

# FAIR data management and AI in research data infrastructures: Semantic Croissant and CDIF

Slava Tykhonov

Head of AI and Interoperability at CODATA

Harvard Dataverse Ambassador

Keynote talk for the Nordic-Baltic Dataverse Hub (NAISH) project kick-off

Kaunas, Lithuania

21 April 2026

## Intro: AI assistant to “collect” and deposit data

### Pale Fire Footnotes

Read [https://docs.google.com/spreadsheets/d/1l15wMlgKUR\\_DkyXiJYSU0xeHTRxcZwKp/edit?gid=250921036#gid=250921036](https://docs.google.com/spreadsheets/d/1l15wMlgKUR_DkyXiJYSU0xeHTRxcZwKp/edit?gid=250921036#gid=250921036) and deposit first 3 rows as datasets in Dataverse.

Ask Pale Fire Footnotes...

Created by Slava Tykhonov (CODATA) | Powered by Gemini CLI

### UC2 - OPENHIDRA - Dataset and Ser...

File Edit View Insert Format Data Tools Help

100% 123 Calibri 11

	A	B	C	D
1	Domain	CLIMATE-ADAPT 4 EOSC Info		
2	Title	Partner	Use Case	Service
3	Description	Indicate the parter responsible for the use case	Location used for Use Case	Indicate the service currently uses the d
4	E.g.	LNEC	OPENHidra	OPENCoasts
5	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPEN
6	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPEN
7	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPEN
8	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPEN
9	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPEN
10	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPEN
11	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPEN
12	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPEN
13	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPEN
14	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPEN
15	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPEN

Sheet1

# Intro: AI Assistant job as data depositor

## Pale Fire Footnotes



Pale Fire Footnotes

<< 144 145 146 147 148 149 150 151 152

Analyze page

Connected

Read [https://docs.google.com/spreadsheets/d/1i15wMlgKUR\\_DkyXiJYSU0xeHTRxcZwKp/edit?gid=250921036#gid=250921036](https://docs.google.com/spreadsheets/d/1i15wMlgKUR_DkyXiJYSU0xeHTRxcZwKp/edit?gid=250921036#gid=250921036) and deposit first 3 rows as datasets in Dataverse.

The first three datasets from the spreadsheet have been successfully deposited and published in Dataverse.

### Deposited Datasets

**Aveiro/OPENHIDRA - Sea level (m)**

PID: doi:10.70122/FK2/RE2TCX

Description: Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological.

**Aveiro/OPENHIDRA - Salinity (PSU - Practical Salinity Unit)**

PID: doi:10.70122/FK2/DV7Q4B

Description: Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological.

**Aveiro/OPENHIDRA - Temperature (degrees C)**

PID: doi:10.70122/FK2/FUU9N9

Description: Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological.

All datasets were created in the slava collection on [demo.dataverse.org](https://demo.dataverse.org) with a CCO 1.0 license.

Ask Pale Fire Footnotes...



Created by Slava Tykhonov (CODATA) | Powered by Gemini CLI

Dataverse



Slava Tykhonov Dataverse

Demo Dataverse >

Contact Share Edit

Search this dataverse...



Advanced Search

+ Add Data

Dataverses (0)

Datasets (149)

Files (139)

Publication Year

2026 (116)

Publication Status

Published (110)

Draft (39)

Unpublished (33)

Author Name

CDIF Expert (49)

Pale Fire Agent (37)

Copernicus (EU) (30)

IPMA (7)

NOAA (7)

More...

Author Affiliation

CODATA (2)

Subject

Other (144)

Social Sciences (11)

Earth and Environmental Sciences (5)

Keyword Term

1 to 10 of 149 Results

Sort

Aveiro/OPENHIDRA - Temperature (degrees C)



Mar 17, 2026

Copernicus (EU), 2026, "Aveiro/OPENHIDRA - Temperature (degrees C)", <https://doi.org/10.70122/FK2/FUU9N9>, Demo Dataverse, V1

Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological. Original Identifier: [yeshttps://doi.org/10.48670/moi-00016](https://doi.org/10.48670/moi-00016).

Aveiro/OPENHIDRA - Salinity (PSU - Practical Salinity Unit)



Mar 17, 2026

Copernicus (EU), 2026, "Aveiro/OPENHIDRA - Salinity (PSU - Practical Salinity Unit)", <https://doi.org/10.70122/FK2/DV7Q4B>, Demo Dataverse, V1

Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological. Original Identifier: [yeshttps://doi.org/10.48670/moi-00016](https://doi.org/10.48670/moi-00016).

Aveiro/OPENHIDRA - Sea level (m)



Mar 17, 2026

Copernicus (EU), 2026, "Aveiro/OPENHIDRA - Sea level (m)", <https://doi.org/10.70122/FK2/RE2TCX>, Demo Dataverse, V1

Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological. Original Identifier: [yeshttps://doi.org/10.48670/moi-00016](https://doi.org/10.48670/moi-00016).

CDIF Dataset: Aveiro/OPENHIDRA - air pressure (Pa)




Mar 13, 2026

Copernicus (EU), 2026, "CDIF Dataset: Aveiro/OPENHIDRA - air pressure (Pa)", <https://doi.org/10.70122/FK2/RE2TCX>, Demo Dataverse, V1

Feedback

DRA - air

# Intro: AI assistant is saving data in Dataverse


 **Dataverse**

[Add Data](#) [Search](#) [About](#) [User Guide](#) [Support](#) [Slava Tykhonov](#) **250**

[Demo Dataverse](#) > [Slava Tykhonov Dataverse](#) >

## Aveiro/OPENHIDRA - Temperature (degrees C)


Version 1.0

 Copernicus (EU), 2026, "Aveiro/OPENHIDRA - Temperature (degrees C)", <https://doi.org/10.70122/FK2/FUU9N9>, Demo Dataverse, V1

[Cite Dataset](#) [Learn about Data Citation Standards.](#)

**Description** Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological. Original Identifier: yeshttps://doi.org/10.48670/moi-00016.

**Subject** Earth and Environmental Sciences

**License/Data Use Agreement**  [CC0 1.0](#)

[Files](#) [Metadata](#) [Terms](#) [Versions](#)

[Add + Edit Metadata](#) [Export Metadata](#)

**Citation Metadata**

<b>Persistent Identifier</b>	doi:10.70122/FK2/FUU9N9
<b>Publication Date</b>	2026-03-17
<b>Title</b>	Aveiro/OPENHIDRA - Temperature (degrees C)
<b>Author</b>	Copernicus (EU)
<b>Point of Contact</b>	Use email button above to contact. Slava Tykhonov
<b>Description</b>	Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological. Original Identifier: yeshttps://doi.org/10.48670/moi-00016.
<b>Subject</b>	Earth and Environmental Sciences

## AI Agent job

1. Read input from any source
2. Analyze Web pages and find data
3. Extract dataset descriptions and form metadata record
4. Create Croissant export for Machine Learning tasks
5. Deposit metadata as dataset in Dataverse data platform
6. Share Croissant back to give AI Agent pointers to metadata and data files



# Intro: AI assistant for Kaunas weather

## Pale Fire Footnotes

Pale Fire Footnotes << 145 146 147 148 149 150 151 152 153 Analyze page Connected

Deposit in  
Dataverse

**Confirming Successful Publication** I've checked the status of the "AccuWeather Kaunas Forecast - April 21, 2026" dataset (PID: doi:10.70122/FK2/X4PQSQ ). It was successfully created, both files were uploaded without any issues, and the publication process was completed. This confirms that the dataset is now officially released and ready for public access. [Thought: true]I have successfully extracted, cached, and deposited the AccuWeather Kaunas Forecast - April 21, 2026 dataset into Dataverse.

**Dataset Title:** AccuWeather Kaunas Forecast - April 21, 2026

**Persistent Identifier (PID):** doi:10.70122/FK2/X4PQSQ

**Status:** RELEASED (Published)

**Files Deposited:**

cdif\_variables\_kaunas\_weather\_20260421.tab (CSV Variable Inventory)

cdif\_variables\_kaunas\_weather\_20260421.jsonld (CDIF Metadata)

The JSON-LD packages and CSV inventory have been successfully cached and deposited.


Do you want to see all variables in CDIF format from this page in a table?

Ask Pale Fire Footnotes...

A


Created by Slava Tykhonov (CODATA) | Powered by Gemini CLI







Address, City or Zip Code

TODAY HOURLY 10-DAY RADAR MINUTECAST\* MONTHLY AIR QUALITY HEALTH & ACTIVITY






38,29 €/mėn.





20,79 €/mėn.




-10%  
28,08 €

TODAY'S WEATHER TUE, APR 21

 Plenty of sunshine Hi: 14°  
 Tonight: Mainly clear and chilly Lo: 0°

CURRENT WEATHER 1:40 PM



11°C  
RealFeel® 11°

RealFeel Shade™ 7°

Wind NNE 20 km/h

Wind Gusts 20 km/h

# Intro: AI assistant for Kaunas weather in Dataverse

## Pale Fire Footnotes

Pale Fire Footnotes << 145 146 147 148 149 150 151 152 153 Analyze page Connected

Deposit in  
Dataverse

**Confirming Successful Publication** I've checked the status of the "AccuWeather Kaunas Forecast - April 21, 2026" dataset (PID: doi:10.70122/FK2/X4PQSQ ). It was successfully created, both files were uploaded without any issues, and the publication process was completed. This confirms that the dataset is now officially released and ready for public access. [Thought: true]I have successfully extracted, cached, and deposited the AccuWeather Kaunas Forecast - April 21, 2026 dataset into Dataverse.

**Dataset Title:** AccuWeather Kaunas Forecast - April 21, 2026

**Persistent Identifier (PID):** doi:10.70122/FK2/X4PQSQ

**Status:** RELEASED (Published)

**Files Deposited:**

cdif\_variables\_kaunas\_weather\_20260421.tab (CSV Variable Inventory)

cdif\_variables\_kaunas\_weather\_20260421.jsonld (CDIF Metadata)  
The JSON-LD packages and CSV inventory have been successfully cached and deposited.

