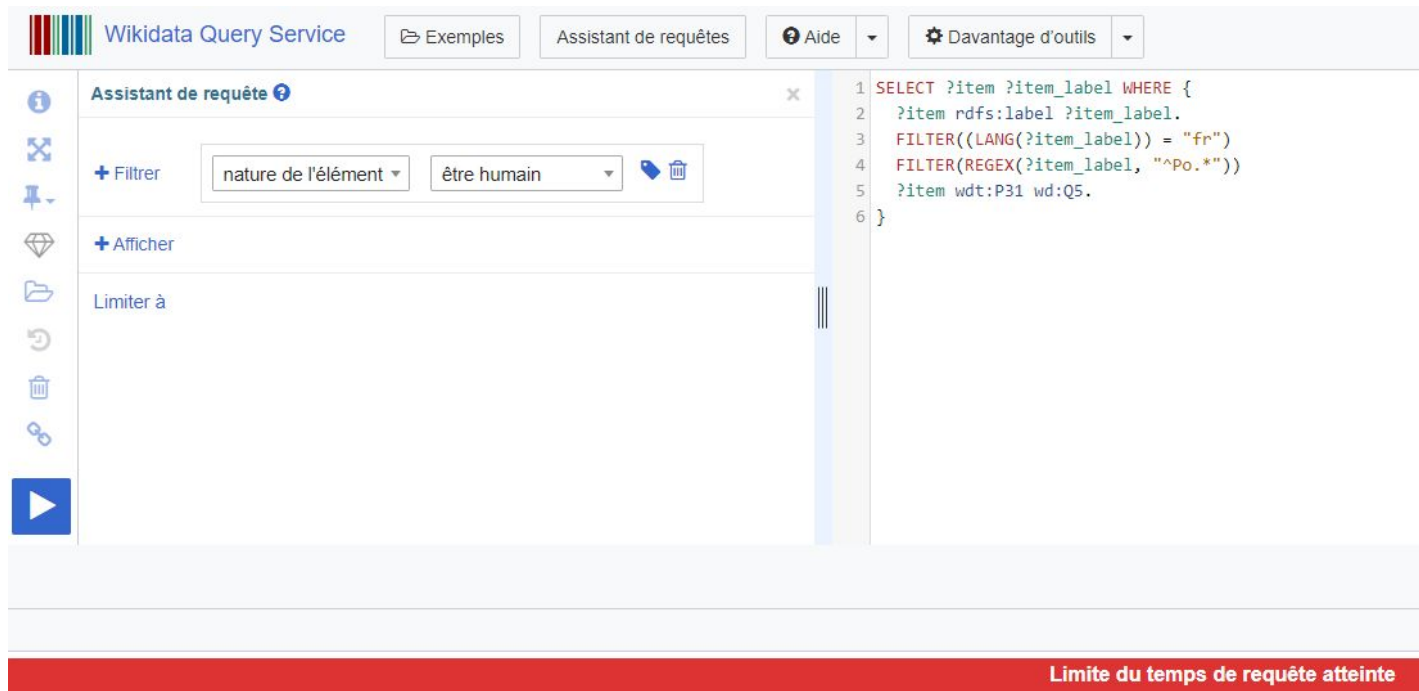
The screenshot displays the Wikidata Query Service interface. At the top, there is a navigation bar with the Wikidata logo, the text "Wikidata Query Service", and buttons for "Exemples", "Assistant de requêtes", "Aide", and "Davantage d'outils". Below this, the "Assistant de requête" panel is visible, featuring a "Filtrer" section with two dropdown menus: "nature de l'élément" and "être humain". To the right of the assistant is a code editor containing a SPARQL query:

```
1 SELECT ?item ?item_label WHERE {  
2   ?item rdfs:label ?item_label.  
3   FILTER((LANG(?item_label)) = "fr")  
4   FILTER(REGEX(?item_label, "^Po.*"))  
5   ?item wdt:P31 wd:Q5.  
6 }
```

At the bottom of the interface, a red banner displays the message "Limite du temps de requête atteinte".

Updating Unitex linguistic resources?

Proper nouns?



The screenshot shows the Wikidata Query Service interface. At the top, there are navigation buttons: "Exemples", "Assistant de requêtes", "Aide", and "Davantage d'outils". The main area is titled "Assistant de requête" and contains a filter section with two dropdown menus: "nature de l'élément" and "être humain". Below the filter is an "Afficher" section and a "Limiter à" section. On the right side, a SPARQL query is displayed in a code editor:

```
1 SELECT ?item ?item_label WHERE {
2   ?item rdfs:label ?item_label.
3   FILTER((LANG(?item_label)) = "fr")
4   FILTER(REGEX(?item_label, "^Po.*"))
5   ?item wdt:P31 wd:Q5.
6 }
```

At the bottom of the interface, a red banner displays the message: "Limite du temps de requête atteinte".

→ Wikidata [dumps](https://wikidata.org/wiki/Wikidata:Database_download/fr): wikidata.org/wiki/Wikidata:Database_download/fr