ENSG Géomatique

ÉCOLE NATIONALE DES SCIENCES GÉOGRAPHIQUES

GÉOGRAPHIE-CITÉS UMR 8504

CONSTRUCTION D'UN EXTRACTEUR D'INFORMATION GÉOGRAPHIQUE AVEC R ET SPARQL

Juliette Delannoy
Projet encadré par Hadrien Commenges et Thomas Louail
Laboratoire de recherche Géographie-cités

Projet de développement informatique 2018

Introduction

Années 2000 : début de l'utilisation du web sémantique

⇨ Structuration de la donnée

Interrogation automatique possible

Exemple : Dbpédia = base de données de Wikipédia

⇨ Interrogation avec mon interface

Plan

- Web sémantique
- Dbpédia
- Langage SPARQL
- Interface
  - Fonctionnalités de base
  - Carte pour résultats spatialisés
  - Informations complémentaires
- Conclusion

Web sémantique

= Web de la donnée
= Web 3.0

“The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries”(W3C)

Fournir un **cadre** pour permettre le **partage** et la **réutilisation des données.**

□ Structurer l’information uniformément pour l’universalité de l’accès et du partage

Web sémantique

= Web de la donnée (brute)
= Web 3.0

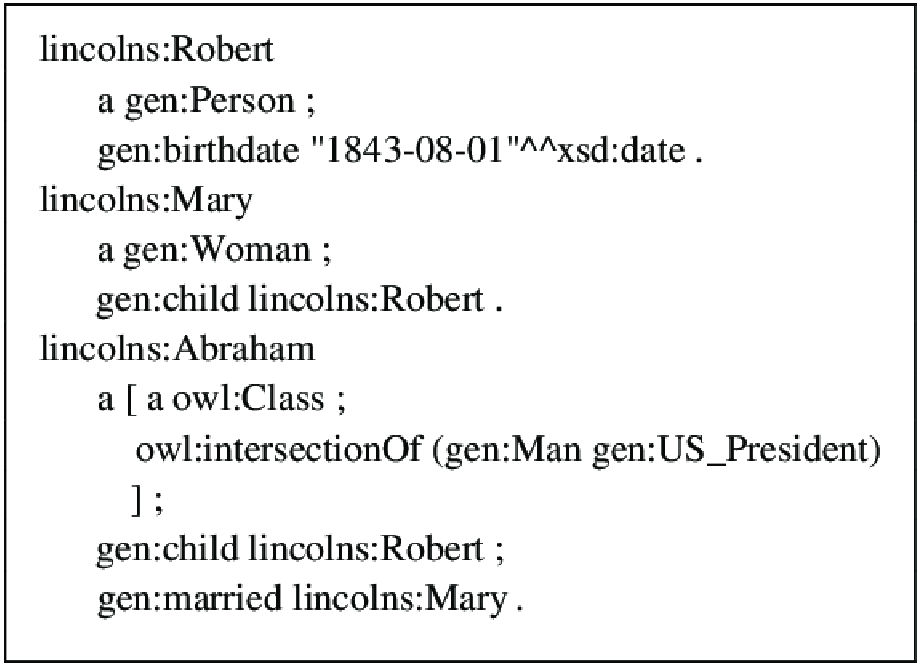
□ Language universel et uniforme :

RDF (Resource Description Framework)

(Cadre de description des ressources)

Sujet → Prédicat → Objet

Le chien s'appelle Bary.

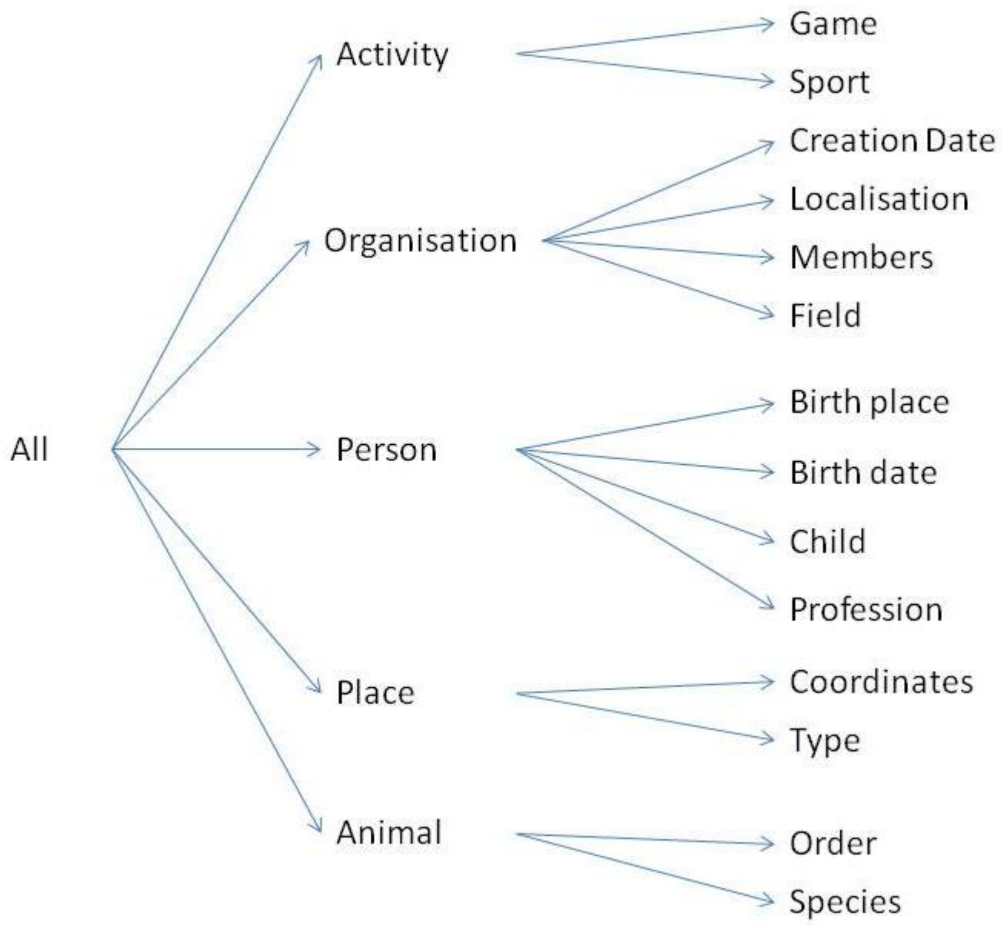


```
lincolns:Robert
     a gen:Person ;
     gen:birthdate "1843-08-01"^^xsd:date .
lincolns:Mary
     a gen:Woman ;
     gen:child lincolns:Robert .
lincolns:Abraham
     a [ a owl:Class ;
        owl:intersectionOf (gen:Man gen:US_President)
       ] ;
     gen:child lincolns:Robert ;
     gen:married lincolns:Mary .
```