Do you want to see all variables in CDIF format from this page in a table?

Ask Pale Fire Footnotes...



Dataverse



Access File ▾

Contact Owner

Share

File Metrics ⓘ

0 Downloads ⓘ

Preview

Metadata

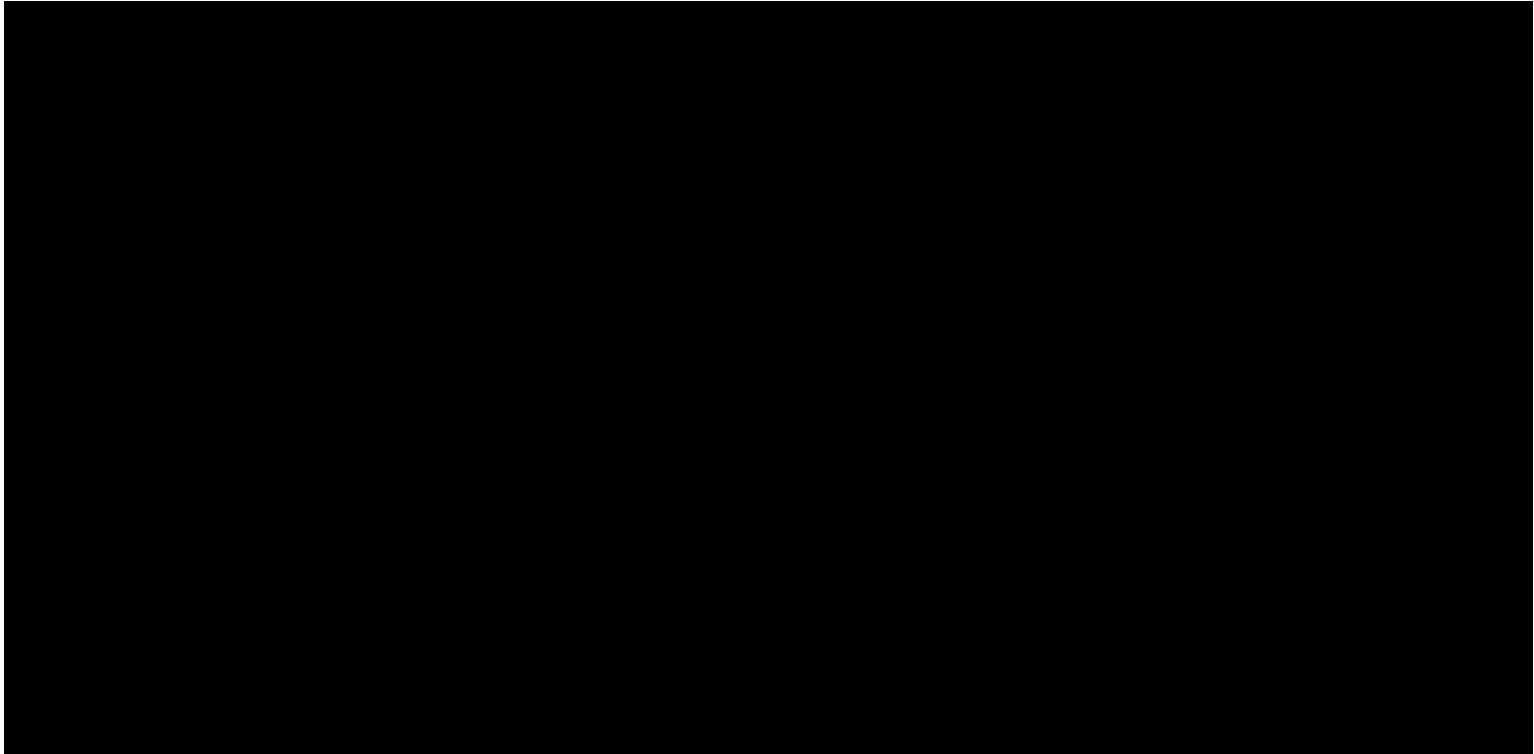
Versions

Open in New Window

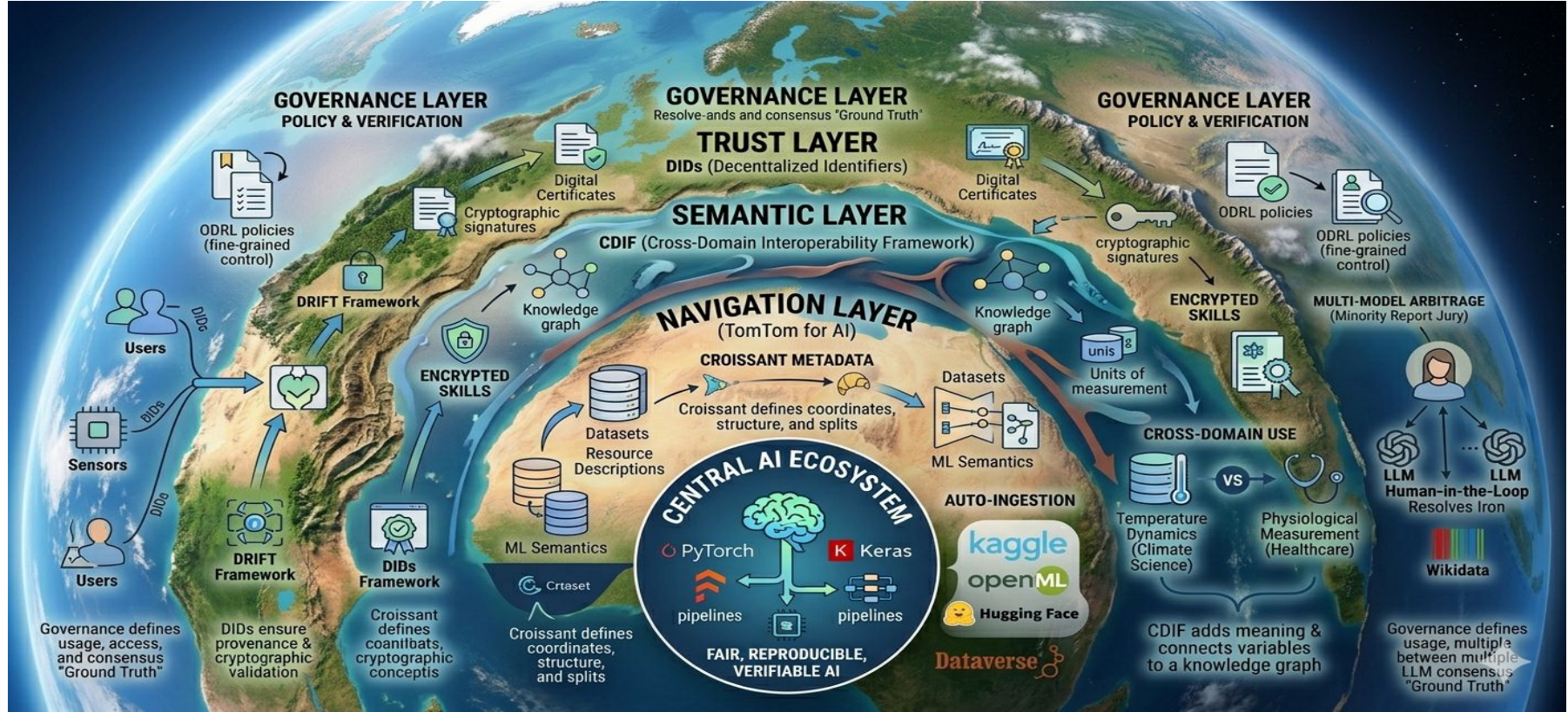
	Name	Value	Unit	
1	Current Temperature	11	°C	Measured at 1:40
2	RealFeel	11	°C	AccuWeather Re
3	RealFeel Shade	7	°C	Temperature felt
4	Today High Temperature	14	°C	Forecasted high
5	Today Low Temperature	0	°C	Forecasted low f
6	Wind Speed	20	km/h	Current wind spe
7	Wind Gusts	20	km/h	Maximum wind c
8	Air Quality	Fair	N/A	Generally accept
9	Daylight Duration	14:31	hours:minutes	Time between su
10	Sunrise	6:08	AM	Time of sunrise
11	Sunset	8:39	PM	Time of sunset
12	Moonrise	7:37	AM	Time of moonrise
13	Moonset	2:57	AM	Time of moonset
14	Precipitation Probability (current)	0	%	Forecasted chan
15	Precipitation Probability (max 10-day)	61	%	Highest chance c
16	Tree Pollen	Low	N/A	Allergy outlook fc

[Dataset reference](#)

What kind of infra you need to make model fully deterministic?  
CODATA Ollama inference engine with CDIF and ODRL extensions



# What is research data infrastructure?



## “Semantic” Croissant ecosystem for AI-Ready data landscape

1. Croissant (Croissant standard for Machine Learning v.1.1)  
Navigation layer for AI - find data (resources) and train new models - led by Omar Benjelloun (*Google, USA*) and Elena Simperl (*King's College London, UK*)
2. CDIF (Cross-Domain Interoperability Framework)  
Semantics for AI - knowledge graph to “understand” and transmit context (*CODATA, France*)
3. DID (Decentralized Identifiers) with digital certificates  
Trust for AI - to “sign” resources digitally - led by Christoph Fabianek (*Frequentis, Austria*)
4. ODRL (Open Digital Rights Language) policies  
Policies for AI - to manage permissions and make resources actionable - led by Darren Bell (*UKDS, UK*)
5. Model Context Protocol (MCP)  
Tools and services to connect AI to resources (*CODATA/Harvard University, USA*)



# AI-ready data - Croissant for Machine Learning spec (2024)

## Croissant Format Specification

Version 1.0

Published: 2024/03/01

<http://mlcommons.org/croissant/1.0>

Authors:

- Omar Benjloun (Google),
- Elena Simperl (King's College London & ODI),
- Pierre Marcenac (Google),
- Pierre Ruyssen (Google),
- Costanza Conforti (Google),
- Michael Kuchnik (Meta),
- Jos van der Velde (Open ML),
- Luis Oala (Dotphoton),
- Steffen Vogler (Bayer),
- Mubashara Akthar (King's College London),
- Nitisha Jain (King's College London),
- Slava Tykhonov (DANS-KNAW)

## Introduction

Datasets are the basis of machine learning (ML). However, a lack of standardization in the description and semantics of ML datasets has made it increasingly difficult for researchers and practitioners to explore, understand, and use all but a small fraction of popular datasets.

