université PARIS-SACLAY

Paris-Saclay Center for Data Science

Input-Ouput Data Science of Paris-Saclay

The data platform

The linked data platform

The linked open data platform

Karima Rafes – BorderCloud
Karima.rafes@bordercloud.com

Cécile Germain – LRI, Université Paris Sud, CNRS, INRIA
cecile.germain@lri.fr

What **must to have** the Data Scientists in 2017 ?

A collaborative Data Science IT Platform for :

- Sharing
- Linking
- Reproducing
- Starting to imagine the futures strategies for the management of Scientific Information

Design principles

- Respectful, volunteer: centralization or intrusiveness is excluded
  - Paris-Saclay is cross section of institutional science: massive heterogeneity in scale, organizations, data policies, resources,...
  - CDS is a scientific collaboration, not an IT service
  - We « build on top » of existing practices and repositories
- **Collaborative**, agile: Scientific knowledge is a continuous process
  - Empower grassroots scientists with the tools to build and re-build the domain vision
  - Following mainstream approaches and standards

Data Science IT Platform, today

Input-Ouput Data Science of Paris-Saclay

https://io.datascience-paris-saclay.fr

For the friends

IODS

- Sharing : Map the existing data repositories operated by CDS partners
  - Better knowledge of the available resources.
  - Helping collaborations to get off the ground.
- Linking : Interoperable data
  - Provide tools, usable across disciplines
  - Build a Linked (Open) Data platform to publish these.
  - Linking data between domains: collective knowledge building
- Infrastructure for reproducibility.
  - Ability to test algorithms on new or extended datasets
  - Provide tools to reuse the queries of interoperable data

université PARIS-SACLAY Paris-Saclay Center for Data Science

CDS Datasets

- 34 datasets from CDS visible
- Objective for CDS2: be reasonably representative of the datasets in the CDS2 perimeter
  - Proprietary or sensitive excluded for now
  - Remember: you keep your data in-house
- Objectives of the demo
  - lods basics: how to publish a dataset - much easier than you imagine
    - A few clicks
    - Any format, eg excel, text, community, sql, nosql, rdf…
  - lods advanced: linking data
- Help available: karima.rafes@gmail.com

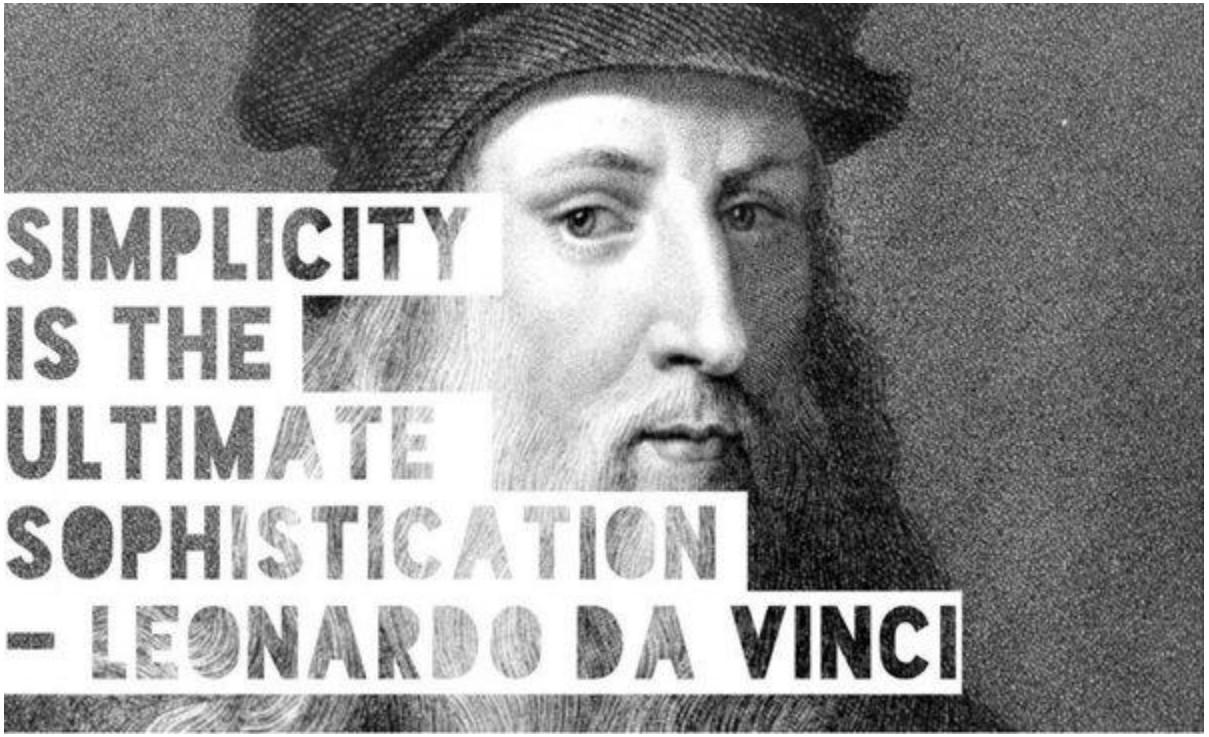
université PARIS-SACLAY
Paris-Saclay Center for Data Science

