

Décrire et Publier des jeux de données sur le web: vocabulaires, catalogues, portails...

ANF APSEM2018 : Apprentissage et sémantique

Toulouse, 12-15 novembre 2018

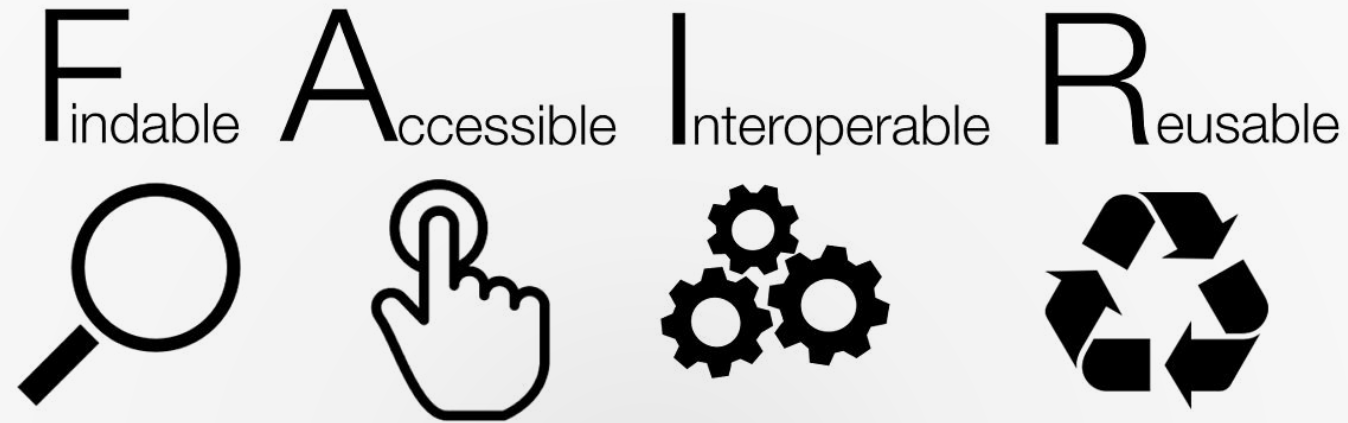
F. Michel

Université Côte d'Azur, CNRS, Inria, I3S, France

UNIVERSITÉ
CÔTE D'AZUR



Inria
inventeurs du monde numérique



*Making datasets **FAIR***

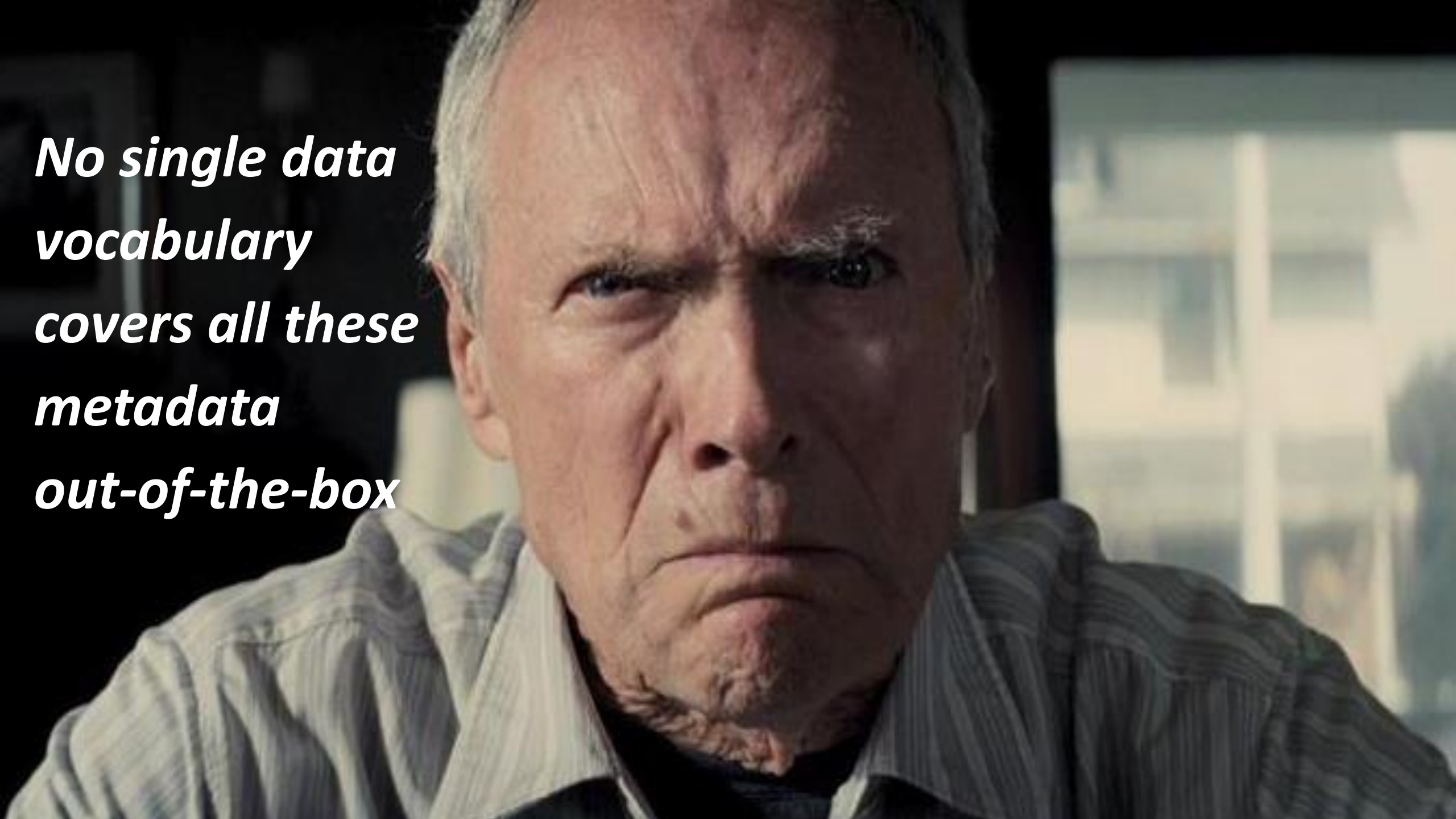
*requires quality, comprehensive **metadata***

