

# Wikidata et le web sémantique

---

Mattia Bunel

Jeudi 19 décembre 2024

Objectif de cette présentation:

- Présenter le projet **Wikidata**
- Présenter les notions et technologies du web sémantique
- Présenter le langage de requête **SPARQL**

# Le web sémantique

---

## Qu'est-ce que le web sémantique ?

*« Je rêve d'un Web [dans lequel les ordinateurs] deviennent capables d'analyser toutes les données du Web : le contenu, les liens, et les transactions entre personnes et ordinateurs. Un « Web Sémantique », qui devrait rendre cela possible, n'a pas encore émergé, mais quand il le fera, les mécanismes journaliers du commerce, de l'administration et de nos vies quotidiennes seront traités par des machines dialoguant avec d'autres machines. Les « agents intelligents » qu'on nous vante depuis longtemps se concrétiseraient enfin. »*

*— Tim Berners-Lee, Weaving the Web (1999).*

Chapô de la page Wikipédia de Paris :

*Paris (/pa.ʁi/a Écouter) est la capitale de la France et une collectivité à statut particulier. Divisée en vingt arrondissements, elle est le chef-lieu de la région Île-de-France et le siège de la métropole du Grand Paris.*

## Transformation en liste de faits

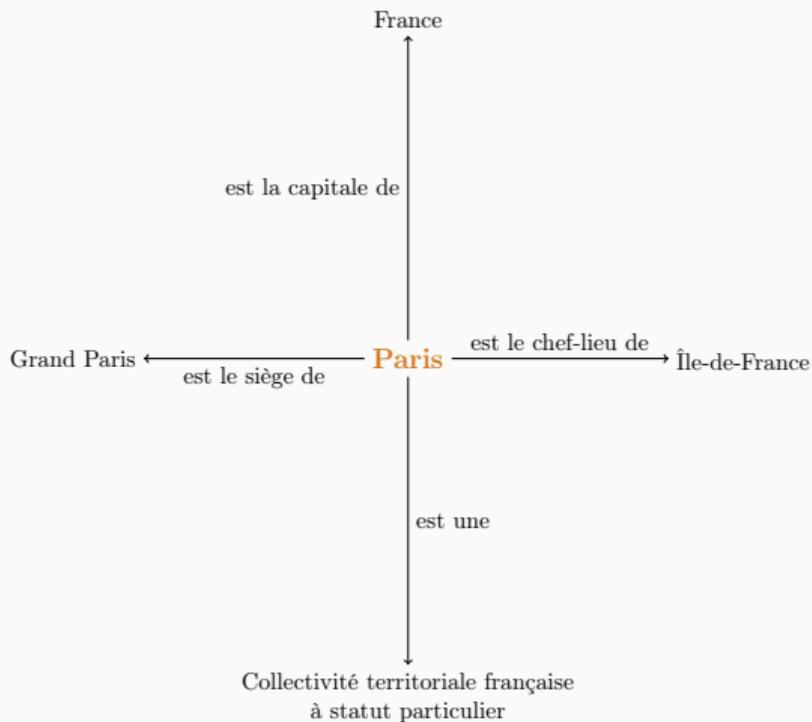
On peut représenter cette information sous la forme d'une **liste de faits** :

- Paris est la capitale de la France
- Paris est une collectivité à statut particulier
- Paris est le chef-lieu de la région Île-de-France
- Paris est le siège de la métropole du Grand Paris.

Tous ces faits ont une structure en triplet:

- **Sujet** : *Paris*
- **Prédicat** : *est la capitale de, est le siège de*
- **Objet** : *La France, métropole du Grand Paris*

# Représentation sous forme de graphe



```
<!-- Sous ensemble du RDF -->
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <rdf:Description rdf:about="http://-/Q90">
    <wdt:P3417>Paris</wdt:P3417>
    <wdt:P361 rdf:resource="http://-/Q13917"/>
    <wdt:P31 rdf:resource="http://-/Q22923920"/>
    <wdt:P1376 rdf:resource="http://-/Q142"/>
    <wdt:P1376 rdf:resource="http://-/Q16665915"/>
    <wdt:P443 rdf:resource="http://-/Paris.wav"/>
  </rdf:Description>
</rdf:RDF>
```