CDS Datasets

Poll

https://goo.gl/forms/OuxdalXIphg0b6fM2

- Link also on the indico page
- Please fill it by noon



DEMO
IODS BASICS

Demo : Sharing any data

- Unique identifier (& free) for each dataset
- Different access level :
  - public, share (for partner of CDS) or private
- Description :
  - What : Card, Licence, size, localisation, audience, etc
  - Distributions : Files (excel, image,…) & API
  - Exemples : queries
  - Links with the knowledge bases

Wikidata

general knowledge | Semantic Web | Wikipedia | Wikidata | Wikivoyage | Wikisource

Description

Wikidata is a free linked database for the structured data of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wikisource and others.
The content is available under a free license, exported using standard formats, and can be interlinked to other open data sets on

https://www.wikidata.org | karima.rafes@gmail.com

Location

Distribution

3★DATA SPARQL endpoint — cc0 srx json osv 2018-01-31 — Details | Go

2★WWW RDF Exports — cc0 nt gz json 2018-01-31 — Details | Go

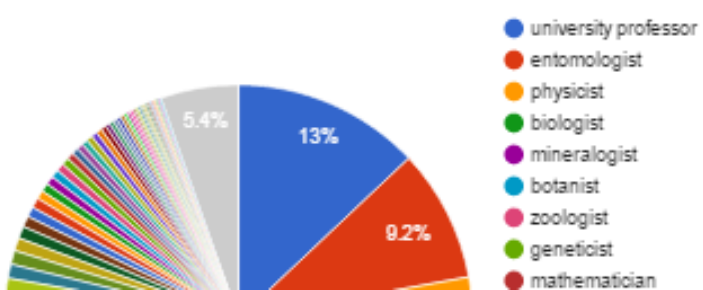
The content is not displayed?

Publishers by Occupation

science | computer science | scientist | profession

This query retrieves the number of scientists publishing in Wikidata.