The Croissant metadata format simplifies how data is used by ML models. It provides a vocabulary for dataset attributes, streamlining how data is loaded across ML frameworks such as PyTorch, TensorFlow or JAX. In doing so, Croissant enables the interchange of datasets between ML frameworks and beyond, tackling a variety of **discoverability**, **portability**, **reproducibility**, and **responsible AI (RAI)** challenges.

<https://mlcommons.github.io/croissant/docs/croissant-spec.html>

The screenshot displays the OpenML data platform interface. At the top, there's a search bar and a 'New Notebook' button. Below this, the 'Global News Engagement on Social Media' dataset is featured, with a 'Download (267 KB)' button and a 'Learn more about Croissant at ML Commons' link. The main section shows the 'monks-problems-2' dataset, which is a verified, public dataset from 2014. It includes a description of the dataset, its source, and a link to the 'monks-problems-2' dataset page. A white arrow points to the 'Download (267 KB)' button.

# Croissant Motivation: helping ML users with datasets

Most Machine Learning datasets are uniquely structured and require specialized handling.

Users spend a lot of efforts on data-related tasks.

What makes Datasets used in Machine Learning special?

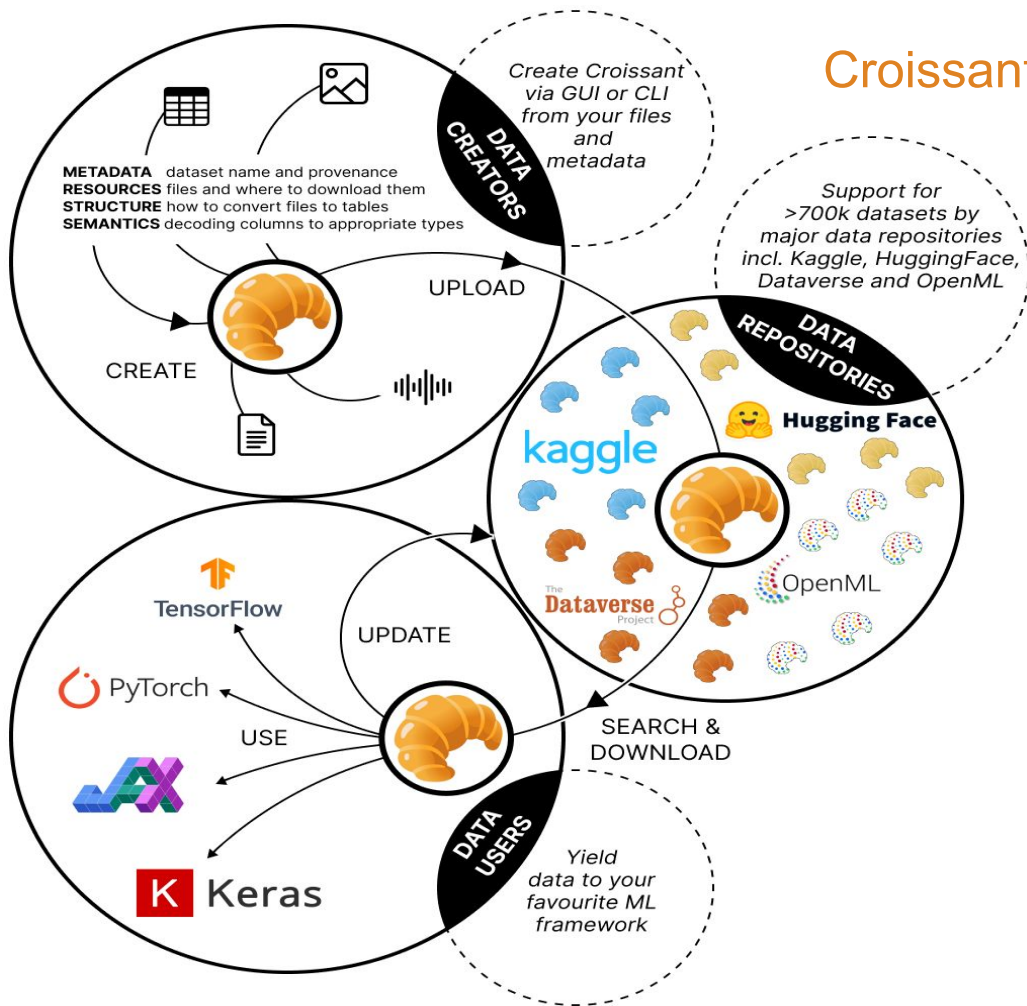
- Often combine **unstructured** (text, image, video) and **structured** (tabular, json) data
- Need to be "**flattened**" / **denormalized** to be used in ML frameworks and tools
- Need **ML-specific metadata**  
(e.g., Responsible AI info, test/train/validation splits, labels)
- Require **versioning** / **checkpointing** to support model snapshots and reproducibility

Credits: Croissant working group



scan to access slides and links

# Croissant for ML: state of art



## Create

- Editor
  - <https://huggingface.co/spaces/MLCommons/croissant-editor>
- Platform auto generate
  - <https://huggingface.co/datasets>
  - <https://www.kaggle.com/datasets>

## Discover and find

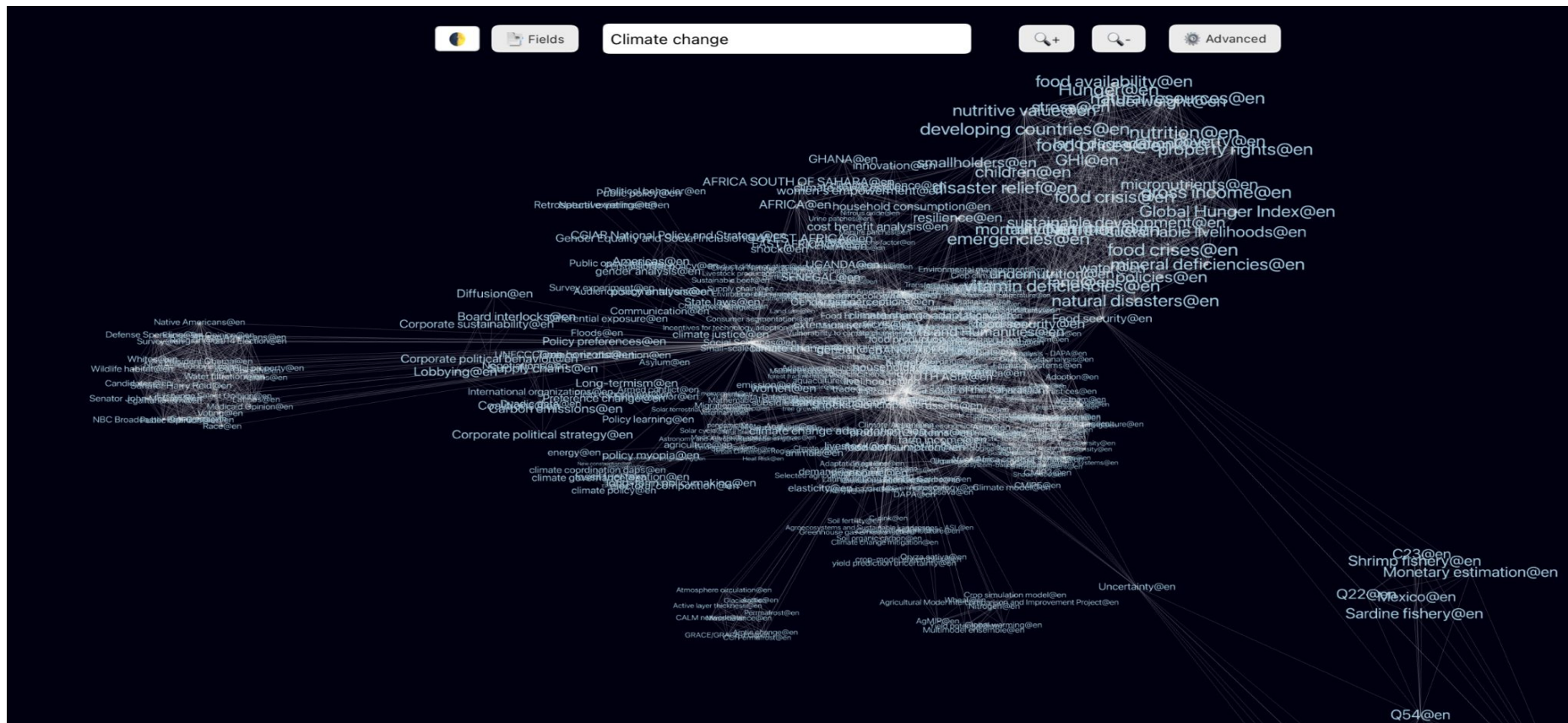
- Google Dataset Search
  - <https://datasetsearch.research.google.com/>
- Crawler in repo
  - [https://github.com/mlcommons/croissant/blob/main/health/visualizer/report\\_huggingface.ipynb](https://github.com/mlcommons/croissant/blob/main/health/visualizer/report_huggingface.ipynb)

## Use it

- Colab
  - [https://github.com/mlcommons/croissant/blob/main/python/mlcroissant/recipes/tfds\\_croissant\\_builder.ipynb](https://github.com/mlcommons/croissant/blob/main/python/mlcroissant/recipes/tfds_croissant_builder.ipynb)
  - Original
    - [https://github.com/mlcommons/croissant/blob/main/python/mlcroissant/recipes/tfds\\_croissant\\_builder.ipynb](https://github.com/mlcommons/croissant/blob/main/python/mlcroissant/recipes/tfds_croissant_builder.ipynb)



# Croissant Graph in Dataverse network



55 million Croissant triples in Qlever triple store.