## Le projet Wikidata

---

# Présentation de Wikidata

Wikidata est:

- un projet de la fondation Wikimedia
- une base de connaissances libre et collaborative
- compatible avec les technologies du web sémantique (*i.e.* RDF et SPARQL)
- unique (pas d'instance / langue)

Aujourd'hui Wikidata est composée de :

- 114 711 518 éléments, de *TYC 4404-1158-1* (Q89383366) à la *guerre des émeus* (Q14665)
- 12 302 propriétés (cf. liste complète)

# À quoi est-ce que ça ressemble ?

English Not logged in Talk Contributions Create account Log in

Item Discussion Read View History Search Wikidata

## Paris (Q90)

capital city of France  
City of Light | City of Love | Lutetia | Paris, Île-de-France | Paris, France

[+ In more languages](#)

### Statements

instance of	commune of France
end time	31 December 2018
<a href="#">+ 1 reference</a>	
department of France	
foundational text	loi du 10 juillet 1964 portant réorganisation de la région parisienne
follows	Seine
start time	1 January 1968
end time	31 December 2018
criterion used	territorial collectivity of France
<a href="#">+ 1 reference</a>	
territorial collectivity of France with special status	
start time	1 January 2019
subject has role	commune of France department of France
<a href="#">+ 1 reference</a>	
metropolis	
of	Paris metropolitan area
<a href="#">+ 0 references</a>	
tourist destination	
<a href="#">+ 0 references</a>	
global city	
<a href="#">+ 0 references</a>	
megacity	
<a href="#">+ 0 references</a>	
largest city	
of	France

## Quel intérêt pour la recherche ?

Le projet [Wikidata](#) permet:

- de récupérer « simplement » un grand volume de données (semi) structurées.
- d'interroger le contenu de Wikipédia.
- de contribuer d'une manière différente (valorisation de résultats).
- de se former aux technologies du web sémantique

Extraire l'information

---

- L'information n'est généralement pas stockée directement en **XML**, mais dans des bases de donnée spécialisées (*triplestore*).
- Les données sont exposées par le biais d'un *endpoint*
- On utilise un langage de requête spécifique, le **SPARQL** pour extraire l'information.
- Parfois, on peut également accéder à un dump **RDF**. Il faut alors utiliser son propre *triplestore* (p. ex. **rdflib** en python ou **virtuoso**).

## Quelques *endpoints*

Organisation	URL
--------------	-----

---

Wikidata	<a href="https://query.wikidata.org/">https://query.wikidata.org/</a>
INSEE	<a href="https://rdf.insee.fr/sparql/">https://rdf.insee.fr/sparql/</a>
BNF	<a href="https://data.bnf.fr/sparql/">https://data.bnf.fr/sparql/</a>
IGN	<a href="https://data.ign.fr/endpoint.html">https://data.ign.fr/endpoint.html</a>
HAL	<a href="https://data.hal.science/doc/sparql">https://data.hal.science/doc/sparql</a>
Persée	<a href="https://data.persee.fr/explorer/sparql-endpoint/">https://data.persee.fr/explorer/sparql-endpoint/</a>

# Exemple de *endpoint* wikidata

Wikidata Query Service

```
1 #Evénements récents
2 #title: Evénements récents
3 SELECT ?event ?eventLabel ?date
4 WITH {
5   SELECT DISTINCT ?event ?date
6   WHERE {
7     # find events
8     ?event wdt:P31wdt:P379* wd:Q1194554.
9     # with a point in time or start date
10    OPTIONAL { ?event wdt:P580 ?date. }
11    OPTIONAL { ?event wdt:P585 ?date. }
12    # but at least one of those
13    FILTER(BOUND(?date) && DATATYPE(?date) = xsd:dateTime).
14    # not in the future, and not more than 31 days ago
15    BIND(NOW() - ?date AS ?distance).
16    FILTER(!= ?distance && ?distance < 31).
17  }
18 LIMIT 150
19 } AS %1
20 WHERE {
21   INCLUDE %1
22   SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],mul,en" . }
23 }
```