5

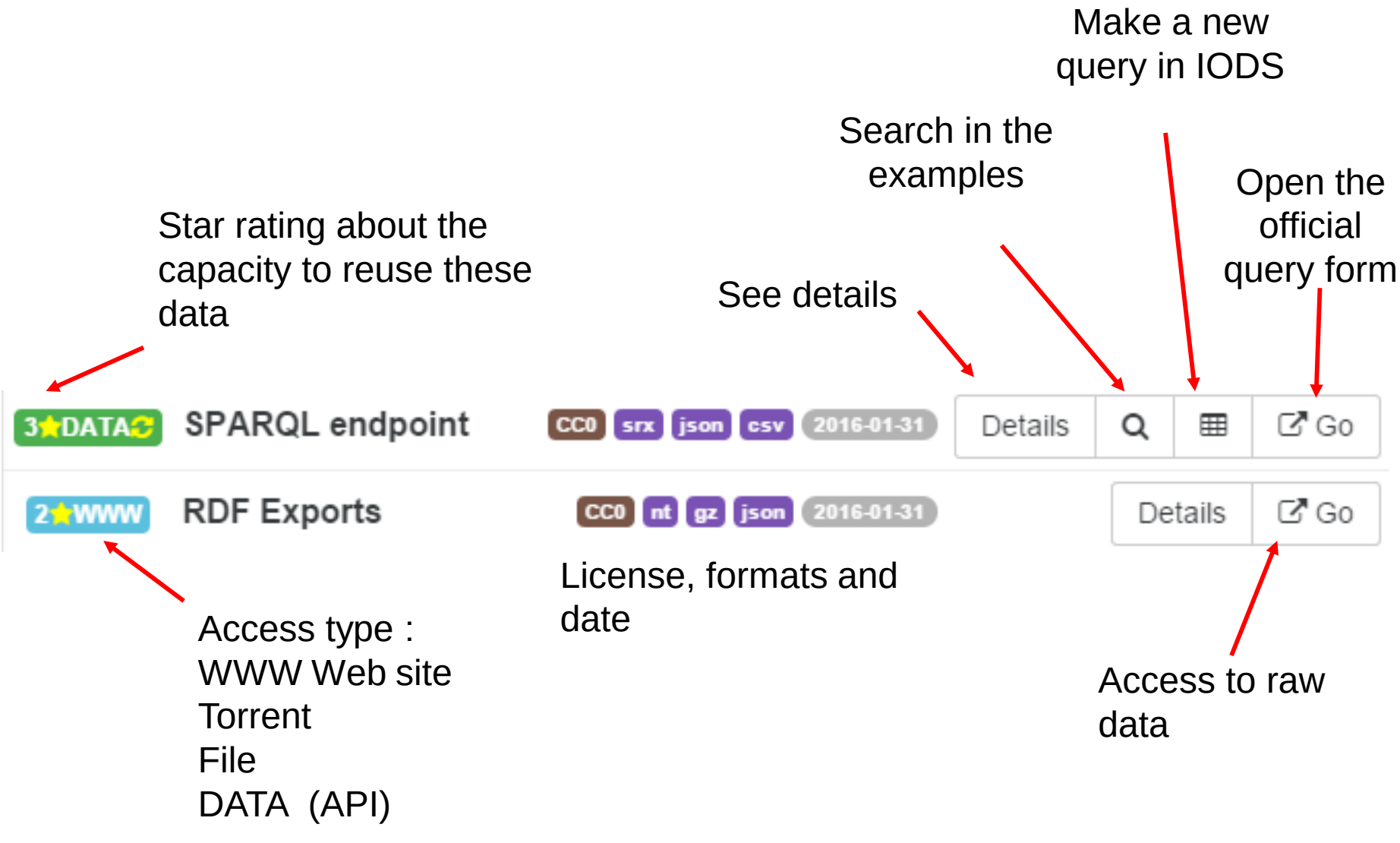


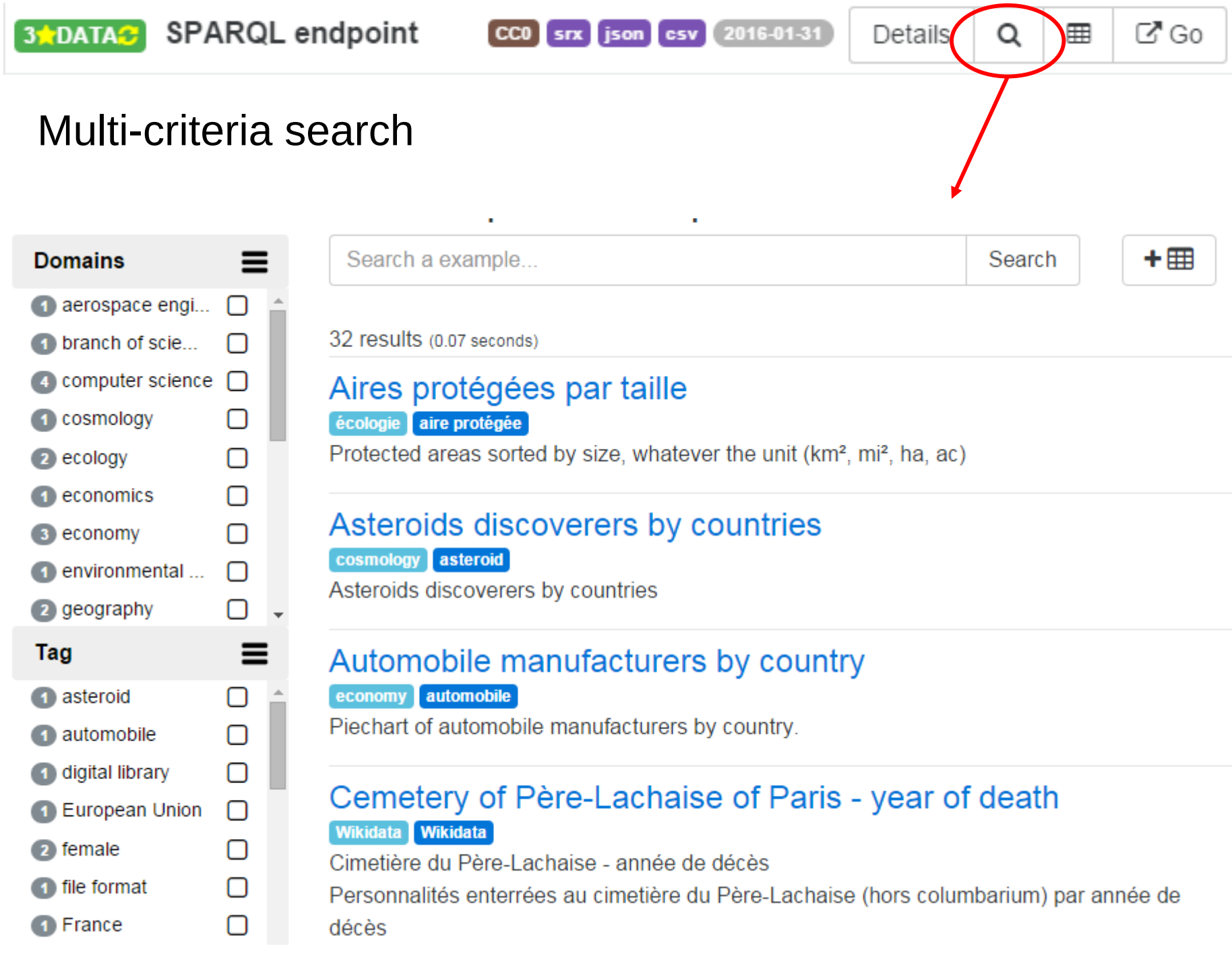
Links to Wikidata.
Scientific topics (light blue)
Tag about the content (dark blue)

Description

Data distribution

Examples to reuse


Make a new query in IODS
Search in the examples
Open the official query form
See details
Star rating about the capacity to reuse these data
3★DATA
SPARQL endpoint
CC0
srx
json
csv
2016-01-31
Details
Go
2★WWW
RDF Exports
CC0
nt
gz
json
2016-01-31
Details
Go
License, formats and date
Access type :
WWW Web site
Torrent
File
DATA (API)
Access to raw data


3★DATA
SPARQL endpoint
CC0
srx
json
csv
2016-01-31
Details
Go
Multi-criteria search
Domains
1 aerospace engi...
1 branch of scie...
4 computer science
1 cosmology
2 ecology
1 economics
3 economy
1 environmental ...
2 geography
Tag
1 asteroid
1 automobile
1 digital library
1 European Union
2 female
1 file format
1 France
Search a example...
Search
32 results (0.07 seconds)
Aires protégées par taille
écologie
aire protégée
Protected areas sorted by size, whatever the unit (km², mi², ha, ac)
Asteroids discoverers by countries
cosmology
asteroid
Asteroids discoverers by countries
Automobile manufacturers by country
economy
automobile
Piechart of automobile manufacturers by country.
Cemetery of Père-Lachaise of Paris - year of death
Wikidata
Wikidata
Cimetière du Père-Lachaise - année de décès
Personnalités enterrées au cimetière du Père-Lachaise (hors columbarium) par année de décès

Demo : Linking the knowledge

With the linked data platform :

You facilitate the discovering of your data.