Web sémantique

= Web de la donnée (brute)
= Web 3.0

- Ontologies



Web sémantique

But : Pouvoir interroger tout le web automatiquement

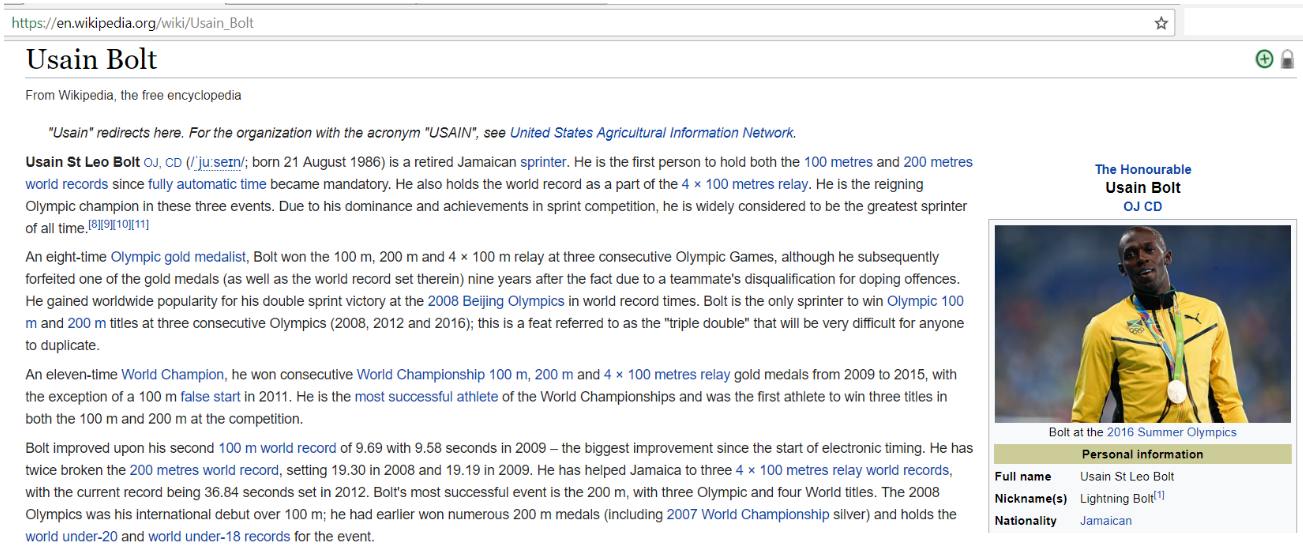
Web scraping inutile

Réalité : Structuration partielle

Ex : Comparateur de billets d’avion

Dbpédia : Base de données structurée

Dbpédia et Wikipédia



https://en.wikipedia.org/wiki/Usain_Bolt

Usain Bolt

From Wikipedia, the free encyclopedia

"Usain" redirects here. For the organization with the acronym "USAIN", see United States Agricultural Information Network.

Usain St Leo Bolt OJ, CD (/ˈjuːseɪn/; born 21 August 1986) is a retired Jamaican sprinter. He is the first person to hold both the 100 metres and 200 metres world records since fully automatic time became mandatory. He also holds the world record as a part of the 4 × 100 metres relay. He is the reigning Olympic champion in these three events. Due to his dominance and achievements in sprint competition, he is widely considered to be the greatest sprinter of all time.[8][9][10][11]

An eight-time Olympic gold medalist, Bolt won the 100 m, 200 m and 4 × 100 m relay at three consecutive Olympic Games, although he subsequently forfeited one of the gold medals (as well as the world record set therein) nine years after the fact due to a teammate's disqualification for doping offences. He gained worldwide popularity for his double sprint victory at the 2008 Beijing Olympics in world record times. Bolt is the only sprinter to win Olympic 100 m and 200 m titles at three consecutive Olympics (2008, 2012 and 2016); this is a feat referred to as the "triple double" that will be very difficult for anyone to duplicate.

An eleven-time World Champion, he won consecutive World Championship 100 m, 200 m and 4 × 100 metres relay gold medals from 2009 to 2015, with the exception of a 100 m false start in 2011. He is the most successful athlete of the World Championships and was the first athlete to win three titles in both the 100 m and 200 m at the competition.

Bolt improved upon his second 100 m world record of 9.69 with 9.58 seconds in 2009 – the biggest improvement since the start of electronic timing. He has twice broken the 200 metres world record, setting 19.30 in 2008 and 19.19 in 2009. He has helped Jamaica to three 4 × 100 metres relay world records, with the current record being 36.84 seconds set in 2012. Bolt's most successful event is the 200 m, with three Olympic and four World titles. The 2008 Olympics was his international debut over 100 m; he had earlier won numerous 200 m medals (including 2007 World Championship silver) and holds the world under-20 and world under-18 records for the event.

His achievements as a sprinter have earned him the media nickname "Lightning Bolt", and his awards include the IAAF W
Field Athlete of the Year, and Laureus World Sportsman of the Year (four times). Bolt retired after the 2017 World Champ
his last solo 100m race.[12]

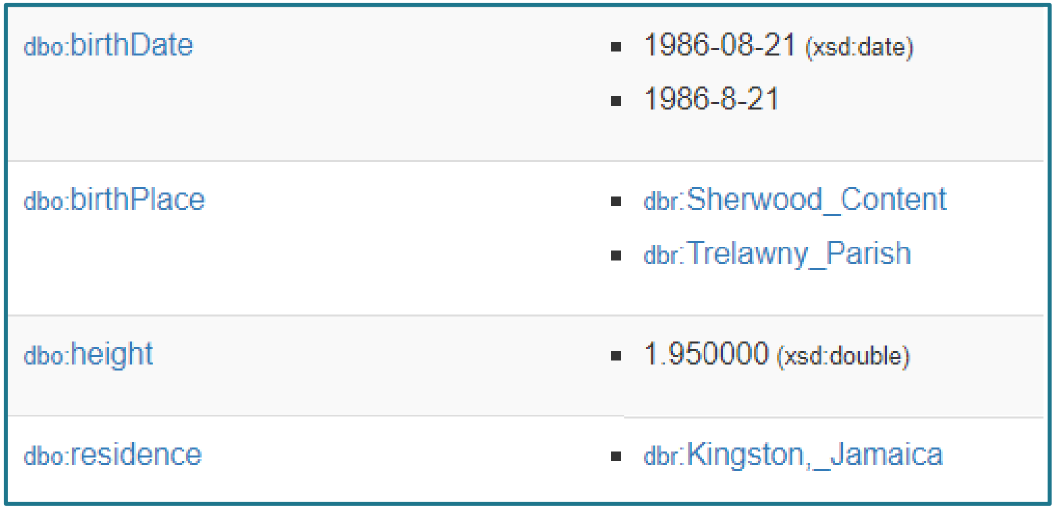
Contents [hide]

The Honourable
Usain Bolt
OJ CD

Bolt at the 2016 Summer Olympics

Personal information	
Full name	Usain St Leo Bolt
Nickname(s)	Lightning Bolt[1]
Nationality	Jamaican

Wikipédia



dbo:birthDate	▪ 1986-08-21 (xsd:date) ▪ 1986-8-21
dbo:birthPlace	▪ dbr:Sherwood_Content ▪ dbr:Trelawny_Parish
dbo:height	▪ 1.950000 (xsd:double)
dbo:residence	▪ dbr:Kingston,_Jamaica

DBpédia

Dbpédia et Wikipédia

- Dbpedia = Fiche signalétique de Wikipédia

- Permet des questions combinées :

Ex : Tous les athlètes nés entre 1810 et 1880.

Remplissage base de données automatique

Problèmes qui en résultent :

- Produit des erreurs
- Mauvaise structuration possible

Interroger Dbpédia nécessite :