# Croissant layers in detail

- **Dataset-level metadata**
  - Based on [schema.org/Dataset](https://schema.org/Dataset)
  - Best practices for required fields, licence choice, etc.
- **Resource description**
  - Flexible data access schemes (files, archives, local / remote directories)
  - Support commonly used file formats (text, images, video, CSV, JSON, etc.)
  - Allow for fine-grained versioning / checkpointing, and resource verification via checksums
- **Content structure**
  - Agnostic to specific file formats
  - Describe structure of tabular and nested data
  - Expressive data type system with support for common semantic types
  - "Join" across structured and unstructured data
  - Define "flattened" / denormalized views that are suitable for ML applications
- **ML Semantics**
  - Mechanisms for data-driven Responsible AI
  - Describe and link ML-specific concepts, e.g. **labels, variables, training/test splits**

## Example for tabular content (CSV table)

```
{
  "identifier": "movies",
  "@type": "ml:RecordSet",
  "source": "#{movies-table}",
  "key": "#{movie_id}",
  "field": [
    {
      "name": "movie_id",
      "@type": "ml:Field",
      "dataType": "sc:Integer",
      "source":
        "#{movies-table/movieId}"
    },
    {
      "name": "title",
      "@type": "ml:Field",
      "dataType": "sc:Text",
      "source":
        "#{movies-table/title}"
    },
  ],
}
```

```
{
  "name": "genre",
  "@type": "ml:Field",
  "dataType": "sc:Text",
  "repeated": "true",
  "source":
    {
      "data": "#{movies-table/genres}",
      "applyTransform": {"separator": "|"}
    }
}
]
```

movieId,title,genres
1,Toy Story (1995),Adventure Animation Children Comedy Fantasy
2,Jumanji (1995),Adventure Children Fantasy
3,Grumpier Old Men (1995),Comedy Romance
4,Waiting to Exhale (1995),Comedy Drama Romance

## ML semantics: Splits

```
{
  "name": "split",
  "@type": "ml:Field",
  "dataType": [
    "sc:Text",
    "ml:Split"
  ],
  "source": {
    "data":
      "#{caption_annotations-files/filename}",
    "applyTransform": {
      "regex": ".*_(val|train)2014\\.json$"
    }
  },
  "references": "#{split_enums/name}"
}
```

name	url
train	<a href="https://mlcommons.org/definitions/training_split">https://mlcommons.org/definitions/training_split</a>
val	<a href="https://mlcommons.org/definitions/validation_split">https://mlcommons.org/definitions/validation_split</a>
test	<a href="https://mlcommons.org/definitions/test_split">https://mlcommons.org/definitions/test_split</a>

# Semantic interoperability on the level of infrastructure

We envision a situation where thousands of data nodes (Dataverse, Kaggle, HuggingFace, OpenML, etc) on the web can be simultaneously queried for dataset and will form shared Data Lake.

The *old dream* of Federated search/Universal catalogue can only be realised if:

- (1) Crosswalks; mapping across different metadata schemes are implemented
- (2) In metadata schemes we seek for ways to enrich indexes with values from controlled vocabularies

Standard response (**centralized**) = standardisation and harmonisation = repository software, certain metadata standards, or certain controlled vocabularies

New response (**distributed**) = explore agile solutions (Proof of Concepts) which can be implemented by different communities (even smaller ones), so we keep variety and still enable integration in the Distributed Data Network by applying Linked Data technologies based on the common metadata exchange standard (Croissant ML). AI is important part of this response.



# Interoperability Frameworks

Among the most important, but most challenging, recommendations of the **Turning FAIR into Reality** report, is R.4:

‘Develop **interoperability frameworks** for FAIR sharing within disciplines and for interdisciplinary research: Research communities need to be supported to develop interoperability frameworks that define their practices for data sharing, data formats, metadata standards, tools and infrastructure. **To support interdisciplinary research, these interoperability frameworks should be articulated in common ways and adopt global standards where relevant.**’

Influential notion of ‘Interoperability Frameworks’.

Led directly to the EOSC Interoperability Framework.

Needs to cover **Legal, Organisational, Technical, and Semantic Interoperability** (LOTS of Interoperability...)

Core driver of CODATA work on the WorldFAIR project and WorldFAIR+ initiative, and on the Cross-Domain Interoperability Framework (CDIF).

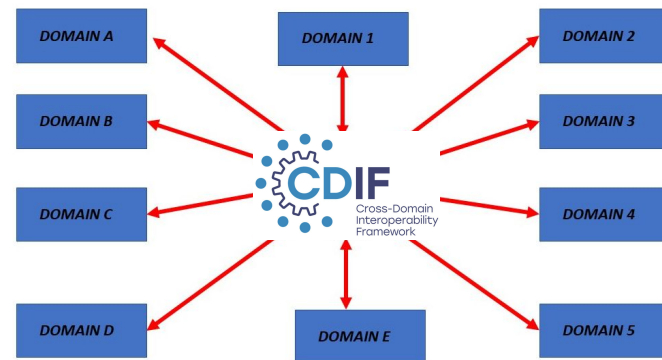


# What is the CDIF (Cross-Domain Interoperability Framework)?

- **Identifies a set of functional requirements for interoperability, particular for steps in data combination, and identifies good practices for each of these requirements.**
- Draws on work with the WorldFAIR case studies and with a number of international initiatives (ODIS, Science on Schema.org, UN Stats KG work, GBIF...)
- **Good web practices:** Significant proportion of CDIF rests on good web practice, domain neutral standards and good practice: disciplines can adopt or map.
- **Use cases:** domain or cross-domain projects or data services that need to combine data for analysis, modelling etc.
- **Directed at implementers:** describes use cases, identifies standards, gives guidance and on how to implement them.
- **Categorically not a new standard.** Rather it is a framework of existing and emerging standards.
- **A framework of standards/specifications to provide a *lingua franca*.**