1. Describe a dataset in IODS
2. Install CDS' APPs in Wikipedia
3. Search a topic in your prefered search engine…
4. Open the Wikipedia's page

➔ Discover the available data and more !

Raman spectroscopy (Q862228 ttl rdf/xml json sparql Export : labels)

Datasets in relation

- Drug Classification : The dataset contain Raman spectra of 4 types of chemotherapeutic agents diluted in 9 different solutions, and having different concentrations. Measures were made by the Lip(Sys)². (source)

Devices in relation

RamanEvolution - Organization : Lip(Sys)²

Inria teams in relation

Hyperspectral imaging | Mutual exclusion

From Wikipedia, the free encyclopedia

DEMO
IODS ADVANCED

Demo : Test and reuse

With the linked data platform :

You facilitate the reusing of your data.

1. Choose an example
2. Copy
3. Change the query
4. Save it
5. Copy/Past the code in your algorithm/program/Wiki

➔ Reuse the Linked Data is simple.

Linked Data example : table, graphs…

Asteroids discoverers by countries

cosmology asteroid

Links to Wikidata.
Scientific topics (light blue)
Tag about the content (dark blue)

Asteroids discoverers by countries

Description

Tools

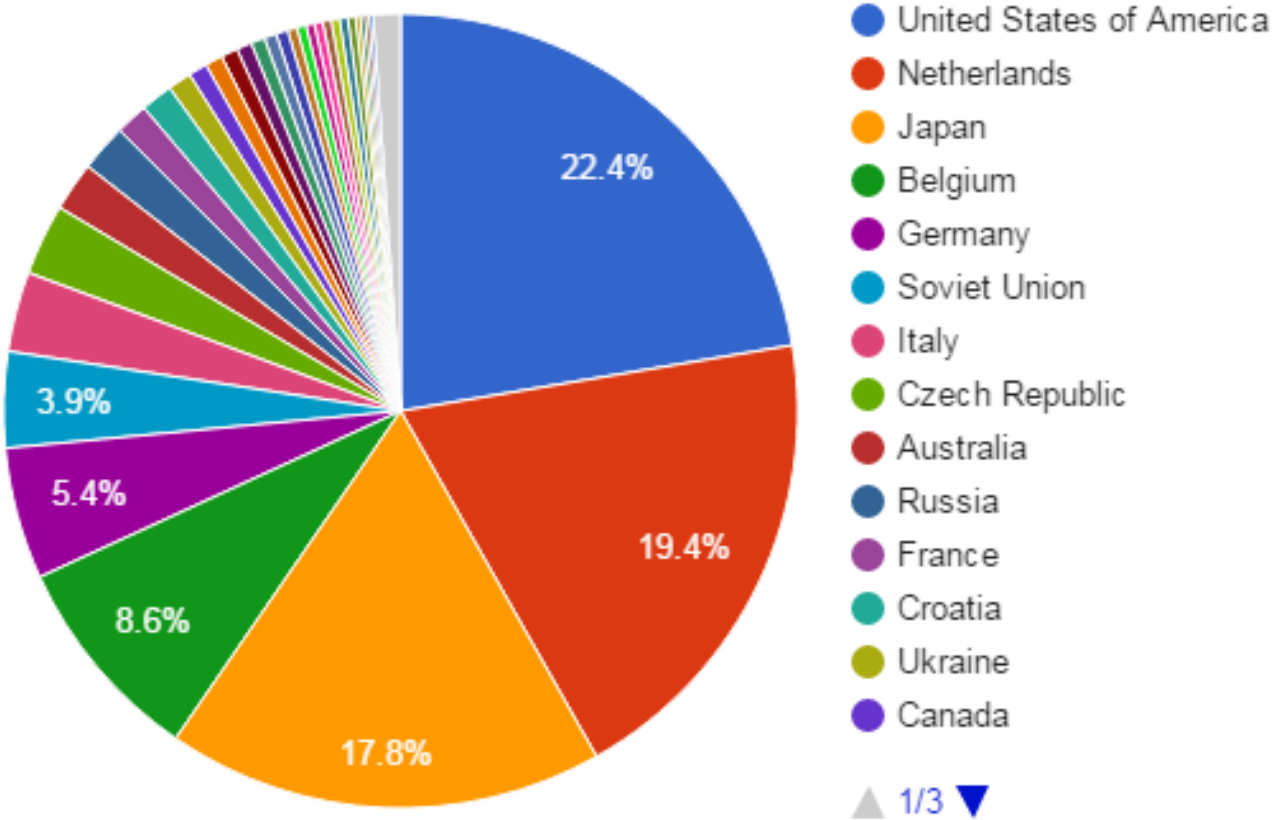


Table ou graph

Share Catch 1 Reuse Full screen Do a copy

Reuse these data in your code

Query, endpoint and code for reusing the same data

2 Choose your langage

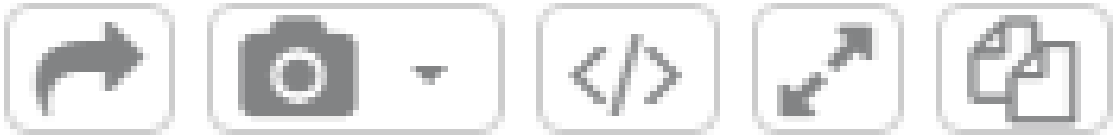
SPARQL | Javascript | HTML | Matlab | Python | R | PHP | Wiki

3 Copy/past the code

Copy

```
from SPARQLWrapper import SPARQLWrapper, JSON

sparql = SPARQLWrapper("https://query.wikidata.org/bigdata/namespace/wdq/sparql")
sparql.setQuery("""
    PREFIX bd: <http://www.bigdata.com/rdf#>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>

select   ?countryLabel (COUNT(?asteroid) as ?nb)
where {
        ?asteroid wdt:P31 wd:Q3863 ;
                  wdt:P61 ?discoverer .
     ?discoverer wdt:P27 ?country .

       SERVICE wikibase:label {
        bd:serviceParam wikibase:language "en" .
      } .

}
GROUP BY  ?countryLabel
ORDER BY DESC(?nb)""")
sparql.setReturnFormat(JSON)
results = sparql.query().convert()

for result in results["results"]["bindings"]:
    print(result["label"]["value"])
```

Howto use SPARQL with Python ? 4 May be also, read the doc ;-)