*published in public **catalogues***

Metadata are the key to “FAIRness”

Context	Identification, authors/organizations, dates, version, reference articles, funding, reuse conditions (licenses, rights), coverage: time, space, topics
Access	Format, structure, location (download), query methods, protocols
Interpretation	What do the data represent? What concepts, entities, semantics? Reference vocabularies, dimensions, variables, time granularity, units (cm or inches, left/right)...
Provenance	Acquired with which equipment, parameters, protocols? Derived from which dataset? With which processing? Dataset-level or entity-level provenance
Statistics	Number of triples per property of class, links to other datasets...
...	

***No single data
vocabulary
covers all these
metadata
out-of-the-box***



Vocabularies to describe datasets and dataset catalogues

Standards about metadata and how to use them

About metadata

DCAT: Data Catalog Vocabulary

VOID: Vocabulary of Interlinked Datasets

Schema.org: Dataset and DataCatalog

About how to use metadata

HCLS: Health Care & Life Sciences Dataset Profile

DCAT, Data Catalog Vocabulary

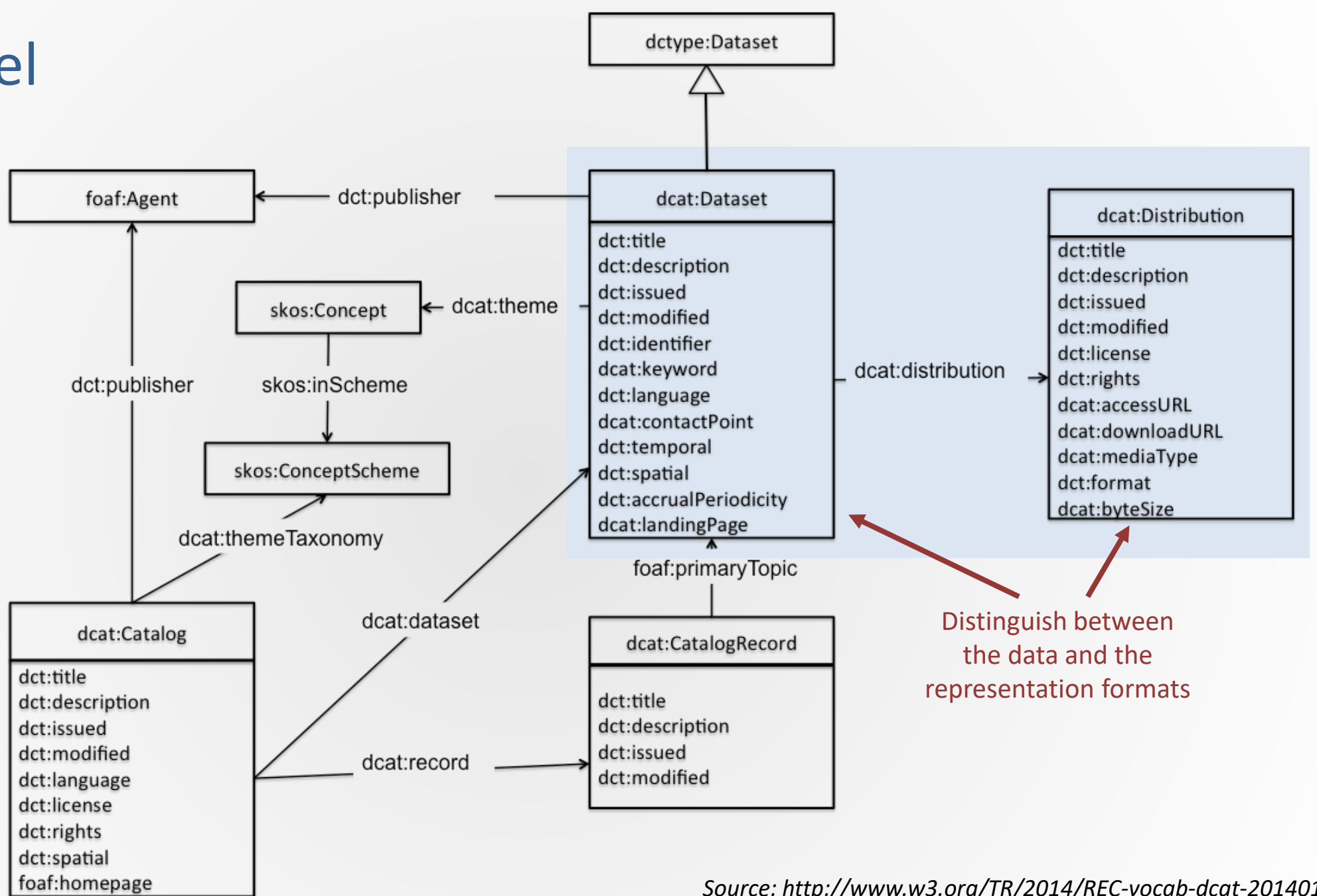
Describe any dataset, in any format, standalone or in a catalog