150 résultats en 7 ms

Evénements récents		
event	eventLabel	date
<a href="#">Q131303105</a>	30th Forgal Awards	14 décembre 2024
<a href="#">Q130756028</a>	17th Asia Pacific Screen Awards	30 novembre 2024
<a href="#">Q130084982</a>	61e célébration des Golden Horse Film Festival and Awards	23 novembre 2024
<a href="#">Q121254115</a>	37e cérémonie des prix du cinéma européen	7 décembre 2024
<a href="#">Q125691384</a>	The Genre Awards 2024	12 décembre 2024
<a href="#">Q131306691</a>	accord de cessez-le-feu de 2024 entre Israël et le Liban	27 novembre 2024
<a href="#">Q131319470</a>	DHL Flight 180	25 novembre 2024
<a href="#">Q131362981</a>	Farewell meeting of the Uruguayan community for 2024	29 novembre 2024
<a href="#">Q131457095</a>	Séisme de 2024 à Port-Vila	17 décembre 2024
<a href="#">Q131393581</a>	2024 Cape Mendocino earthquake	6 décembre 2024
<a href="#">Q131449870</a>	2024 Kerch Strait oil spill	15 décembre 2024
<a href="#">Q131457000</a>	Q131457000	11 décembre 2024
<a href="#">Q131393984</a>	Wikiconary workshop for WikiKata competition 2024	11 décembre 2024
<a href="#">Q131393997</a>	Q131393997	6 décembre 2024
<a href="#">Q131393984</a>	Wikiconary workshop for WikiKata competition 2024	6 décembre 2024
<a href="#">Q131393992</a>	Q131393992	6 décembre 2024
<a href="#">Q131393984</a>	Wikiconary workshop for WikiKata competition 2024	6 décembre 2024
<a href="#">Q131392544</a>	How to use interactive maps in Wikipedia articles?	26 novembre 2024

# Constructeur de requetes

## WIKIDATA QUERY BUILDER

français

**À propos de cet outil**

Le Constructeur de requêtes de Wikidata fournit une interface visuelle pour construire une requête Wikidata simple. Il est idéal pour les utilisateurs avec peu ou pas d'expérience en SPARQL, le puissant langage de requêtage. Le Constructeur de requêtes n'offre pas toutes les fonctionnalités de SPARQL, mais vous pouvez toujours ouvrir votre requête dans le Service de requête, où vous pouvez afficher, la modifier ou l'étendre via le lien au-dessus des résultats. [Les avis sont bienvenus ici.](#)

**Requête**

Trouver tous les éléments...

**Avec** **Sans** Propriété  correspondant à Valeur  Références

Inclure également les valeurs liées lors de la recherche (recommandé)

**Ajouter une condition**

**Paramètres**

Limiter le nombre de résultats à

Montrer les identifiants plutôt que les libellés (peut éviter une erreur d'exécution hors détail).

**Exécuter la requête** Obtenir un lien partageable

[Afficher la requête dans le service de requête](#)

**Résultats**

Item	ItemLabel
<a href="#">wd:Q33321</a>	Vichy
<a href="#">wd:Q30</a>	Paris
<a href="#">wd:Q321</a>	Versailles
<a href="#">wd:Q1479</a>	Bordeaux
<a href="#">wd:Q3561</a>	Alger

- Le langage **SPARQL** permet d'interagir avec des données **RDF**.
- Il permet notamment :
  - D'extraire des informations (**SELECT**).
  - De modifier ou d'ajouter des éléments (**CONSTRUCT**).
- Malgré leur (relative) ressemblance, **SPARQL** et **SQL** sont très différents.
- Langage assez simple, mais difficile à appréhender dans un premier temps.