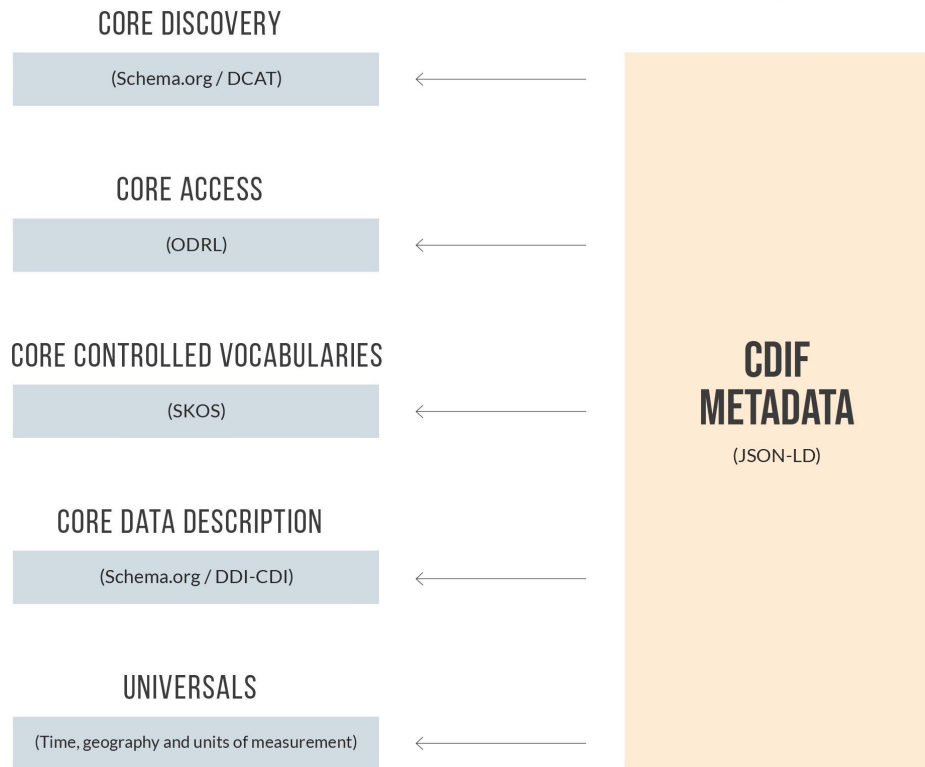
Source: xkcd.com



# What is CDIF?

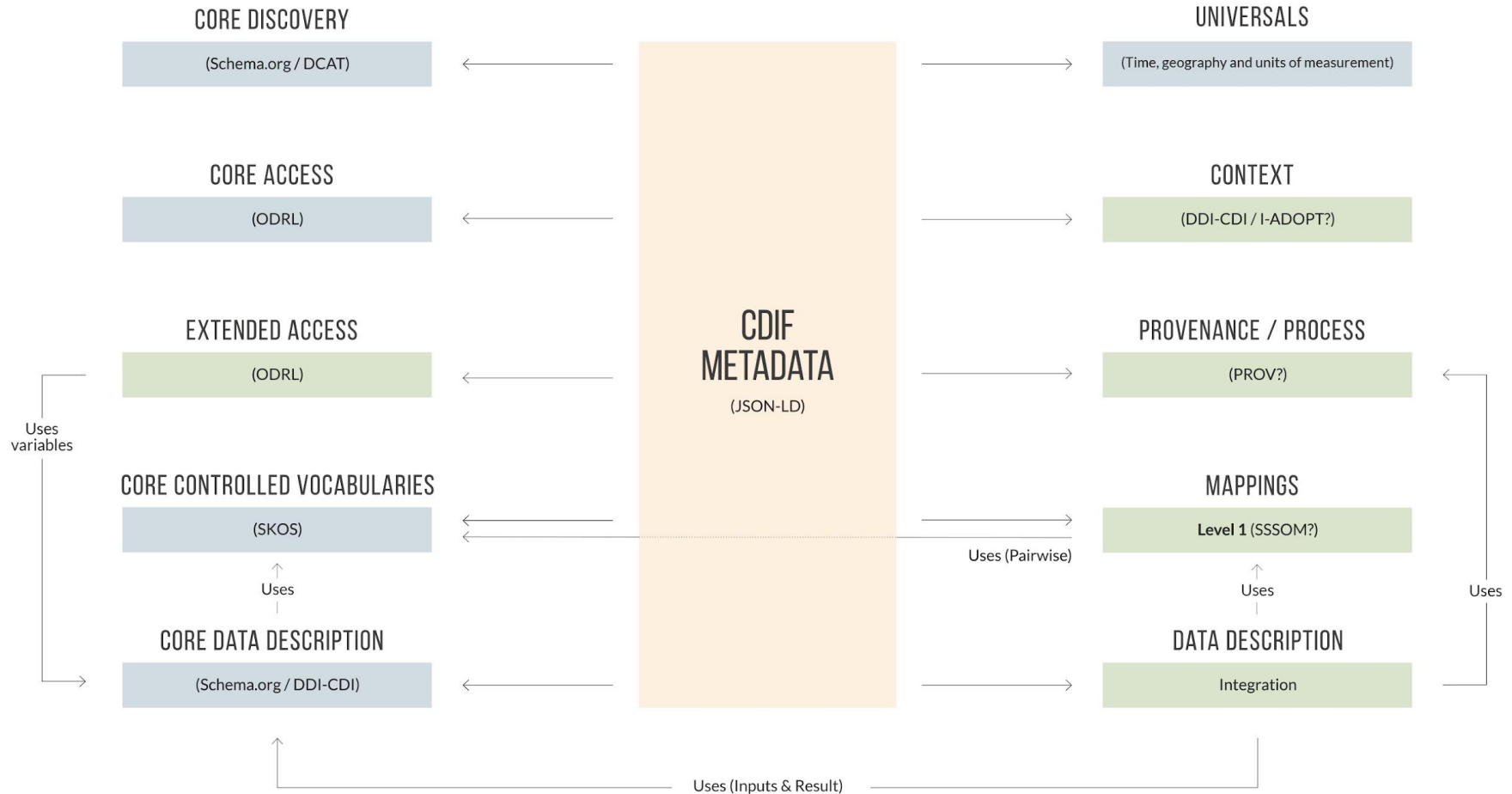


- The Cross Domain Interoperability Framework (CDIF) is a set of practical, **implementation-level** principles designed to improve data management practices within any community and lower the barriers to **cross-domain data reuse**. CDIF offers standards and methodologies for achieving different types of interoperability necessary for reusing data across diverse domains. It is (currently) built around five core profiles that address the essential functions for implementing cross-domain FAIR principles.
- Serves a number of use cases, but above all those of finding and combining / integrating data for subsequent research.
- CDIF was first released in May 2024 as an output of the WorldFAIR project:  
<https://doi.org/10.5281/zenodo.11236871>
- The point of reference for CDIF and its component profiles is now the CDIF Book: <https://bit.ly/CDIF-Book>






# CDIF, Next Steps



## Data layer: why Dataverse is suitable data repository for AI?

- ❖ Open source project developed by Institute for Quantitative Social Sciences (IQSS) at the Harvard University
- ❖ Published on github with a history back to 2006 (which is quite long for collective software development)
- ❖ Very dynamic and experienced development team working in an agile environment (e.g., community call scheduled once in two weeks)
- ❖ Clear vision and understanding of research communities requirements, public roadmap
- ❖ Strong community (mix of software developers, Dataverse instance providers, and partly user communities) behind of Dataverse is helping to improve the basic functionality and develop it further
- ❖ Dataverse has been selected as a data repository infrastructure by countries from all continents
- ❖ State-of-the art well developed architecture with rich API endpoints to build further application layers around Dataverse

# Croissant ML export in Dataverse


 **Dataverse** Search User Guide Support Sign Up Log In

Cars  
(Dataverse.org)

Root > Cars >

## Cars


Version 1.0

 Admin, Dataverse, 2024, "Cars", <https://doi.org/10.5072/FK2/DZRHUP>, Root, V1, UNF:6:RPd9EWHSZwqUvRZuKTJmQg== [fileUNF]

[Cite Dataset](#) Learn about [Data Citation Standards](#).

[Access Dataset](#)  
Contact Owner Share

Dataset Metrics  
0 Downloads

**Description** Data about cars.  
**Subject** Other  
**License/Data Use Agreement**  CC0 1.0

[Files](#) [Metadata](#) [Terms](#) [Versions](#)

**Citation Metadata**

<b>Persistent Identifier</b>	doi:10.5072/FK2/DZRHUP
<b>Publication Date</b>	2024-03-14
<b>Title</b>	Cars
<b>Author</b>	Admin, Dataverse (Dataverse.org)
<b>Point of Contact</b>	Use email button above to contact.
	Admin, Dataverse (Dataverse.org)
<b>Description</b>	Data about cars.

[Export Metadata](#)

- OAI\_ORE
- DataCite
- OpenAIRE
- Schema.org JSON-LD
- DDI
- Dublin Core
- Croissant**
- DDI HTML Codebook
- JSON

## Mappings

Croissant  
exporter  
Code

```
@context:
  @language: "en"
  @vocab: "https://schema.org/"
  citeAs: "cr:citeAs"
  column: "cr:column"
  conformsTo: "dct:conformsTo"
  cr: "http://nlcommons.org/croissant/"
  data:
    @id: "cr:data"
    @type: "@json"
    dataBiases: "cr:dataBiases"
    dataCollection: "cr:dataCollection"
  dataType:
    @id: "cr:dataType"
    @type: "@vocab"
    dct: "http://purl.org/dc/terms/"
    extract: "cr:extract"
    field: "cr:field"
    fileProperty: "cr:fileProperty"
    fileObject: "cr:fileObject"
    fileSet: "cr:fileSet"
    format: "cr:format"
    includes: "cr:includes"
    isEnumeration: "cr:isEnumeration"
    jsonPath: "cr:jsonPath"
    key: "cr:key"
    md5: "cr:md5"
    parentField: "cr:parentField"
    path: "cr:path"
    personalSensitiveInformation: "cr:personalSensitiveInformation"
    recordSet: "cr:recordSet"
    references: "cr:references"
    regex: "cr:regex"
    repeated: "cr:repeated"
    replace: "cr:replace"
    sc: "https://schema.org/"
    separator: "cr:separator"
    source: "cr:source"
    subField: "cr:subField"
    transform: "cr:transform"
    wd: "https://www.wikidata.org/wiki/"
  @type: "sc:Dataset"
  conformsTo: "http://nlcommons.org/croissant/1.0"
  name: "Cars"
  version: "1.0"
  citeAs: "https://doi.org/10.5072/FK2/DZRHUP"
  distribution:
    0:
      @type: "cr:FileObject"
      @id: "stata13-auto.dta"
      name: "stata13-auto.dta"
      encodingFormat: "application/x-stata-13"
      md5: "7b1281ce6b469796837a835377338c5a"
      contentUrl: "stata13-auto.dta"
  recordSet:
    0:
      @type: "cr:RecordSet"
      field:
        0:
          @type: "cr:Field"
          name: "make"
          description: "Make and Model"
          dataType: "sc:Text"
          source:
            @id: "6"
            fileObject:
              @id: "stata13-auto.dta"
```

# Model Context Protocol (MCP) for Dataverse

README

## MCP (Model Context Protocol) server for Dataverse

Credits: this work is funded by the [SSHOC-NL](#) project developing [Semantic Croissant](#). The first version of [Croissant](#) export for Dataverse was implemented by Philip Durbin (Harvard IQSS) and Slava Tykhonov (DANS-KNAW).

Croissant is a special language for machines, built on top of Schema.org. With Croissant, we aim to solve multilingual challenges and finally speak the same language across the planet. Even if it's artificial.

### Getting started with [mcp.dataverse.org](https://mcp.dataverse.org)

When getting started, we recommend the public MCP server for Dataverse at <https://mcp.dataverse.org>. (Below you'll also find instructions on how to run the MCP server locally.) You can visit <https://mcp.dataverse.org/tools> for an inventory of available tools.

You will need an MCP client with AI agent support such as [Cursor](#), [Visual Studio Code](#), [Windsurf Editor](#), or [Zed](#).

### (Optional) Command line test

Before you get too far into configuring your MCP client, you could try this quick test to get information about a dataset by passing its DOI.


```
curl -X POST "https://mcp.dataverse.org/tools/get_croissant_record" -H "Content-Type: applic
```


### Configuring your MCP Client

You'll be using <https://mcp.dataverse.org/sse> as the URL and SSE (Server-Sent Events) as the type of MCP server. Click the arrow to expand instructions for your MCP client.

- Cursor
- Visual Studio Code
- Windsurf

Contributors 2

 4tikhonov Vyacheslav Tykhonov

 pdurbin Philip Durbin


Languages

Python 98.5%

Dockerfile 1.5%


Suggested workflows

Based on your tech stack

 **Python package**


Configure

Create and test a Python package on multiple Python versions.

 **Python application**

Configure

Create and test a Python application.

 **Pylint**

Configure

Lint a Python application with pylint.

[More workflows](#) [Dismiss suggestions](#)

Install it and try in your IDE! <https://mcp.dataverse.org>

<https://github.com/gdcc/mcp-dataverse>

# MCP protocol for Dataverse powered by Croissant 1.0 ML

## Dataverse MCP Use Cases

The **Dataverse MCP** (Model Context Protocol) enables automated tools and LLM agents to explore and analyze the global Dataverse data network.



### Onboarding: Before You Start

To get started, complete onboarding at:



<https://mcp.dataverse.org>

This prepares your agent or system to interact with the MCP tools programmatically.



### Available Tools

Tool	Description
<code>fetch</code>	Fetch a website and return its content.
<code>onboarding</code>	Get onboarding instructions for an LLM or system to act independently.
<code>get_croissant_record</code>	Convert a dataset to Croissant ML format using DOI or handle.
<code>datatool</code>	Process and inspect a file in a dataset by DOI and filename.
<code>overview</code>	Get an overview of all Dataverse installations worldwide.
<code>overview_datasets</code>	Get dataset statistics for a specific Dataverse host.
<code>overview_files</code>	Get file statistics for a specific Dataverse host.
<code>search_datasets</code>	Search datasets within a specific Dataverse installation.