Describe any catalog of datasets

Facilitate interoperability between catalogs published on the web

<http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/>

DCAT model



Source: <http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/>

DCAT limitations: partial metadata coverage

- ✘ No versioning
- ✘ No properties for provenance, organizations, projects, funding, interpretation...

Yes, but DCAT is just a framework:

“Other complementary vocabularies may be used (...) to provide more detailed format-specific information”

(from DCAT recommandation)

DCAT is extensible with *Application Profiles*

- DCAT-AP: DCAT profile for data portals in Europe

“(...) enable cross-data portal search of data sets and make public sector data better searchable across borders and sectors”

(<https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe>)

- Europe: DCAT-AP-IT, DCAT-AP-NO, DCAT-AP.de, TransportDCAT-AP, StatDCAT-AP
- RING DCAT extension: agriculture, agrifood
- BotDCAT-AP: datasets for Chatbot Systems
- ...

At DCAT 1.1 will fix many limitations (2nd Q 2019)

Draws on the experience of DCAT,
the multiple use cases and Application Profiles

Multiple extensions:

- Versioning
- Qualitative information
- Formal description (coverage, interpretation...)
- ...

<https://w3c.github.io/dxwg/ucr/>



Collaborative community project led by:

Define a common vocabulary to **markup web pages**

- Make data understandable to search engines
- Improve discoverability, ranking
- Lightweight descriptions to allow for summarizations

Markup formats



RDFa



Microdata

Dataset

Canonical URL: <http://schema.org/Dataset>

[Thing](#) > [CreativeWork](#) > [Dataset](#)

A body of structured information describing some topic(s) of interest.

Property	Expected Type	Description
Properties from Dataset		
distribution	DataDownload	A downloadable form of this dataset, at a specific location, in a specific format.
includedInDataCatalog	DataCatalog	A data catalog which contains this dataset. Supersedes catalog , includedDataCatalog . Inverse property: dataset .
issn	Text	The International Standard Serial Number (ISSN) that identifies this serial publication. You can repeat this property to identify different formats of, or the linking ISSN (ISSN-L) for, this serial publication.
measurementTechnique	Text or URL	A technique or technology used in a Dataset (or DataDownload), such as "spectroscopy" or "colorimetry" or "immunofluorescence".

Not as comprehensive as DCAT, but benefits from large adoption of schema.org

schema.org
Home Schemas Documentation

DataCatalog

Canonical URL: <http://schema.org/DataCatalog>

[Thing](#) > [CreativeWork](#) > [DataCatalog](#)

A collection of datasets.

Property	Expected Type	Description
Properties from DataCatalog		
dataset	Dataset	A dataset contained in this catalog. Inverse property: includedInDataCatalog .
measurementTechnique	Text or URL	A technique or technology used in a Dataset (or DataDownload), such as "spectroscopy" or "colorimetry" or "immunofluorescence".

VOID: Vocabulary of Interlinked Datasets

*“VOID is an RDF Schema vocabulary for expressing metadata **about RDF datasets**”*

(from VOID recommendation)

General metadata

Terms from Dublin Core, FOAF: topics, web page, contacts, license, ~~version~~...

Access metadata

Data dump, RDF serialization formats, SPARQL endpoint, example

Structural metadata

URIs scheme, vocabularies, statistics (no. triples, properties, classes, subjects, objects), dataset partitions, root resource

Links from/to other datasets

No. link triples, link predicates (e.g. owl:sameAs, skos:exactMatch, owl:equivalentClass)