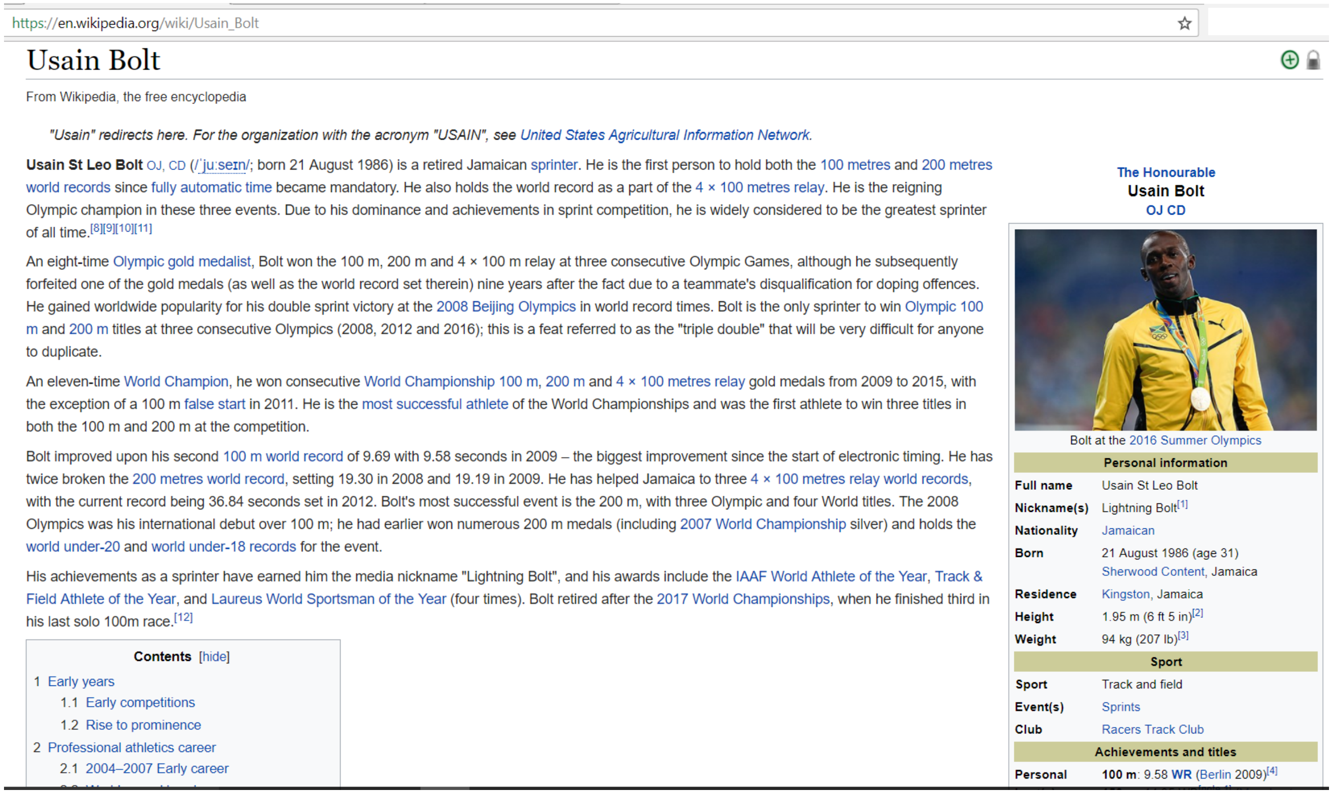
- De comprendre ce qu'est le **web sémantique** (notamment pour les ontologies et préfixes)
- D'être familiariser avec **l'organisation de Dbpédia**
- De faire des explorations préalables pour savoir quels **prédicats de Dbpédia** sont généralement utilisés dans le cas souhaité
- Apprendre le **langage SPARQL**
- De savoir quel **point de terminaison** utiliser

  (et savoir ce que c'est !)

Intérêt de mon interface

- Pas de connaissance du web sémantique nécessaire
- Pas de connaissance de Dbpédia
- Les ontologies sont déjà sélectionnées
- Pas de langage à connaitre
- Le point de terminaison est déjà fixé

Schéma de l’organisation



https://en.wikipedia.org/wiki/Usain_Bolt

Usain Bolt

From Wikipedia, the free encyclopedia

"Usain" redirects here. For the organization with the acronym "USAIN", see United States Agricultural Information Network.

Usain St Leo Bolt OJ, CD (/ˈjuːseɪn/; born 21 August 1986) is a retired Jamaican sprinter. He is the first person to hold both the 100 metres and 200 metres world records since fully automatic time became mandatory. He also holds the world record as a part of the 4 × 100 metres relay. He is the reigning Olympic champion in these three events. Due to his dominance and achievements in sprint competition, he is widely considered to be the greatest sprinter of all time.[8][9][10][11]

An eight-time Olympic gold medalist, Bolt won the 100 m, 200 m and 4 × 100 m relay at three consecutive Olympic Games, although he subsequently forfeited one of the gold medals (as well as the world record set therein) nine years after the fact due to a teammate's disqualification for doping offences. He gained worldwide popularity for his double sprint victory at the 2008 Beijing Olympics in world record times. Bolt is the only sprinter to win Olympic 100 m and 200 m titles at three consecutive Olympics (2008, 2012 and 2016); this is a feat referred to as the "triple double" that will be very difficult for anyone to duplicate.

An eleven-time World Champion, he won consecutive World Championship 100 m, 200 m and 4 × 100 metres relay gold medals from 2009 to 2015, with the exception of a 100 m false start in 2011. He is the most successful athlete of the World Championships and was the first athlete to win three titles in both the 100 m and 200 m at the competition.

Bolt improved upon his second 100 m world record of 9.69 with 9.58 seconds in 2009 – the biggest improvement since the start of electronic timing. He has twice broken the 200 metres world record, setting 19.30 in 2008 and 19.19 in 2009. He has helped Jamaica to three 4 × 100 metres relay world records, with the current record being 36.84 seconds set in 2012. Bolt's most successful event is the 200 m, with three Olympic and four World titles. The 2008 Olympics was his international debut over 100 m; he had earlier won numerous 200 m medals (including 2007 World Championship silver) and holds the world under-20 and world under-18 records for the event.

His achievements as a sprinter have earned him the media nickname "Lightning Bolt", and his awards include the IAAF World Athlete of the Year, Track & Field Athlete of the Year, and Laureus World Sportsman of the Year (four times). Bolt retired after the 2017 World Championships, when he finished third in his last solo 100m race.[12]

Contents [hide]