Copy and change a query

Copy of Asteroids discoverers by countries

Title: [COPY] Asteroids discoverers by countries

Domain: cosmology x Search domain

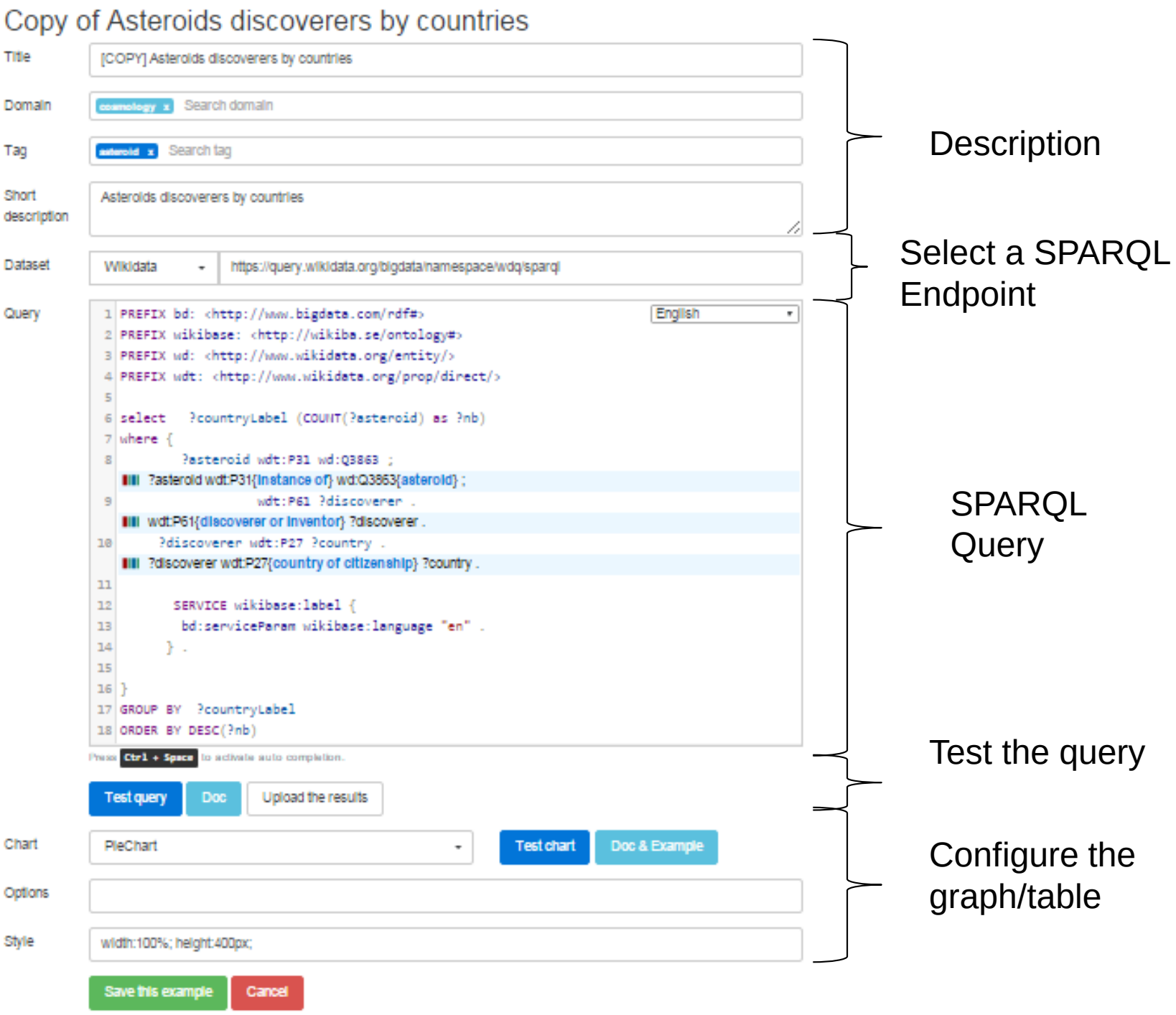
Tag: asteroid x Search tag

Short description: Asteroids discoverers by countries

Dataset: Wikidata | https://query.wikidata.org/bigdata/namespace/wdq/sparql

Query: English

```
1 PREFIX bd: <http://www.bigdata.com/rdf#>
2 PREFIX wikibase: <http://wikiba.se/ontology#>
3 PREFIX wd: <http://www.wikidata.org/entity/>
4 PREFIX wdt: <http://www.wikidata.org/prop/direct/>
5
6 select   ?countryLabel (COUNT(?asteroid) as ?nb)
7 where {
8         ?asteroid wdt:P31 wd:Q3863 ;
  ?asteroid wdt:P31{instance of} wd:Q3863{asteroid} ;
9                   wdt:P61 ?discoverer .
```