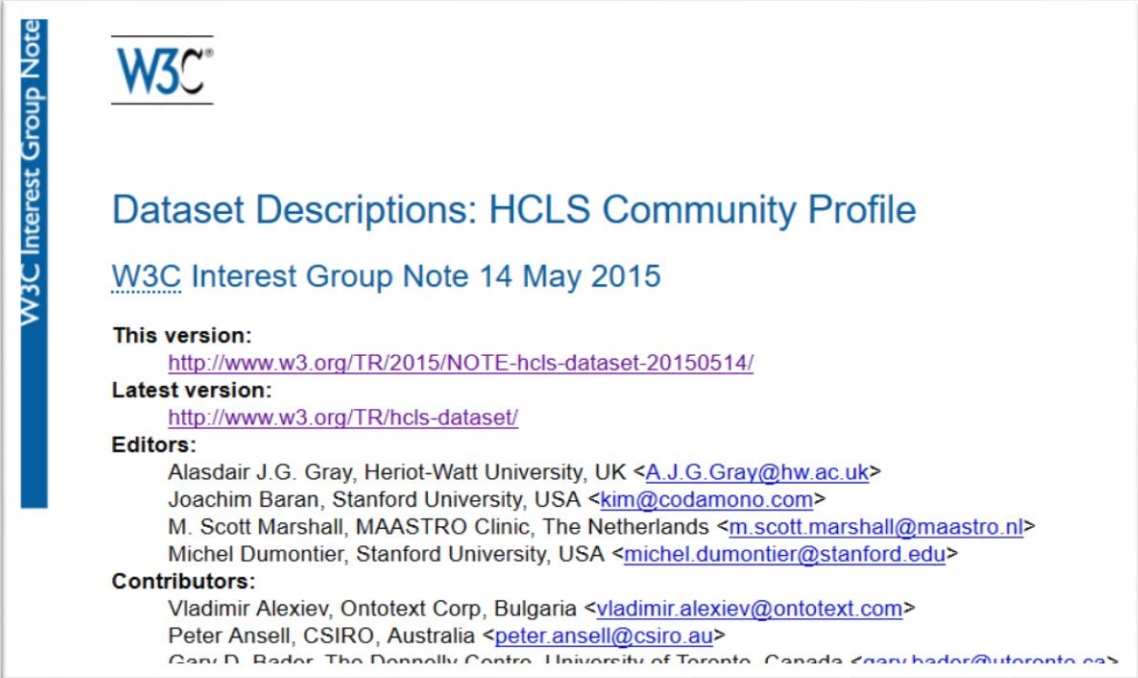
Health Care and Life Sciences Profile (HCLS)

“Develop a guidance note for reusing existing vocabularies to describe datasets with RDF”

(Michel Dumontier, <https://www.slideshare.net/micheldumontier/w3c-hcls-dataset-description>)

Consensus among participating stakeholders on the RDF description of HC and LS datasets

Detailed requirements have been drawn from a wide range of use cases



The image shows a slide from a W3C Interest Group Note. On the left side, there is a vertical blue bar with the text 'W3C Interest Group Note' written vertically. The main content of the slide is as follows:

W3C

Dataset Descriptions: HCLS Community Profile

W3C Interest Group Note 14 May 2015

This version:
<http://www.w3.org/TR/2015/NOTE-hcls-dataset-20150514/>

Latest version:
<http://www.w3.org/TR/hcls-dataset/>

Editors:
Alasdair J.G. Gray, Heriot-Watt University, UK <A.J.G.Gray@hw.ac.uk>
Joachim Baran, Stanford University, USA <kim@codamono.com>
M. Scott Marshall, MAASTRO Clinic, The Netherlands <m.scott.marshall@maastro.nl>
Michel Dumontier, Stanford University, USA <michel.dumontier@stanford.edu>

Contributors:
Vladimir Alexiev, Ontotext Corp, Bulgaria <vladimir.alexiev@ontotext.com>
Peter Ansell, CSIRO, Australia <peter.ansell@csiro.au>
Gary D. Bader, The Donnelly Centre, University of Toronto, Canada <gary.bader@utoronto.ca>

Vocabularies used in HCLS

Data Catalog (DCAT)

Citation Typing Ontology

Dublin Core Types/Terms

Friend-of-a-Friend (FOAF)

Collection Description Frequency Vocabulary

Identifiers.org vocabulary (IdoT)

Lexical Vocabulary (Lexvo)

Provenance Authoring and Versioning ontology (PAV)