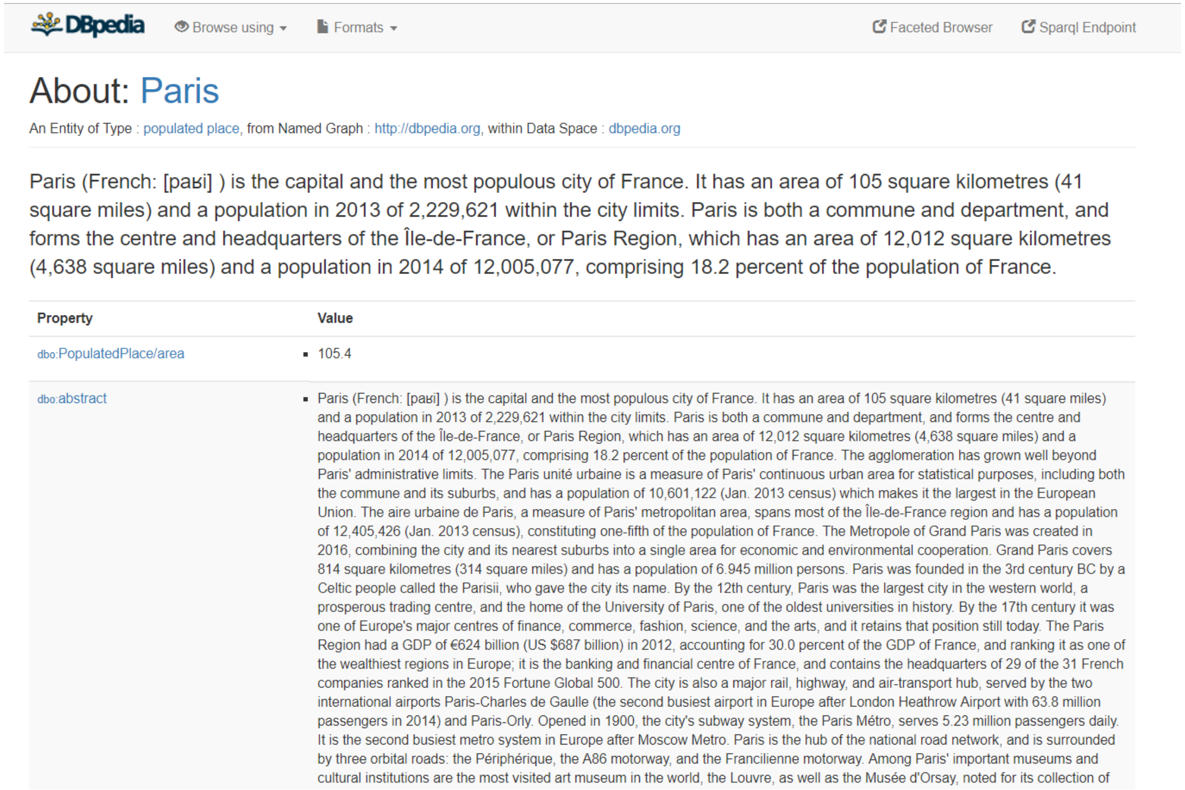
The Honourable
Usain Bolt
OJ CD

Bolt at the 2016 Summer Olympics

Personal information	
Full name	Usain St Leo Bolt
Nickname(s)	Lightning Bolt[1]
Nationality	Jamaican
Born	21 August 1986 (age 31) Sherwood Content, Jamaica
Residence	Kingston, Jamaica
Height	1.95 m (6 ft 5 in)[2]
Weight	94 kg (207 lb)[3]
Sport	
Sport	Track and field
Event(s)	Sprints
Club	Racers Track Club
Achievements and titles	
Personal	**100 m**: 9.58 **WR** (Berlin 2009)[4]

Wikipédia

Schéma de l’organisation

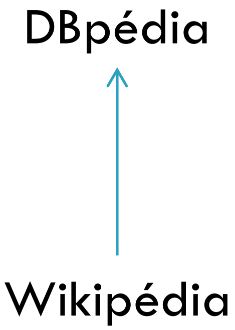


About: Paris

An Entity of Type : populated place, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Paris (French: [paʁi]) is the capital and the most populous city of France. It has an area of 105 square kilometres (41 square miles) and a population in 2013 of 2,229,621 within the city limits. Paris is both a commune and department, and forms the centre and headquarters of the Île-de-France, or Paris Region, which has an area of 12,012 square kilometres (4,638 square miles) and a population in 2014 of 12,005,077, comprising 18.2 percent of the population of France.

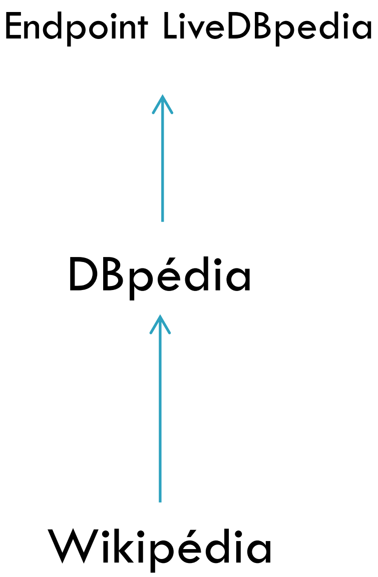
Property	Value
dbo:PopulatedPlace/area	▪ 105.4
dbo:abstract	▪ Paris (French: [paʁi]) is the capital and the most populous city of France. It has an area of 105 square kilometres (41 square miles) and a population in 2013 of 2,229,621 within the city limits. Paris is both a commune and department, and forms the centre and headquarters of the Île-de-France, or Paris Region, which has an area of 12,012 square kilometres (4,638 square miles) and a population in 2014 of 12,005,077, comprising 18.2 percent of the population of France. The agglomeration has grown well beyond Paris' administrative limits. The Paris unité urbaine is a measure of Paris' continuous urban area for statistical purposes, including both the commune and its suburbs, and has a population of 10,601,122 (Jan. 2013 census) which makes it the largest in the European Union. The aire urbaine de Paris, a measure of Paris' metropolitan area, spans most of the Île-de-France region and has a population of 12,405,426 (Jan. 2013 census), constituting one-fifth of the population of France. The Metropole of Grand Paris was created in 2016, combining the city and its nearest suburbs into a single area for economic and environmental cooperation. Grand Paris covers 814 square kilometres (314 square miles) and has a population of 6.945 million persons. Paris was founded in the 3rd century BC by a Celtic people called the Parisii, who gave the city its name. By the 12th century, Paris was the largest city in the western world, a prosperous trading centre, and the home of the University of Paris, one of the oldest universities in history. By the 17th century it was one of Europe's major centres of finance, commerce, fashion, science, and the arts, and it retains that position still today. The Paris Region had a GDP of €624 billion (US $687 billion) in 2012, accounting for 30.0 percent of the GDP of France, and ranking it as one of the wealthiest regions in Europe; it is the banking and financial centre of France, and contains the headquarters of 29 of the 31 French companies ranked in the 2015 Fortune Global 500. The city is also a major rail, highway, and air-transport hub, served by the two international airports Paris-Charles de Gaulle (the second busiest airport in Europe after London Heathrow Airport with 63.8 million passengers in 2014) and Paris-Orly. Opened in 1900, the city's subway system, the Paris Métro, serves 5.23 million passengers daily. It is the second busiest metro system in Europe after Moscow Metro. Paris is the hub of the national road network, and is surrounded by three orbital roads: the Périphérique, the A86 motorway, and the Francilienne motorway. Among Paris' important museums and cultural institutions are the most visited art museum in the world, the Louvre, as well as the Musée d'Orsay, noted for its collection of