# Common questions for Dataverse MCP

## Common Questions

- Do onboarding and give me overview of all Dataverses
- List Dataverses from the US
- How many datasets in dataverse.nl?
- How many datasets on {query} exist in the whole Dataverse network?
- How many Dataverse installations were created over the last 10 years, by country?
- How many datasets exist in France?
- How many datasets on economics are in dataverse.nl?
- Which countries have added the most new Dataverse nodes since 2015?
- What kinds of files are in energy consumption datasets from dataverse.nl?
- How many datasets were published in France in 2024?
- Compare number of datasets between Johns Hopkins and Harvard Dataverse.
- I'm studying gender inequality in education. What datasets could help?
- give me overview of dataset doi:10.17026/dans-x8n-hfvr
  - a. where this coin was found?
  - b. what is the age of the coin?



## Summary: Dutch Dataverses (June 2025)

- **DataverseNL:** 8,161 datasets
- **DANS Data Station Archaeology:** 162,435 datasets
- **DANS SSH:** 7,932 datasets
- **DANS Life Sciences:** 997 datasets
- **DANS Physical & Technical Sciences:** 845 datasets
- **IISH Dataverse:** 358 datasets
- **ODISSEI Portal:** 10,163 datasets

These are the major Dutch Dataverse nodes covering a wide range of research areas.

## MCP powered by Croissant ML 2.0

MCP primitives are “moving pieces” modular, dynamic, and designed to evolve in time:

- LLM prompts can be created both by human and machine
- Resources (data points, metadata) can be dynamic such as streaming
- MCP tools can be connected and disconnected (offline)

MCP is about structuring the context in which machine learning models, especially LLMs, operate. This includes everything that informs, constrains, or modifies model behavior.

Open question: How to make workflows persistent, sustainable and FAIR?



# Decentralized Resource Identity & Framework of Trust (DRIFT)

We introduce “DRIFT” - decentralized identity and trust layer that propagates **traceable context** - such as datasets (Croissant), semantic (CDIF), policies (ODRL), sessions, prompts, user metadata, and spans — across AI tools and services in **event-driven environments** like the Model Context Protocol (MCP).

In software engineering, **data drift** is a key concern that can affect system reliability and trust in data processing pipelines. DRIFT helps mitigate these risks by making identity and changes traceable.

Three common types of drift include:

- **Infrastructure Drift** – Changes in the software environment that can break or invalidate infrastructure configurations.
- **Structural Drift** – Changes to the data schema that may invalidate databases or data contracts.
- **Semantic Drift** – Changes in the meaning of data without altering its structure, often caused by multiple developers independently modifying system components.

By assigning decentralized identifiers (DIDs) and enforcing consistent, signed metadata, DRIFT supports systems in managing and auditing these types of drift.

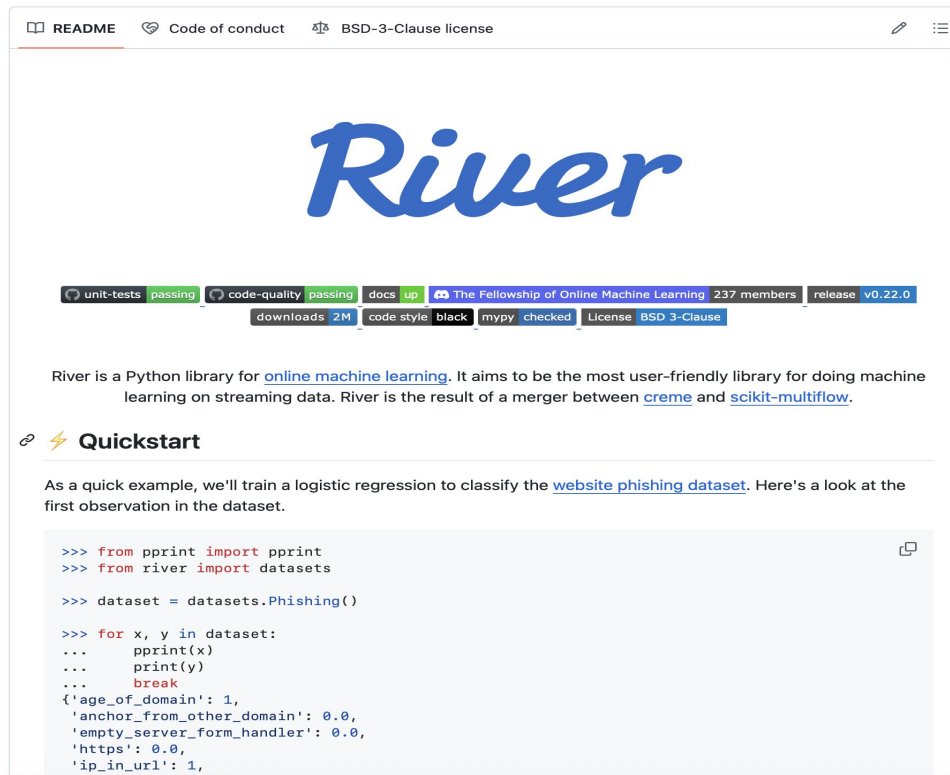
DRIFT enables trusted interoperability between distributed AI components by providing unique, verifiable identities for all parts of an AI pipeline, such as sessions, prompts, users and tools.

“In [computer science](#), **online machine learning** is a method of [machine learning](#) in which data becomes available in a sequential order and is used to update the best predictor for **future data at each step**, as opposed to batch learning techniques which generate the best predictor by learning on the entire [training data set](#) at once.”

Source: [Wikiland](#)



# Why it's important for AI? DRIFT has another nature



The screenshot shows the GitHub repository for the 'River' Python library. At the top, there are links for 'README', 'Code of conduct', and 'BSD-3-Clause license'. The main heading is 'River' in a large, blue, cursive font. Below the heading, there are several status badges: 'unit-tests passing', 'code-quality passing', 'docs up', 'The Fellowship of Online Machine Learning 237 members', 'release v0.22.0', 'downloads 2M', 'code style black', 'mypy checked', and 'License BSD 3-Clause'. A paragraph describes River as a Python library for online machine learning, aiming to be the most user-friendly library for doing machine learning on streaming data, and mentions it is a merger between 'creme' and 'scikit-multiflow'. Below this is a 'Quickstart' section with a code example for training a logistic regression model on the 'website phishing dataset'.

unit-tests passing code-quality passing docs up The Fellowship of Online Machine Learning 237 members release v0.22.0 downloads 2M code style black mypy checked License BSD 3-Clause

River is a Python library for [online machine learning](#). It aims to be the most user-friendly library for doing machine learning on streaming data. River is the result of a merger between [creme](#) and [scikit-multiflow](#).

### Quickstart

As a quick example, we'll train a logistic regression to classify the [website phishing dataset](#). Here's a look at the first observation in the dataset.

```
>>> from pprint import pprint
>>> from river import datasets

>>> dataset = datasets.Phishing()

>>> for x, y in dataset:
...     pprint(x)
...     print(y)
...     break
{'age_of_domain': 1,
 'anchor_from_other_domain': 0.0,
 'empty_server_form_handler': 0.0,
 'https': 0.0,
 'ip_in_url': 1,
 ...}
```

DRIFT in fact is dedicated for Machine Learning on streaming data where knowledge graphs are continuously evolving.

We need to change current AI paradigm to give all rights back to content creators and publishers:

- identifiers assigned to MCP primitives should guarantee reproducibility.
- Reliable mechanism to filter out low quality (“poison”) materials
- Users should get answers based on trustworthy and verifiable information
- New generation of AI models should be fully decentralized

# Decentralized FAIR data network

## Decentralized identifier documents [\[ edit \]](#)

---

Decentralized identifier documents or **DIDs** are a type of [identifier](#) that enables a verifiable, decentralized [digital identity](#).<sup>[1]</sup> They are based on the [self-sovereign identity](#) paradigm. A DID identifies any subject (e.g., a person, organization, thing, data model, abstract entity, etc.) that the controller of the DID decides that it identifies. These identifiers are designed to enable the controller of a DID to prove control over it and to be implemented independently of any centralized registry, [identity provider](#), or [certificate authority](#). DIDs are [URIs](#) that associate a DID subject with a DID document allowing trustable interactions associated with that subject. Each DID document can express cryptographic material, verification methods, or service endpoints, which provide a set of mechanisms enabling a DID controller to prove control of the DID. Service endpoints enable trusted interactions associated with the DID subject. A DID document might contain semantics about the subject that it identifies. A DID document might contain the DID subject itself (e.g. a data model).<sup>[3][2]</sup>

source:  
Wikipedia

We're considering experimental implementation of the decentralized identifiers for various content types extension to archive various types of content.

DIDs can be assigned to any artefacts (not only MCP primitives) including images, audio and video, for example, to store and link metadata records and provenance information together with their digitized content. ODRL policies can be also digitally signed and receive own DID documents.

**DID can be private (invisible and not resolvable for public) but available for access with cryptokey.**

# FAIR decentralized identifiers for AI

We envision the near future where it will be possible to create a decentralized system which will not depend on any specific registry, one provider, one authority, etc., so all connections will be established in a peer-to-peer network, and but will be persistent at the same time. This solution should support AI workflows and infrastructure with FAIR principles of findability, accessibility, interoperability, and reusability (FAIR).

This can be achieved with global decentralized identifier - DID.

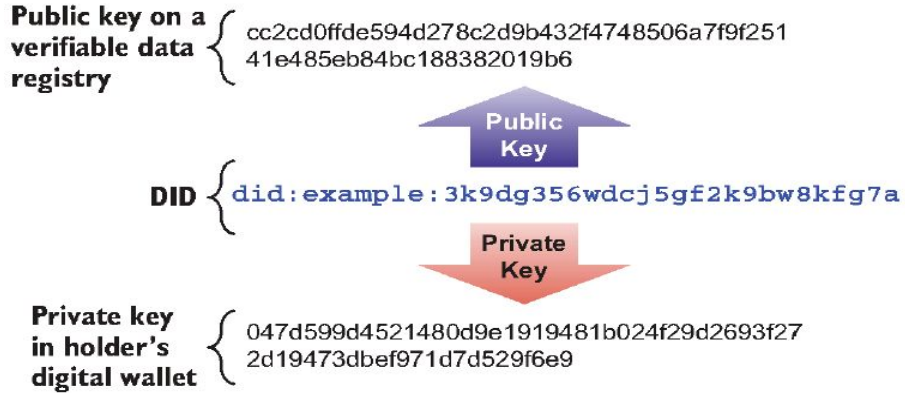
The resolution of the decentralized identifier (DID) is **cryptographically verifiable to prove the identity and the ownership** of that identifier and can support Model Context Protocol with sustainable infrastructure to keep provenance and origins of prompts, resources and tools.