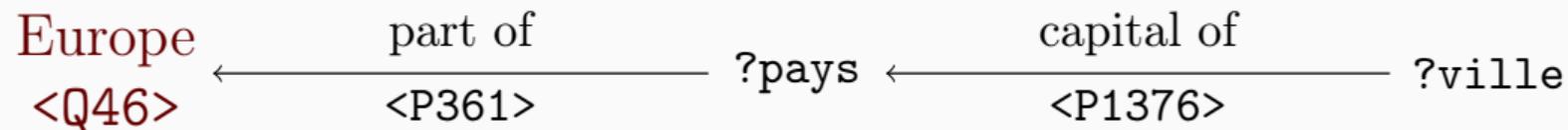
## Exemple de requête SPARQL (1)

Comment identifier les capitales européennes dans ce graphe ?



## Exemple de requête SPARQL (2)

On recherche le motif suivant :



Avec `?pays` et `?ville` des variables.

## Exemple de requête SPARQL (3)

En **SPARQL** ce motif s'écrit de la manière suivante :

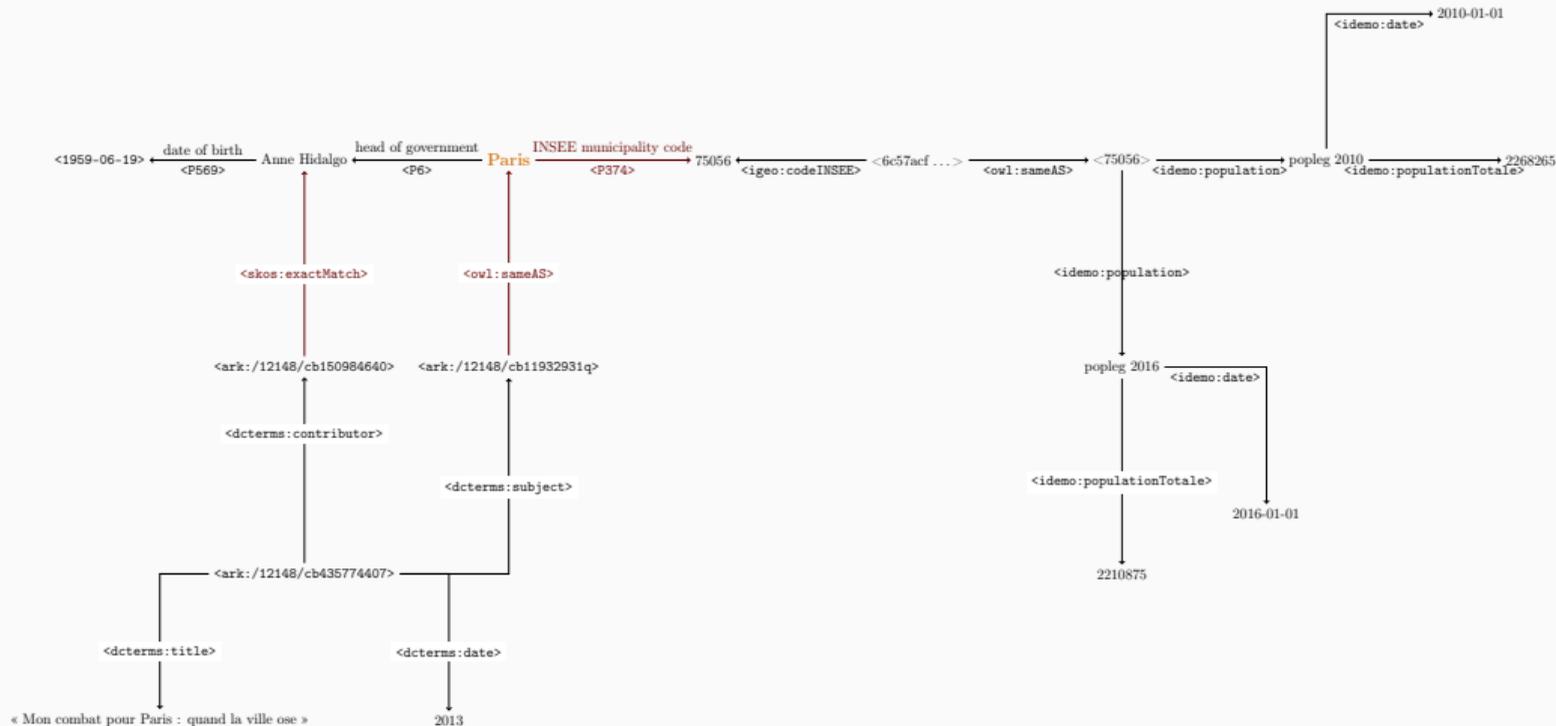
```
SELECT *  
WHERE {  
  ?ville P1376 ?pays.  
  ?pays P361 Q46.  
}
```