PROV Ontology

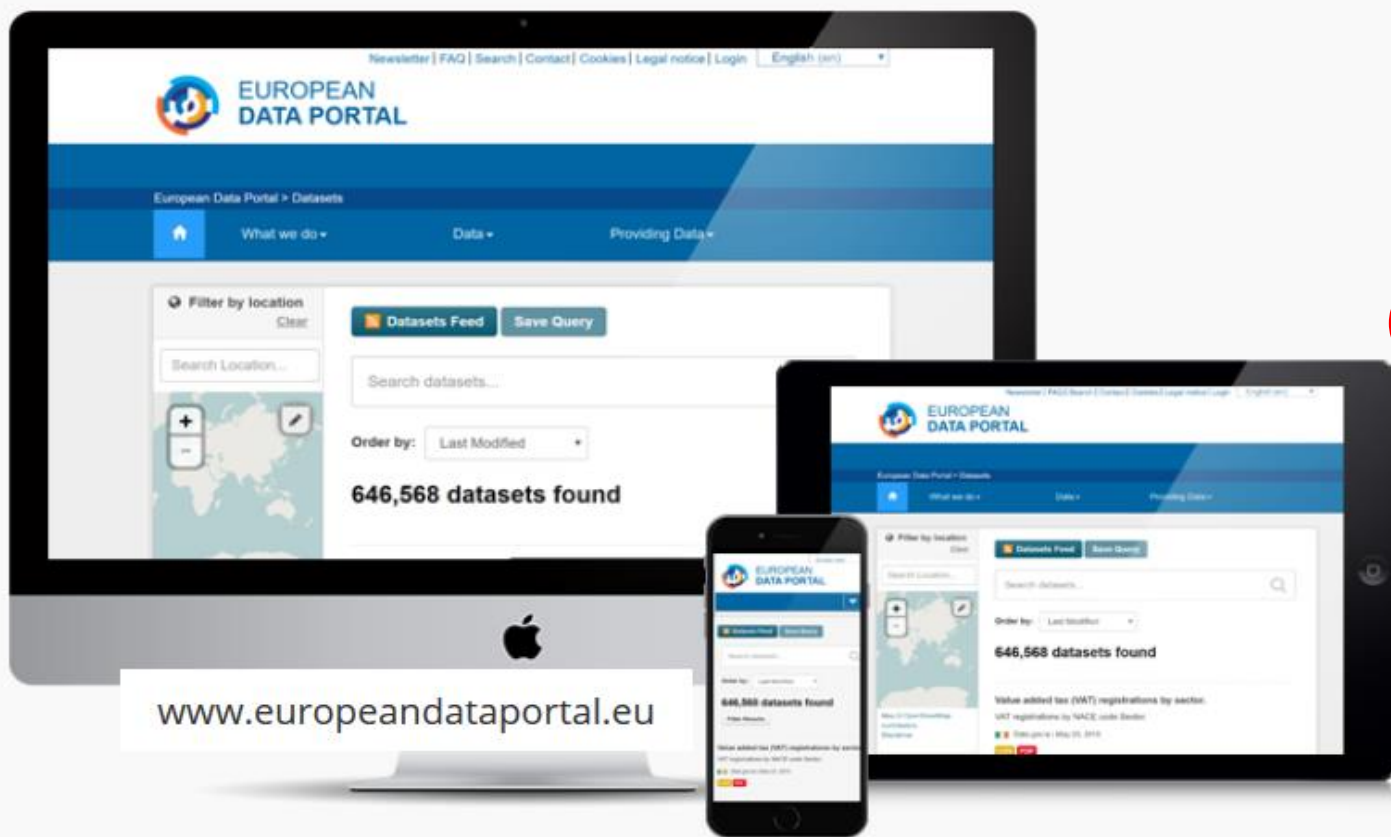
Schema.org

SPARQL Service Description (SD)

Semantic science Integrated Ontology (SIO)

Vocabulary of Interlinked Datasets (VOID)

Organize, host, search datasets with data portals



CKAN, the world's leading Open Source data portal platform

CKAN is a powerful data management system that makes data accessible – by providing tools to streamline publishing, sharing, finding and using data.

WHAT IS CKAN?

CKAN is aimed at data publishers (national and regional governments, companies and organizations) wanting to make their data open and available. [Learn more](#)

WHY CKAN?

CKAN is open source, free software. This means that you can use it without any license fees, and you retain all rights to the data and metadata you enter.

HOW TO CONTRIBUTE?

CKAN version releases are coordinated, tested and deployed by the tech team. Being an open source project, CKAN and its extensions are developed by a large community of people.

CKAN Data Management Platform

- Functions: allow to publish, share, search datasets
- Faceted search by topics, publishers, keywords, formats, languages...
- 197 instances as of Oct. 2018 (<https://ckan.org/about/instances/>)
 - 93 in Europe, 39 national government instances
 - data.gov.*
- Native exporting of records in DCAT and harvesting DCAT records from other catalogues

“Powered by CKAN”