DBpédia

Wikipédia

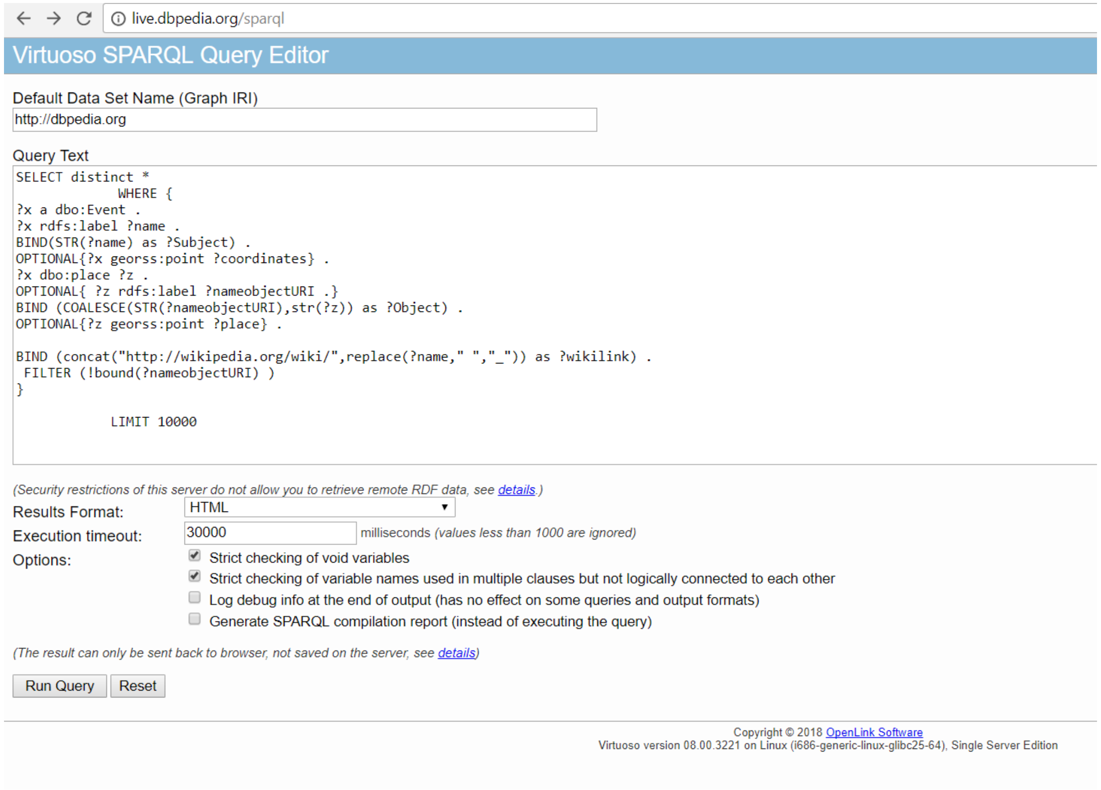
Schéma de l'organisation



Endpoint LiveDBpedia

DBpédia

Wikipédia



Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

http://dbpedia.org

Query Text

```
SELECT distinct *
             WHERE {
?x a dbo:Event .
?x rdfs:label ?name .
BIND(STR(?name) as ?Subject) .
OPTIONAL{?x georss:point ?coordinates} .
?x dbo:place ?z .
OPTIONAL{ ?z rdfs:label ?nameobjectURI .}
BIND (COALESCE(STR(?nameobjectURI),str(?z)) as ?Object) .
OPTIONAL{?z georss:point ?place} .

BIND (concat("http://wikipedia.org/wiki/",replace(?name," ","_")) as ?wikilink) .
 FILTER (!bound(?nameobjectURI) )
}

            LIMIT 10000
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see details.)

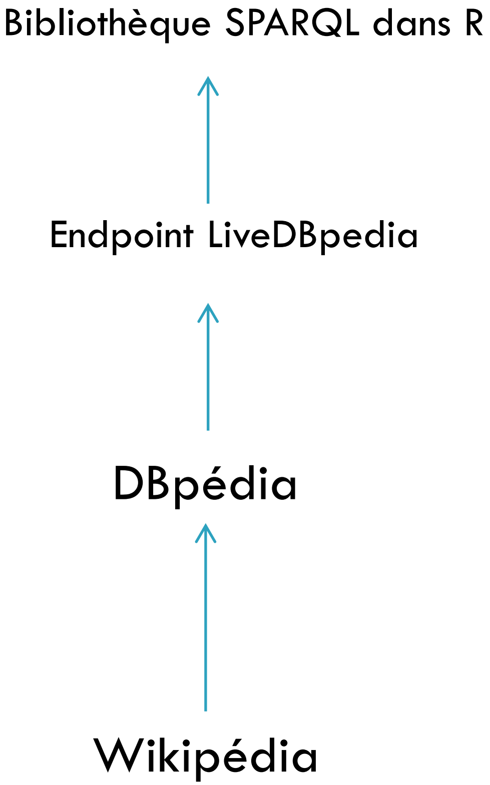
Results Format: HTML ▾

Execution timeout: 30000 milliseconds *(values less than 1000 are ignored)*

Options:
☑ Strict checking of void variables
☑ Strict checking of variable names used in multiple clauses but not logically connected to each other
□ Log debug info at the end of output (has no effect on some queries and output formats)
□ Generate SPARQL compilation report (instead of executing the query)

(The result can only be sent back to browser, not saved on the server, see details)

Schéma de l’organisation



Bibliothèque SPARQL dans R

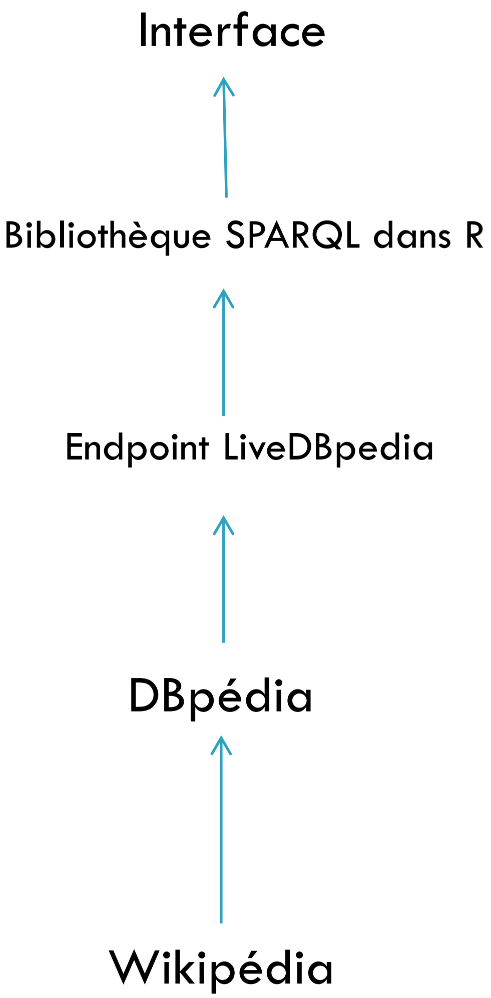
Endpoint LiveDBpedia

DBpédia

Wikipédia

SPARQL Protocol and RDF Query Language

Schéma de l'organisation

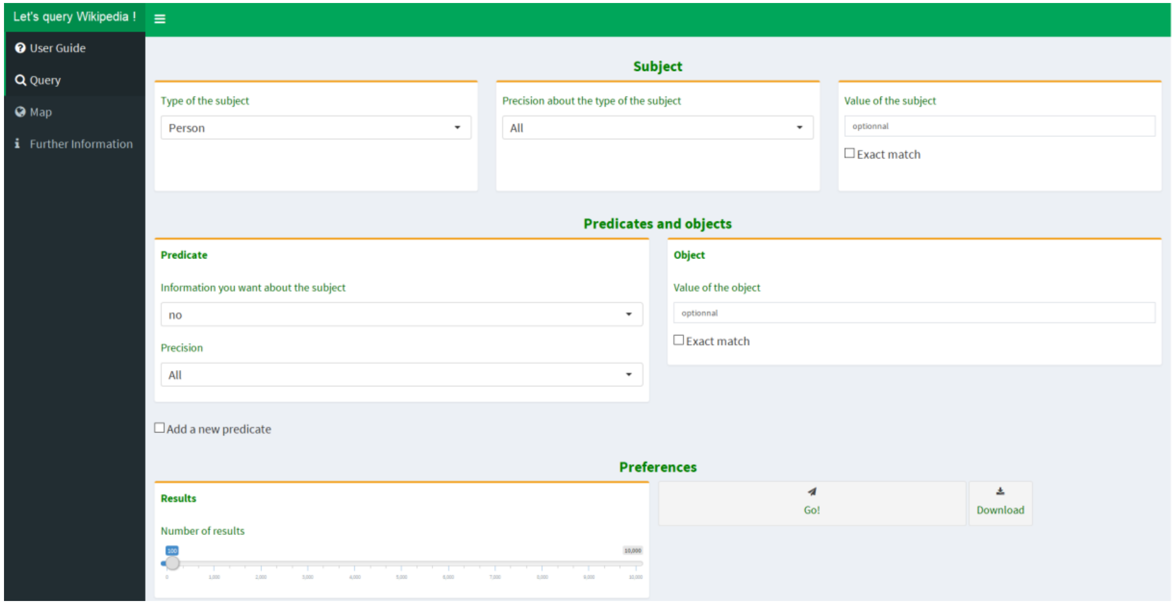

Interface

Bibliothèque SPARQL dans R

Endpoint LiveDBpedia

DBpédia

Wikipédia



Subject

Type of the subject: Person

Precision about the type of the subject: All

Value of the subject: optionnal

□ Exact match

Predicates and objects

Predicate

Information you want about the subject: no

Precision: All

Object

Value of the object: optionnal

□ Exact match

□ Add a new predicate

Preferences

Results

Number of results: 100 (0 – 10,000)

SPARQL

- □ langage universel verbeux
- □ Permet d'interroger un triplestore

Triplestore = base de données au format RDF

SPARQL

Chercher toutes les personnes,
Qui sont des athlètes.
Limiter à 300