## Croiser l'information

---

- **Wikidata** n'est pas le seul endpoint.
- Un même objet peut être décrit à plusieurs endroits. Exemple *Paris*:
  - Sur wikidata : Q90
  - Sur data.bnf : `ark:/12148/cb11932931q`
  - sur rdf.insee.fr : 75056
  - sur data.ign.fr : 75056
- Si l'on *sait* que ces identifiants décrivent le même objet, alors **on peut croiser ces sources de données**.

# Un exemple de graphe de connaissance fédéré



## Requête fédérée

un exemple de requête SPARQL fédérée (Wikidata / INSEE) :

```
SELECT *
WHERE {
  SERVICE <https://query.wikidata.org/sparql> {
    wd:Q90 wdt:P374 ?codeinsee.
  }
  SERVICE <https://rdf.insee.fr/sparql> {
    ?s_insee igeo:codeINSEE ?codeinsee;
             a          igeo:Commune.
  }
}
```

## Dans les faits (1)

- Réservé à des usages avancés
- Nécessite de connaître la structuration des données pour chaque source
- Pour faire une requête fédérée il faut que l'endpoint l'autorise.
  - L'endpoint de wikidata autorise les requêtes fédérées avec une petite centaine d'endpoints. Notamment :
    - `data.europa.eu`
    - `data.idref.fr`
    - `data.bnf.fr`
    - `rdf.insee.fr`
  - Beaucoup d'endpoints ne sont accessibles par wikidata (ex. `data.ign.fr`)

## Dans les faits (2)

- Les requêtes fédérées ne sont pas implémentées dans beaucoup de clients (p. ex. `rdflib`).
- La solution la plus « simple », déployer localement un *triplestore* et autoriser les requêtes fédérées :

# Déploiement de virtuoso CE avec docker

```
docker run \  
  --name my_virtddb --interactive --tty \  
  --env DBA_PASSWORD=mysecret \  
  --publish 1111:1111 --publish 8890:8890 \  
  --volume `pwd`::/database \  
  openlink/virtuoso-opensource-7:latest
```

## Pour conclure

- Le projet **Wikidata** offre de nombreuses possibilités (cf. exemples du *endpoint*).
- Les premiers pas sont difficiles :
  - Technologie austère
  - Petite communauté
  - Peu d'implémentations (beaucoup de **Java**)
- De nombreux points non abordés:
  - Comment définir des propriétés → **RDFS** et **OWL**
  - Construction d'ontologies → **OWL**
  - Comment s'assurer de la cohérence → Raisonnement formel
  - Comment inférer des connaissances → Raisonnement formel
- Quid des **LLM** ?

## Pour aller plus loin !

- Formations de *Pierre-Yves Beaudouin* sur Wikidata:
  - « Introduction à wikidata ». 2023.  
<https://doi.org/10.5281/zenodo.8032969>.
  - « Introduction à SPARQL ». 2024.  
<https://doi.org/10.5281/zenodo.13986391>.
- Cours en ligne:
  - MOOC « Web sémantique et Web de données ». <https://www.fun-mooc.fr/fr/cours/web-semantique-et-web-de-donnees/>.
- Spécifications du W3C :
  - « RDF 1.1 Primer ». 2014. <https://www.w3.org/TR/rdf11-primer/>.
  - « SPARQL 1.1 Overview ». 2013.  
<https://www.w3.org/TR/sparql11-overview/>.
  - « Publications of the W3C Semantic Web Activity ». 2013.  
<https://www.w3.org/2001/sw/Specs.html>.

Merci de votre attention