Copy of Asteroids discoverers by countries
Title
[COPY] Asteroids discoverers by countries
Domain
cosmology x
Search domain
Tag
asteroid x
Search tag
Short description
Asteroids discoverers by countries
Dataset
Wikidata
https://query.wikidata.org/bigdata/namespace/wdq/sparql
Query
English
PREFIX bd: <http://www.bigdata.com/rdf#>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
select ?countryLabel (COUNT(?asteroid) as ?nb)
where {
?asteroid wdt:P31 wd:Q3863 ;
?asteroid wdt:P31{instance of} wd:Q3863{asteroid} ;
wdt:P61 ?discoverer .
wdt:P61{discoverer or inventor} ?discoverer .
?discoverer wdt:P27 ?country .
?discoverer wdt:P27{country of citizenship} ?country .
SERVICE wikibase:label {
bd:serviceParam wikibase:language "en" .
} .
}
GROUP BY ?countryLabel
ORDER BY DESC(?nb)
Press Ctrl + Space to activate auto completion.
Test query
Doc
Upload the results
Chart
PieChart
Test chart
Doc & Example
Options
Style
width:100%; height:400px;
Save this example
Cancel
Description
Select a SPARQL Endpoint
SPARQL Query
Test the query
Configure the graph/table

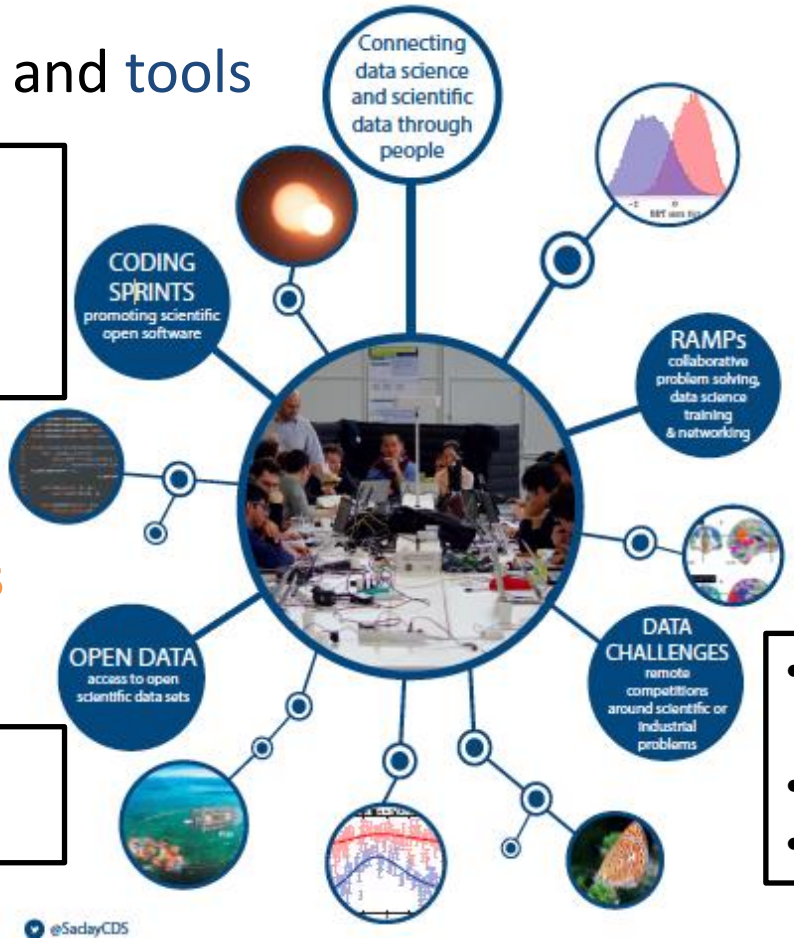
What next in 2018 ?

Connecting experts and problems

- Prototyping
- Training
- Collaboration building

Connecting experts and tools

- State-of-the-art data science in easy-to-use tools
- High-quality software



Connecting experts and data

- Data as a Service
- Linked (Open) Data

- Impact on science
- Visibility
- Benchmarks

université PARIS-SACLAY | Paris-Saclay Center for Data Science

Thanks, questions ?

Karima.rafes@bordercloud.com

Hangout : Karima.rafes@gmail.com

Twitter : @karima_rafes

« Long live the open science and proper »

Poll