```
SELECT distinct *
WHERE {
?person a dbo:Person .
?person a dbo:Athlete .
}
LIMIT 300
```

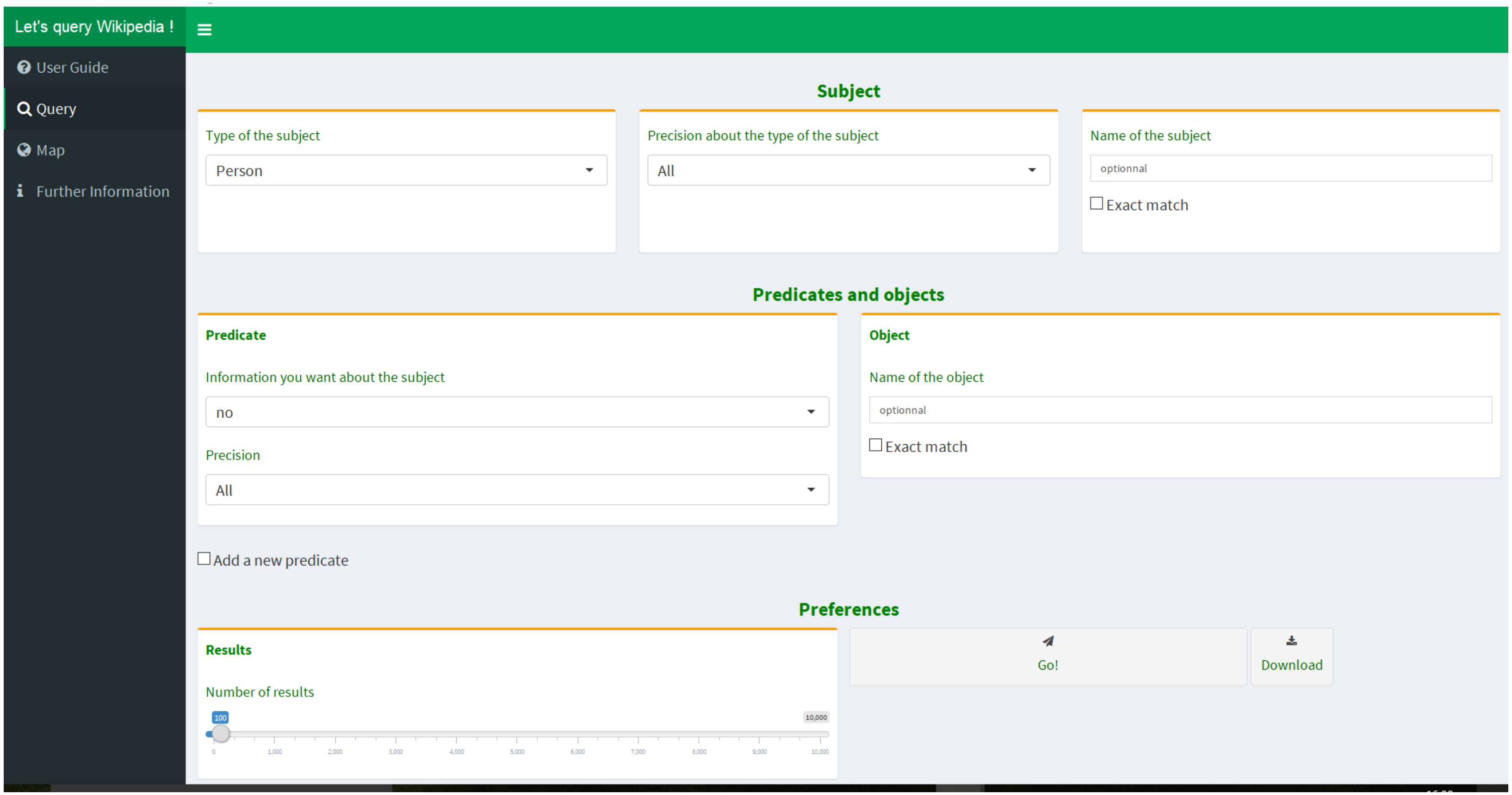
SPARQL

Chercher le lieu de naissance de Teddy Riner.