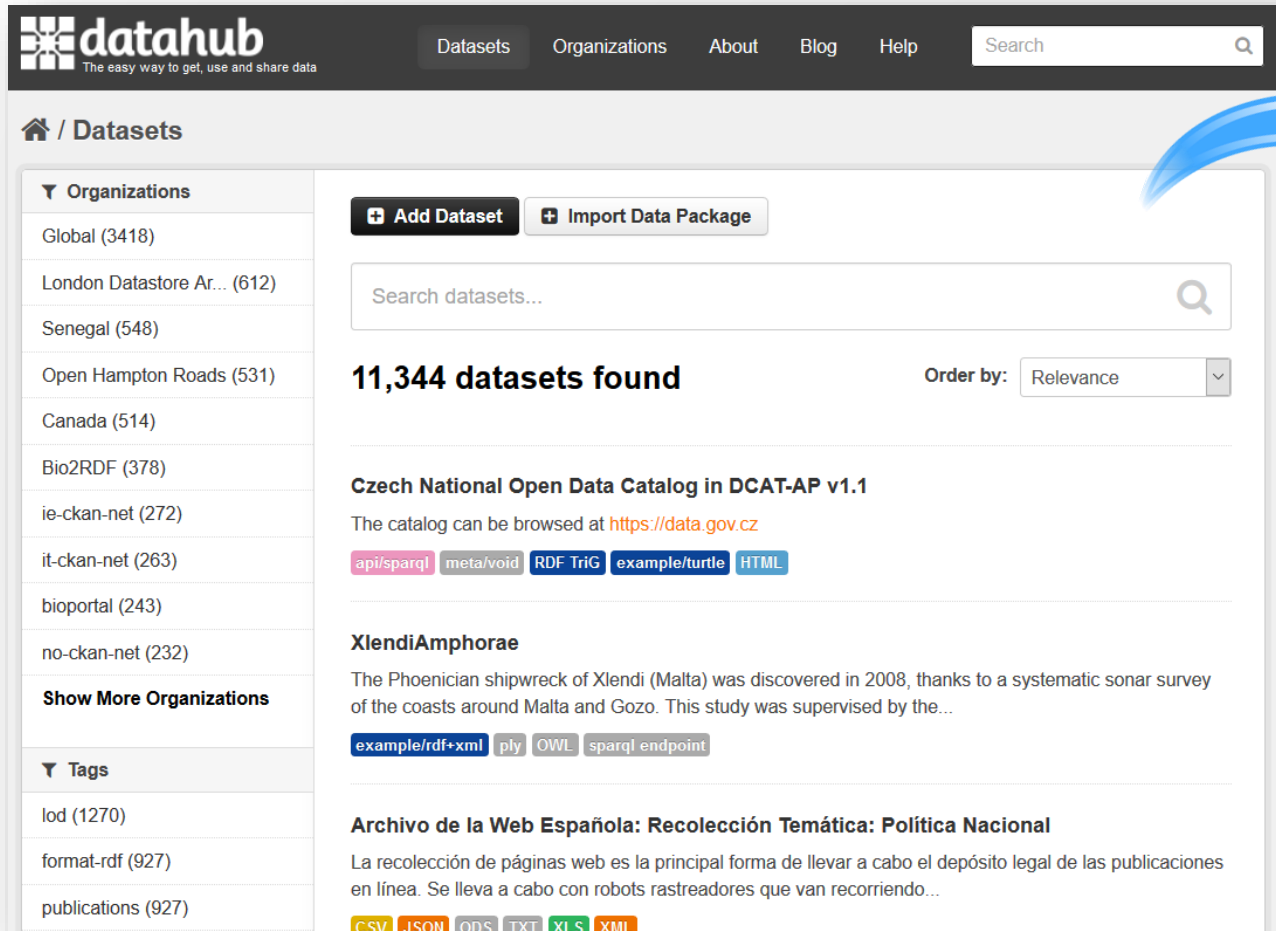
https://www.europeandataportal.eu/

The screenshot shows the European Data Portal interface. At the top, there is a search bar and navigation tabs for 'Home', 'Data', 'Applications', 'Linked Data', 'Visualisation Catalog', 'Developers' corner', and 'About'. Below the navigation, there is a 'Filter by location' section with a world map and a search bar for location. The main content area displays '840,874 datasets found' and a list of datasets, including 'Katastrální mapa pro katastrální území - Nový' and 'Cadastral map for cadastral areas — Hrazín a'. The interface is clean and modern, with a blue and white color scheme.

The screenshot shows the EU Open Data Portal interface. At the top, there is a search bar and navigation tabs for 'Home', 'Data', 'Applications', 'Linked Data', 'Visualisation Catalog', 'Developers' corner', and 'About'. Below the navigation, there is a 'Search datasets...' section with a search bar and a 'Suggest a dataset' section. The main content area displays 'Total datasets available: 12636' and a 'Most viewed datasets' section. The interface is clean and modern, with a blue and white color scheme.