Core DID features are listed below:

1. A permanent (persistent) identifier (never change)
2. A resolvable identifier (you can look it up to discover metadata)
3. A cryptographically-verifiable identifier (with private and public keys)
4. A decentralized identifier (no centralized authority)

DID should bring control of all provenance and metadata back to their owners instead of giving them away. In the same time public part will/could not be very different from other persistent identifiers like DOIs and even replace them for the specific use cases like sharing sensitive data.

# The role of private and public key, and service endpoints in DID



## EXAMPLE 20: Usage of the service property

```
{
  "service": [{
    "id": "did:example:123#linked-domain",
    "type": "LinkedDomains", // external (property value)
    "serviceEndpoint": "https://bar.example.com"
  }]
}
```

Service endpoints can tell how exactly to interact with the subject, what kind of protocols, what kind of network endpoints are available to connect, for example, to an agent that represents the data subjects so that you can then exchange credentials or some other messages.

# Policies: demo of Open Digital Right Language (ODRL) infrastructure

The screenshot displays the ODRL Infra web application interface. At the top left is the ODRL Infra logo with the tagline "PROTOCOL LAYER". A navigation sidebar on the left lists several modules: Live Demo (highlighted), Dashboard, DID Manager, VC Wallet, Policy Builder, Prompts Manager, Variables, Groups, and Croissants. The top right corner features a "Login" button and a dark mode toggle. The main content area is titled "Live Demo: FAIR Data Spaces managed by ODRL" and includes a subtitle: "Automated end-to-end verification of Open Digital Rights Language policies and DID artifacts." Below the title are tabs for "Test Group", "Test Policy", "Test Prompt", "Test Variable", and "Test Croissant" (which is selected). The interface is divided into two main panels. The left panel contains a green "Run Test Croissant" button and a vertical flowchart with three steps: "Fetch Croissant" (Get JSON-LD from Dataverse), "Anchor DID" (Assign DID to Croissant metadata), and "Verify DID" (Resolve and validate metadata). Below the flowchart is an "EXECUTION LOGS" section showing the status "Ready to start...". The right panel, titled "Live Data Output", contains a large play button icon and the text "Run the scenario to see live API data." At the bottom left of the interface, a green dot indicates the "SYSTEM ONLINE" status.

**ODRL Infra**  
• PROTOCOL LAYER

**Live Demo**  
Test all applications

**Dashboard**  
Explore infrastructure

**DID Manager**  
Bookmarks and resolver

**VC Wallet**  
Google, Github, ORCID a...

**Policy Builder**  
Create your policy

**Prompts Manager**  
FAIR LLM prompts

**Variables**  
Cross-Domain Interopera...

**Groups**  
Organization Ontology

**Croissants**  
AI-Ready data

**Live Demo: FAIR Data Spaces managed by ODRL**  
Automated end-to-end verification of Open Digital Rights Language policies and DID artifacts.

Test Group Test Policy Test Prompt Test Variable **Test Croissant**

**Run Test Croissant**

- Fetch Croissant  
Get JSON-LD from Dataverse
- Anchor DID  
Assign DID to Croissant metadata
- Verify DID  
Resolve and validate metadata

EXECUTION LOGS  
Ready to start...

**Live Data Output**

Run the scenario to see live API data.

SYSTEM ONLINE

# Universal Resolver for AI primitives

## Registrar output:

```
{
  "didState": {
    "did": "did:oyd:zQmZNDWhhr3cHaL2pFLCpwXXbK5WVKY5wJHtC3RvrgQIG5G",
    "state": "finished",
    "secret": {
      "documentKey": "z155Z4B25YCWgKYuFwApp531iOGK3BcDc2HkR2ckXvQae4",
      "revocationKey": "z155YBrmeShGQwGPNyTEGxfGNLIeL2HN9zy1Y4XmtBjIAfC",
      "revocationLog": {
        "ts": 1749461892,
        "op": 1,
        "doc": "zQmP9FPnN8HWHtHmJExcXvGhBzGjGsDjSMfCxiRgNzdKcCV",
        "sig": "z4pmbit5ijGHkvqYt1qGkqYemkrXDBMSpS37cv56banJ2L9Lc4wUaAdCjwsBca8VJG7VmRezzmVHVgnLcJcGzq6rR"
      }
    },
    "didDocument": {
      "@context": [
        "https://www.w3.org/ns/did/v1",
        "https://w3id.org/security/suites/ed25519-2020/v1"
      ],
      "id": "did:oyd:zQmZNDWhhr3cHaL2pFLCpwXXbK5WVKY5wJHtC3RvrgQIG5G",
      "verificationMethod": [
        {
          "id": "did:oyd:zQmZNDWhhr3cHaL2pFLCpwXXbK5WVKY5wJHtC3RvrgQIG5G#key-doc",
          "type": "Ed25519VerificationKey2020",
          "controller": "did:oyd:zQmZNDWhhr3cHaL2pFLCpwXXbK5WVKY5wJHtC3RvrgQIG5G",
          "publicKeyMultibase": "z6MuxxwvcChKb2wSLBeXRV594k5KtqAZa6H7YXFRXC54Hjx"
        }
      ],
      "service": [
        {
          "id": "did:oyd:zQmZNDWhhr3cHaL2pFLCpwXXbK5WVKY5wJHtC3RvrgQIG5G#payload",
          "type": "Custom",
          "serviceEndpoint": "https://oydid.ownyourdata.eu",
          "payload": {
            "prompt": "hello MCP"
          }
        }
      ]
    },
    "log_hash": "zQmT6wtciv8n1d4maP6BcyLrXFZyk6D7LzjVCxjfnQ7hZH",
    "log": {
      "ts": 1749461892,
      "op": 2,
      "doc": "zQmZNDWhhr3cHaL2pFLCpwXXbK5WVKY5wJHtC3RvrgQIG5G",
      "sig": "z3av6sHSPzZH2Loqq3GuGwMApHpkPgYLYZXnzbuKbGckRTrA4WwX4o6LNCvAZxeD8ueuWCFsDUySpceo7hmKs9S",
      "previous": [
        {
          "ts": 1749461892,
          "op": 0,
          "doc": "zQm5Ufr1jKnWCGa5eDmEsBuwS5fTyngkTcyPGQVwM45hyD",
          "sig": "z2kyjucLNUvgJEKBCx3isfkvfNpYIdanT5bzHXbXha5y8KbQ1knz1wMaPwEugCA1PPS5wptYHPWYAew3KR6h",
          "previous": [
            {}
          ]
        }
      ]
    }
  }
}
```

did-url

did:oyd:zQmZNDWhhr3cHaL2pFLCpwXXbK5WVKY5wJHtC3RvrgQIG5G

Resolve

Clear

Examples

Copy link to result

Check Compliance

RESULT

DID DOCUMENT

RESOLUTION METADATA

DOCUMENT METADATA

```
{
  "@context": [
    "https://www.w3.org/ns/did/v1",
    "https://w3id.org/security/suites/ed25519-2020/v1"
  ],
  "id": "did:oyd:zQmZNDWhhr3cHaL2pFLCpwXXbK5WVKY5wJHtC3RvrgQIG5G",
  "verificationMethod": [
    {
      "id": "did:oyd:zQmZNDWhhr3cHaL2pFLCpwXXbK5WVKY5wJHtC3RvrgQIG5G#key-doc",
      "type": "Ed25519VerificationKey2020",
      "controller": "did:oyd:zQmZNDWhhr3cHaL2pFLCpwXXbK5WVKY5wJHtC3RvrgQIG5G",
      "publicKeyMultibase": "z6MuxxwvcChKb2wSLBeXRV594k5KtqAZa6H7YXFRXC54Hjx"
    }
  ],
  "service": [
    {
      "id": "did:oyd:zQmZNDWhhr3cHaL2pFLCpwXXbK5WVKY5wJHtC3RvrgQIG5G#payload",
      "type": "Custom",
      "serviceEndpoint": "https://oydid.ownyourdata.eu",
      "payload": {
        "prompt": "hello MCP"
      }
    }
  ]
}
```

## Decentralized identifier (DID):

did:oyd:zQmZNDWhhr3cHaL2pFLCpwXXbK5WVKY5wJHtC3RvrgQIG5G

curl

<https://dev.uniresolver.io/1.0/identifiers/did:oyd:zQmZNDWhhr3cHaL2pFLCpwXXbK5WVKY5wJHtC3RvrgQIG5G>

```
"service": [
  {
    "id":
"did:oyd:zQmZNDWhhr3cHaL2pFLCpwXXbK5WVKY5wJHtC3RvrgQIG5G#payload",
    "type": "Custom",
    "serviceEndpoint":
"https://oydid.ownyourdata.eu",
    "payload": {
      "prompt": "hello MCP"
    }
  }
]
```

Try this! <https://dev.uniresolver.io>

# Nectar Publisher as a “human in the loop” - CDIF

Root >

## Earthquake dataset

Version 1.0



User, Test, 2025, "Earthquake dataset", <https://doi.org/10.5072/FK2/BXSHPO>, Root, V1,  
UNF:6:bkDZ89grt9xKHlGoW5hkZW== [fileUNF]

[Cite Dataset](#)

[Learn about Data Citation Standards.](#)

[Access Dataset](#)

[Contact Owner](#)

[Share](#)

[Dataset Metrics](#)

26 Downloads

### Description

Earthquake data from Kaggle

### Subject

Astronomy and Astrophysics

### License/Data Use Agreement



[Files](#)

[Metadata](#)

[Terms](#)

[Versions](#)