```
SELECT distinct *
WHERE {
?person a dbo:Person .
?person rdfs:label « Teddy Riner »@en .
.?person dbo:birthPlace ?place .
}
```

Interface

□ Présentation



Let's query Wikipedia !

Subject

Type of the subject: Person

Precision about the type of the subject: All

Name of the subject: optionnal

□ Exact match

Predicates and objects

Predicate

Information you want about the subject: no

Precision: All

Object

Name of the object: optionnal

□ Exact match

□ Add a new predicate

Preferences

Results

Number of results: 100

Conclusion

- Facile d’utilisation
- Complète
- Disponible en ligne

- Perspectives :
  - Complexifier le schéma des prédicats
  - Choisir lieu sur la carte
  - Carte de la galaxie

Apports

- Découverte du Web sémantique et de ses concepts
- Projet à portée très large
- Pas de limites d’utilisation

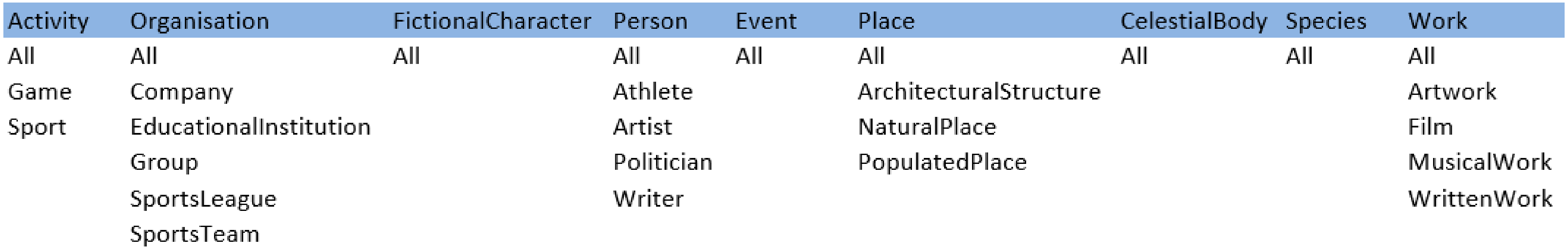
Apports

- !!! est un groupe de rock indépendant américain
- !!! est le nom d’un de leur album
- Will Smith est aussi chanteur de rap
- Et beaucoup d’autres choses !

Merci de votre attention !

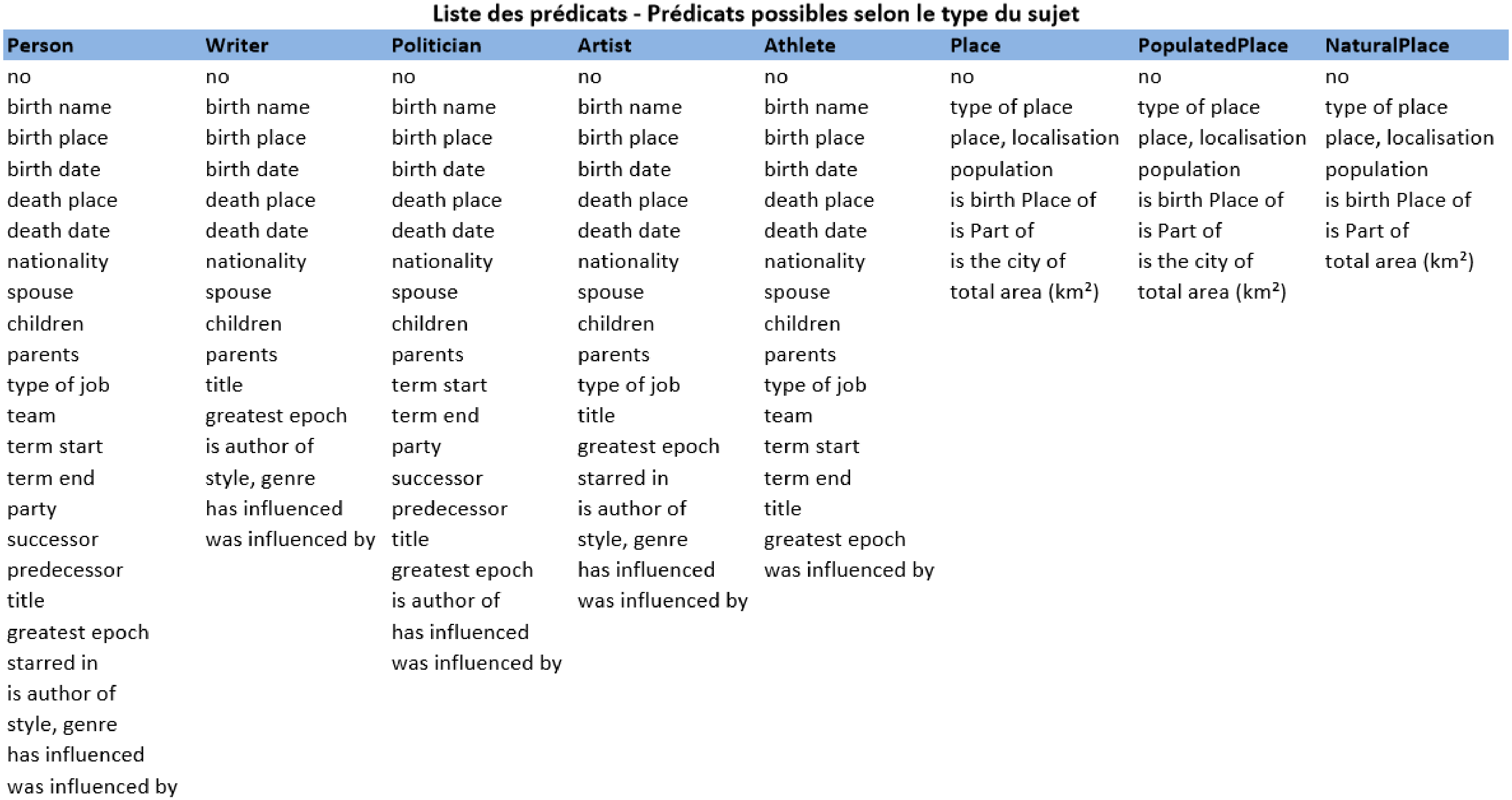
Des questions ?

CSV



Activity	Organisation	FictionalCharacter	Person	Event	Place	CelestialBody	Species	Work
All	All	All	All	All	All	All	All	All
Game	Company		Athlete		ArchitecturalStructure			Artwork
Sport	EducationalInstitution		Artist		NaturalPlace			Film
	Group		Politician		PopulatedPlace			MusicalWork
	SportsLeague		Writer					WrittenWork
	SportsTeam							

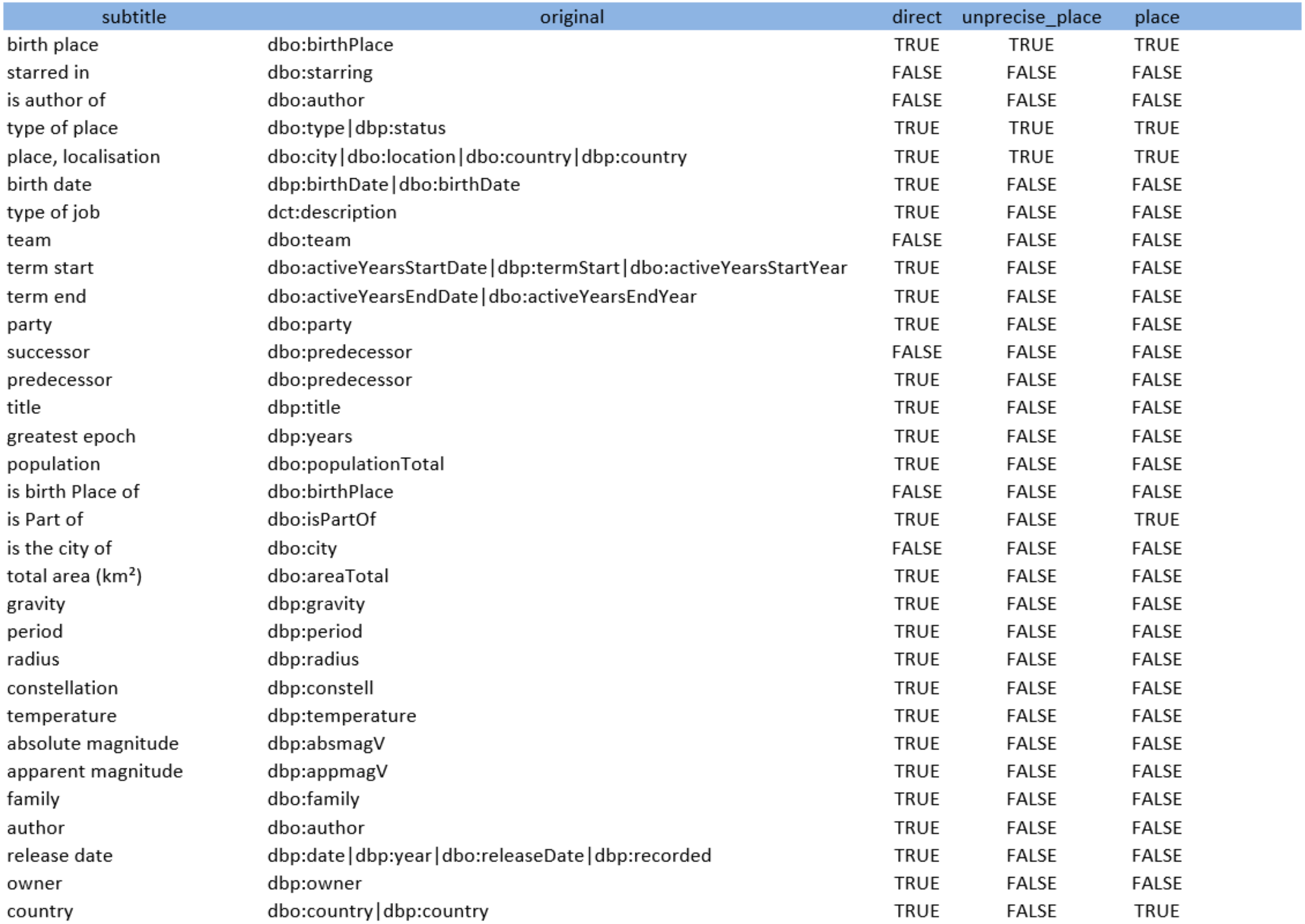
CSV



Liste des prédicats - Prédicats possibles selon le type du sujet

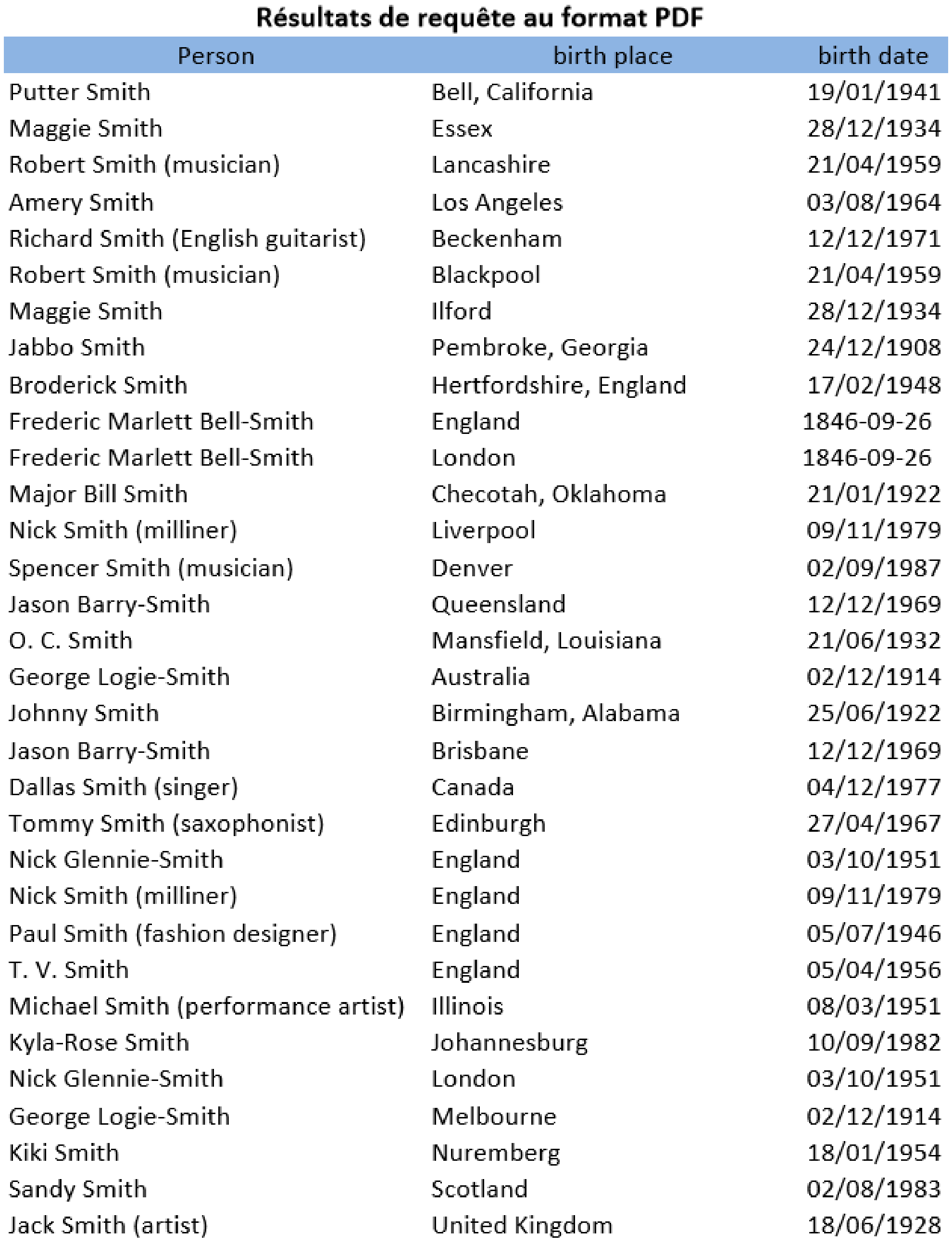
Person	Writer	Politician	Artist	Athlete	Place	PopulatedPlace	NaturalPlace
no	no	no	no	no	no	no	no
birth name	birth name	birth name	birth name	birth name	type of place	type of place	type of place
birth place	birth place	birth place	birth place	birth place	place, localisation	place, localisation	place, localisation
birth date	birth date	birth date	birth date	birth date	population	population	population
death place	death place	death place	death place	death place	is birth Place of	is birth Place of	is birth Place of
death date	death date	death date	death date	death date	is Part of	is Part of	is Part of
nationality	nationality	nationality	nationality	nationality	is the city of	is the city of	total area (km^2)
spouse	spouse	spouse	spouse	spouse	total area (km^2)	total area (km^2)	
children	children	children	children	children			
parents	parents	parents	parents	parents			
type of job	title	term start	type of job	type of job			
team	greatest epoch	term end	title	team			
term start	is author of	party	greatest epoch	term start			
term end	style, genre	successor	starred in	term end			
party	has influenced	predecessor	is author of	title			
successor	was influenced by	title	style, genre	greatest epoch			
predecessor		greatest epoch	has influenced	was influenced by			
title		is author of	was influenced by				
greatest epoch		has influenced					
starred in		was influenced by					
is author of							