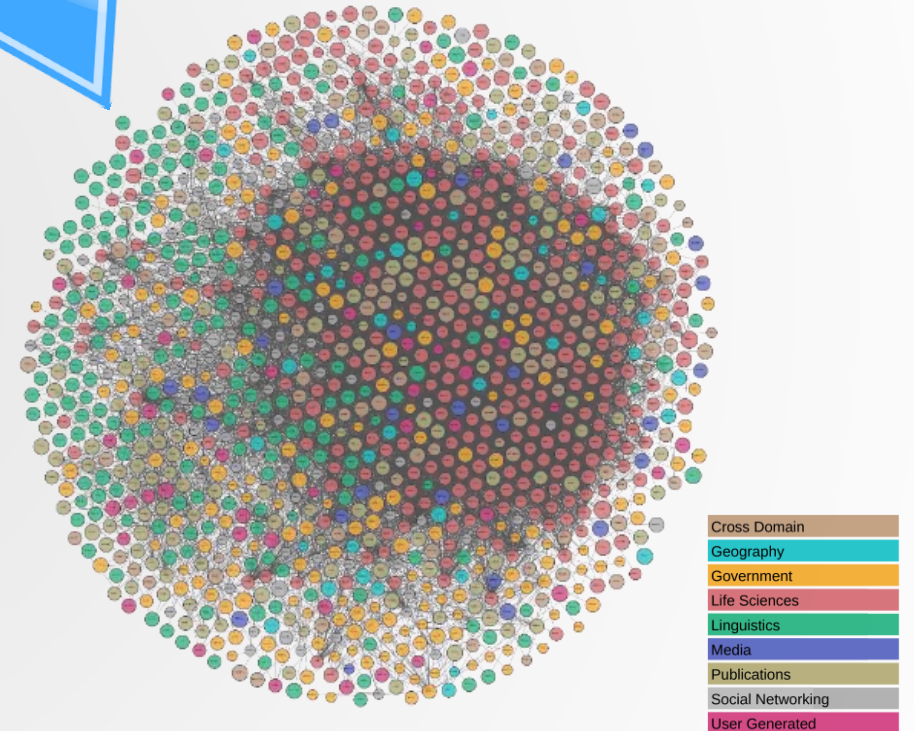
The screenshot shows the data.gouv.fr interface. At the top, there is a search bar and navigation tabs for 'Découvrez l'OpenData', 'Données', and 'Tableau de bord'. Below the navigation, there is a 'Recherche' section with a search bar and a 'Thématiques' dropdown menu. The main content area displays a grid of statistics: '36 647 Jeux de données', '155 646 Ressources', '1 797 Réutilisations', '31 155 Utilisateurs', '1 848 Organisations', and '2 785 Discussions'. At the bottom, there are two line charts: 'Derniers jeux de données envoyés' and 'Dernières réutilisations envoyées'. The interface is clean and modern, with a blue and white color scheme.

“Powered by CKAN”



The screenshot shows the Datahub.io website interface. At the top, there is a navigation bar with the Datahub logo and the tagline "The easy way to get, use and share data". The main navigation menu includes "Datasets", "Organizations", "About", "Blog", and "Help". A search bar is located on the right side of the navigation bar. Below the navigation bar, the page title is "Home / Datasets". On the left side, there is a sidebar with "Organizations" and "Tags" sections. The "Organizations" section lists various organizations with their respective dataset counts. The "Tags" section lists various tags with their respective dataset counts. The main content area features a search bar with the text "Search datasets...". Below the search bar, it displays "11,344 datasets found" and an "Order by: Relevance" dropdown menu. The first dataset listed is "Czech National Open Data Catalog in DCAT-AP v1.1", which includes a description, a URL, and several format options. The second dataset listed is "XlendiAmphorae", which includes a description and several format options. The third dataset listed is "Archivo de la Web Española: Recolección Temática: Política Nacional", which includes a description and several format options.

Datahub.io: Source for the Linked Open Data cloud



Linking Open Data cloud diagram, 2018. J.P. McCrae, A. Abele, P. Buitelaar, A. Jentzsch, V. Andryushechkin and R. Cyganiak. <http://lod-cloud.net/>

SEARCH

Making it easier to discover datasets

Natasha NoyResearch Scientist, Google
AI


Published Sep 5, 2018

In today's world, scientists in many disciplines and a growing number of journalists live and breathe data. There are many thousands of data repositories on the web, providing access to millions of datasets; and local and national governments around the world publish their data as well. To enable easy access to this data, we launched [Dataset Search](#), so that scientists, data journalists, data geeks, or anyone else can [find the data](#) required for their work and their stories, or simply to satisfy their intellectual curiosity.

Similar to how [Google Scholar](#) works, Dataset Search lets you find datasets wherever they're hosted, whether it's a publisher's site, a digital library, or an author's personal web page. To create Dataset search, we developed [guidelines for dataset providers](#) to describe their data in a way that Google (and other search engines) can better understand the content of their pages. These guidelines include salient information about datasets: who created the dataset, when it was



Google Dataset Search Bêta

Essayer [boston education data](#) ou [weather site:noaa.gov](#)



INPN - Référentiel taxonomique TAXREF

www.data.gouv.fr
data.wu.ac.at

Dernière mise à jour : 31 janv. 2017



TAXREF-LD: Linked Data French Taxonomic Register

data.wu.ac.at
datahub.ckan.io

Dernière mise à jour : Jun 21, 2018



Inventaire de la faune

next.data.gouv.fr
opendata.hauts-de-seine.fr
+3plus

Dernière mise à jour : 6 avr. 2017



INPN - Espèces protégées et réglementées

www.data.gouv.fr
data.wu.ac.at

Dernière mise à jour : 13 avr. 2016



TAXREF-LD: Linked Data French Taxonomic Register



data.wu.ac.at



datahub.ckan.io

Ensemble de données mis à jour le Jun 21, 2018

Ensemble de données publié le Jan 27, 2017

Ensemble de données fourni par

French Muséum National d'Histoire Naturelle

Licence

<http://creativecommons.org/licenses/by-nc/2.0/>

Formats de téléchargement disponibles auprès des fournisseurs

SPARQL , DCAT , VOID , HTML , TURTLE

Description

TAXREF-LD is the Linked Data representation of TAXREF (<https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref?lg=en>), the French national taxonomical register for fauna, flora and fungus, that covers mainland France and overseas territories. It accounts for over 500000 scientific names.

TAXREF-LD is a joint initiative of the National Museum of Natural History (<http://www.mnhn.fr/>) and the I3S laboratory (<http://www.i3s.unice.fr/>) (Université Côte d'Azur, CNRS, Inria). Its model is described in [1].

[1] Michel F., Gargominy O., Terceire S. & Faron-Zucker C. (2017). A Model to Represent Nomenclatural and Taxonomic Information as Linked Data. Application to the French Taxonomic Register TAXREF. In *Proceedings of the 2nd International Workshop on Semantics for Biodiversity*

- Introduction
- Structured data**
 - About Search features
 - Search feature gallery
 - Introduction to structured data
 - Enhance your site's attributes
 - Mark up your content items
 - Build, test, & release structured data
 - Structured data general guidelines
 - Feature guides
 - Enhancements**
 - Breadcrumb
 - Sitelinks searchbox
 - Corporate contact
 - Logo
 - Social profile
 - Carousel
 - Content Types**
 - Article
 - Book

tool for processing

- Images capturing data
- Files relating to machine learning, such as trained parameters or neural network structure definitions
- Anything that looks like a dataset to you

Our approach to dataset discovery

We can understand structured data in Web pages about datasets, using either [schema.org Dataset markup](#) [↗](#), or equivalent structures represented in [W3C's Data Catalog Vocabulary \(DCAT\) format](#) [↗](#). We also exploring [experimental support for structured data based on W3C CSVW](#) [↗](#), and expect to evolve and adapt our approach as best practices for dataset description emerge. For more information about our approach to dataset discovery, see [Facilitating the discovery of public datasets](#) [↗](#).

Examples

Here's an example for datasets using JSON-LD syntax (preferred) in the Structured Data Testing Tool. The same vocabulary can also be used in RDFa 1.1, Microdata, or W3C DCAT vocabulary. The following example is based on a [real-world dataset description](#) [↗](#).

JSON-LD

RDFA

Sommaire

Our approach to dataset discovery

Examples

Guidelines

Sitemap best practices

Source and provenance best practices

Known Errors and Warnings

Structured data type definitions

Dataset

DataCatalog

DataDownload

Tabular datasets

Help and tools

TAXREF-LD: Linked Data French Taxonomic Register

Followers

0

Follow

Organization



French Muséum National d'Histoire Naturelle

Dataset

Groups

Activity Stream

Manage

TAXREF-LD: Linked Data French Taxonomic Register

TAXREF-LD is the Linked Data representation of TAXREF (<https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref?lg=en>), the French national taxonomical register for fauna, flora and fungus, that covers mainland France and overseas territories. It accounts for over 500000 scientific names.

TAXREF-LD is a joint initiative of the National Museum of Natural History (<http://www.mnhn.fr>) and the I3S laboratory (<http://www.i3s.unice.fr/>) (Université Côte d'Azur, CNRS, Inria). Its model is described in [1].

[1] Michel F., Gargominy O., Terceire S. & Faron-Zucker C. (2017). A Model to Represent Nomenclatural and Taxonomic Information as Linked Data. Application to the French Taxonomic Information. *Proceedings of the 2nd International Workshop on Semantics for Biodiversity (S4Bi) ISWC 2017* vol. 1933. Vienna, Austria. CEUR.

Data and Resources

- Turtle Example**
Link to the RDF Turtle description of species Delphinus [More information](#)
- SPARQL endpoint**
SPARQL endpoint [More information](#)

Makup embedded in the web page as JSON-LD



```

@id:
  "https://datahub.ckan.io/dataset/709597bd-9fcf-4c06-8845-2c74c51000c2"
@type:
  "schema:Dataset"
schema:dateModified:
  "2018-06-21T08:26:23.362605"
schema:datePublished:
  "2017-01-27T09:10:18.569429"
schema:description:
  "TAXREF-LD is the Linked Data representation of TAXREF (https://inpn.mnhn.fr/programme/referentiel-taxonomique-taxref?lg=en), the French national taxonomical register for mainland France and overseas territories. It accounts for over 500000 a joint initiative of the National Museum of Natural History (http://www.i3s.unice.fr/) (Université Côte d'Azur, CNRS, Inria). Its model is described in [1]. Its model is described in [1]. Michel F., Gargominy O., Terceire S. & Faron-Zucker C. (2017). A Model to Represent Nomenclatural and Taxonomic Information as Linked Data. Application to the French Taxonomic Information. Proceedings of the 2nd International Workshop on Semantics for Biodiversity (S4Bi) ISWC 2017 vol. 1933. Vienna, Austria. CEUR."
schema:distribution:
  0:
    @id:
      "https://datahub.ckan.io/dataset/709597bd-9fcf-4c06-8845-2c74c51000c2b48d5da7e25d"
  1:
    @id:
      "https://datahub.ckan.io/dataset/709597bd-9fcf-4c06-8845-2c74c51000c2e31107d48572"
  2:
    @id:
      "https://datahub.ckan.io/dataset/709597bd-9fcf-4c06-8845-2c74c51000c2
  
```

Thank you!