style, genre							
has influenced							
was influenced by							

CSV



subtitle	original	direct	unprecise_place	place
birth place	dbo:birthPlace	TRUE	TRUE	TRUE
starred in	dbo:starring	FALSE	FALSE	FALSE
is author of	dbo:author	FALSE	FALSE	FALSE
type of place	dbo:type|dbp:status	TRUE	TRUE	TRUE
place, localisation	dbo:city|dbo:location|dbo:country|dbp:country	TRUE	TRUE	TRUE
birth date	dbp:birthDate|dbo:birthDate	TRUE	FALSE	FALSE
type of job	dct:description	TRUE	FALSE	FALSE
team	dbo:team	FALSE	FALSE	FALSE
term start	dbo:activeYearsStartDate|dbp:termStart|dbo:activeYearsStartYear	TRUE	FALSE	FALSE
term end	dbo:activeYearsEndDate|dbo:activeYearsEndYear	TRUE	FALSE	FALSE
party	dbo:party	TRUE	FALSE	FALSE
successor	dbo:predecessor	FALSE	FALSE	FALSE
predecessor	dbo:predecessor	TRUE	FALSE	FALSE
title	dbp:title	TRUE	FALSE	FALSE
greatest epoch	dbp:years	TRUE	FALSE	FALSE
population	dbo:populationTotal	TRUE	FALSE	FALSE
is birth Place of	dbo:birthPlace	FALSE	FALSE	FALSE
is Part of	dbo:isPartOf	TRUE	FALSE	TRUE
is the city of	dbo:city	FALSE	FALSE	FALSE
total area (km²)	dbo:areaTotal	TRUE	FALSE	FALSE
gravity	dbp:gravity	TRUE	FALSE	FALSE
period	dbp:period	TRUE	FALSE	FALSE
radius	dbp:radius	TRUE	FALSE	FALSE
constellation	dbp:constell	TRUE	FALSE	FALSE
temperature	dbp:temperature	TRUE	FALSE	FALSE
absolute magnitude	dbp:absmagV	TRUE	FALSE	FALSE
apparent magnitude	dbp:appmagV	TRUE	FALSE	FALSE
family	dbo:family	TRUE	FALSE	FALSE
author	dbo:author	TRUE	FALSE	FALSE
release date	dbp:date|dbp:year|dbo:releaseDate|dbp:recorded	TRUE	FALSE	FALSE
owner	dbp:owner	TRUE	FALSE	FALSE
country	dbo:country|dbp:country	TRUE	FALSE	TRUE

CSV



Résultats de requête au format PDF

Person	birth place	birth date
Putter Smith	Bell, California	19/01/1941
Maggie Smith	Essex	28/12/1934
Robert Smith (musician)	Lancashire	21/04/1959
Amery Smith	Los Angeles	03/08/1964
Richard Smith (English guitarist)	Beckenham	12/12/1971
Robert Smith (musician)	Blackpool	21/04/1959
Maggie Smith	Ilford	28/12/1934
Jabbo Smith	Pembroke, Georgia	24/12/1908
Broderick Smith	Hertfordshire, England	17/02/1948
Frederic Marlett Bell-Smith	England	1846-09-26
Frederic Marlett Bell-Smith	London	1846-09-26
Major Bill Smith	Checotah, Oklahoma	21/01/1922
Nick Smith (milliner)	Liverpool	09/11/1979
Spencer Smith (musician)	Denver	02/09/1987
Jason Barry-Smith	Queensland	12/12/1969
O. C. Smith	Mansfield, Louisiana	21/06/1932
George Logie-Smith	Australia	02/12/1914
Johnny Smith	Birmingham, Alabama	25/06/1922
Jason Barry-Smith	Brisbane	12/12/1969
Dallas Smith (singer)	Canada	04/12/1977
Tommy Smith (saxophonist)	Edinburgh	27/04/1967
Nick Glennie-Smith	England	03/10/1951
Nick Smith (milliner)	England	09/11/1979
Paul Smith (fashion designer)	England	05/07/1946
T. V. Smith	England	05/04/1956
Michael Smith (performance artist)	Illinois	08/03/1951
Kyla-Rose Smith	Johannesburg	10/09/1982
Nick Glennie-Smith	London	03/10/1951
George Logie-Smith	Melbourne	02/12/1914
Kiki Smith	Nuremberg	18/01/1954
Sandy Smith	Scotland	02/08/1983
Jack Smith (artist)	United Kingdom	18/06/1928