1 File



earthquakes.tab


Tabular Data - 1.6 MB

Published Nov 27, 2025

26 Downloads

23 Variables, 7717 Observations UNF:6:bkDZ...kZW==





Nectar Publisher

pre-Dataverse 0.1

import data

import metadata

import data via service

about

Dataset Information
















Variables

Export documentation

Mappings

Endpoint

Variable cascade

	Name	Label	Definition	Type	Coded	Details
0	time	time	Time is the continuous, unidirectional	String / Text	<input type="checkbox"/>	
1	latitude	latitude	The angular coordinate that specifies	Decimal	<input type="checkbox"/>	
2	longitude	longitude	A geographic coordinate that specifies	Decimal	<input type="checkbox"/>	
3	depth	depth	Depth is the vertical distance from a	Decimal	<input type="checkbox"/>	
4	mag	mag	Magnitude (mag) is a quantitative measure	Decimal	<input type="checkbox"/>	
5	magType	magType	The magnitude type is a categorical variable	String / Text	<input type="checkbox"/>	
6	nst	nst	The variable 'nst' represents the number of	Decimal	<input type="checkbox"/>	
7	gap	gap	A gap is the measurable space or separation	Decimal	<input type="checkbox"/>	
8	dmin	dmin	The minimum spatial or temporal separation	Decimal	<input type="checkbox"/>	
9	rms	rms	Root Mean Square (RMS) is a statistical measure	Decimal	<input type="checkbox"/>	
10	net	net	A derived numeric quantity that represents	String / Text	<input type="checkbox"/>	
11	id	id	The measure of the warmth or coldness	String / Text	<input type="checkbox"/>	
12	updated	updated	The measure of the average kinetic energy	String / Text	<input type="checkbox"/>	
13	place	place	A variable that captures the geographic	String / Text	<input type="checkbox"/>	
14	type	type	The ambient air temperature measurement	String / Text	<input type="checkbox"/>	

Demo



# Responsible AI: Croissant and DDI

## (Data Documentation Initiative)

### Responsible AI


“As AI advances at rapid speed there is increased recognition among researchers, practitioners and policy makers that we need to explore, understand, manage, and assess [its economic, social, and environmental impacts](#). One of the main instruments to operationalise responsible AI (RAI) is dataset documentation.

This is how Croissant helps address RAI:

1. It proposes a machine-readable way to capture and publish metadata about ML datasets – this makes existing documentation solutions like [Data Cards](#) easier to publish, share, discover, and reuse;
2. It records at a granular level how a dataset was created, processed and enriched throughout its lifecycle – this process is meant to be automated as much as possible by integrating Croissant with popular ML frameworks. By allowing the metadata to be loaded automatically, Croissant also enables developers to compute RAI metrics automatically and systematically, identifying potential data quality issues to be fixed.

Croissant is designed to be modular and extensible. One such extension is the Croissant RAI vocabulary, which addresses 7 specific use cases, starting with the data life cycle, data labeling, and participatory scenarios to AI safety and fairness evaluation, traceability, regulatory compliance and inclusion. More details are available in the . We welcome additional extensions from the community to meet the needs of specific data modalities (e.g. audio or video) and domains (e.g. geospatial, life sciences, cultural heritage).”

Croissant spec v1.0

 CESSDA DATA CATALOGUE (CDC) DDI3.3 PROFILE					
DDI_XPath	Required	UI Label	Type	Repeatable	Usage note
/ddi:DDIInstance/@xmi:lang	Recommended	Study description available in	Attribute		ISO 639-1 codes are strongly encouraged to be used
/ddi:DDIInstance/@xsi:schemaLocation	Recommended		Attribute		System-required information - nearly always "ddi:instance-3.3 http://www.ddialliance.org/Specification/DDI-Lifecycle/3.3/XMLSchema/instance.xsd"
/ddi:DDIInstance/r:Citation/r:Title/r:String	Optional		Content element	No	Title of the CDC XML document - note that the Study Title goes in /ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String
/ddi:DDIInstance/r:Citation/r:Title/r:String/@xmi:lang	Mandatory if 'r:Title/r:String' element is present		Attribute		ISO 639-1 codes are strongly encouraged to be used
/ddi:DDIInstance/r:ResourcePackage/pi:PhysicalInstance/r:Citation/r:Language	Recommended	Language of data file(s)	Content element	Yes	ISO 639-1 codes are strongly encouraged to be used if available for the language.
/ddi:DDIInstance/s:StudyUnit/r:UserID	Mandatory	Study number / PID	Content element	Yes	This element serves two use cases. (1) Unique archival number. If no such number available, a PID may be used alternatively (and in such a case, the PID must be captured in this element "as well as" in "/ddi:DDIInstance/s:StudyUnit/r:Citation/r:InternationalIdentifier/r:IdentifierContent"). The @typeOfUserID attribute must be "StudyNumber". (2) A URL or URN (a reference to a web resource that specifies its location) linking to the study description on the SP website. CDC provides a link from the study information there to the study description on the SP website, to allow users to access the data. [This element not visible in the UI, it is only needed for the link]. The @typeOfUserID attribute must be "URLServiceProvider".
/ddi:DDIInstance/s:StudyUnit/r:UserID/@typeOfUserID	Mandatory		Attribute		Must be specified when the "ddi:DDIInstance/s:StudyUnit/r:UserID" element is used for the unique archival number.
/ddi:DDIInstance/s:StudyUnit/r:UserID/@typeOfUserID	Mandatory		Attribute		Must be specified when the "ddi:DDIInstance/s:StudyUnit/r:UserID" element is used for the URL of the study description at the SP website.
/ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String	Mandatory	Study title	Content element	Yes	Title of the Study (as opposed to the title of the XML document).
/ddi:DDIInstance/s:StudyUnit/r:Citation/r:Title/r:String/@xmi:lang	Mandatory		Attribute		Language of the study title. ISO 639-1 codes are strongly encouraged to be used.
/ddi:DDIInstance/s:StudyUnit/r:Citation/r:Creator/r:CreatorReference	Recommended	Creator	Container element	Yes	Reference to Principal investigator Person OR Institution
/ddi:DDIInstance/s:StudyUnit/r:Citation/r:Creator/r:CreatorReference/r:TypeOfObject	Mandatory of 'r:CreatorReference' is present.	Creator	Content element	Yes	Type of object being referenced - in this case, it should always be either "Individual" or "Organization".
/ddi:DDIInstance/s:StudyUnit/r:Citation/r:Publisher/r:PublisherReference	Recommended	Publisher	Container element	Yes	A reference to the name of the institution publishing the metadata, i.e. the name of the actual CESSDA Service Provider providing the metadata information. Only on study level.
/ddi:DDIInstance/s:StudyUnit/r:Citation/r:Publisher/r:PublisherReference/r:TypeOfObject	Mandatory if 'r:PublisherReference' is present	Publisher	Content element	Yes	Type of object being referenced - in this case, it should always be either "Individual" or "Organization".

[CESSDA DDI profile](#)



# DDI variables in Croissant - breaking the bias

(attributes, categories, units of measurements, ...)



Export Metadata ▼

OAI\_ORE

DataCite

OpenAIRE

Schema.org JSON-LD

DDI

Dublin Core

DDI HTML Codebook

JSON

```
<var ID="v63828" name="gender" intrvl="contin">
  <location fileid="f181579"/>
  <category>
    <catValu>1</catValu>
    <labl level="category">male, or</labl>
  </category>
  <catValu>2</catValu>
  <labl level="category">female</labl>
  </category>
  <varFormat type="numeric"/>
  <notes subject="Universal Numeric Fingerprint" level="variable" type="Dataverse:UNF13:MDMaBJEgQoqgVvUPlSKwA=="/>
</var>
<var ID="v63998" name="marit_b" intrvl="contin">
  <location fileid="f181579"/>
  <category>
    <catValu>1</catValu>
    <labl level="category">no</labl>
  </category>
  <catValu>2</catValu>
  <labl level="category">RF</labl>
  </category>
  <catValu>3</catValu>
  <labl level="category">OK</labl>
  </category>
  <catValu>4</catValu>
  <labl level="category">yes</labl>
  </category>
  <varFormat type="numeric"/>
  <notes subject="Universal Numeric Fingerprint" level="variable" type="Dataverse:UNF13:oueOKfY5yBdpXf9YnIGM0=="/>
</var>
<var ID="v63876" name="marit_b" intrvl="contin">
  <location fileid="f181579"/>
  <category>
    <catValu>1</catValu>
    <labl level="category">divorced, or</labl>
  </category>
  <catValu>2</catValu>
  <labl level="category">never married</labl>
  </category>
  <catValu>3</catValu>
  <labl level="category">widowed</labl>
  </category>
```

## Dataverse-Data-Curation-Tool

The Data Curation Tool (DCT) allows data owners and curators to view summary statistics for variables and to create and edit variable-level metadata for any tabular file in a data set. This stand-alone component is built to complement [The Dataverse Project](#). The Data Curation tool is integrated into dataverse for .tab files under the configure button.

The DCT is an Angular application and uses the Angular Material Design component library.

## Responsible AI in Croissant?

Search

Items Per Page: 5 1 - 7 of 21

Edit Variable

<input type="checkbox"/>	ID	Name	Label	Weight	
<input type="checkbox"/>	v457580	RID	RpsRespondent	No Weight	
<input type="checkbox"/>	v457594	Q1	Have you, or has anyone in your household had a fever, that is, a temperature above 38 degrees Celsius or about 100 degrees Fahrenheit, in the past week?	No Weight	
<input type="checkbox"/>	v457599	Q2	Are you, or is anyone in your household currently suffering from a new cough in the past week?	No Weight	
<input type="checkbox"/>	v457596	Q3	Are you, or is anyone in your household currently suffering from new headaches in the past week?	No Weight	
<input type="checkbox"/>	v457593	Q4	Are you, or is anyone in your household suffering from a new sore throat in the past week?	No Weight	
<input type="checkbox"/>	v457587	Q5	Are you, or is anyone in your household suffering from a loss of taste or smell in the past week?	No Weight	
<input type="checkbox"/>	v457595	Q6	Are you, or is anyone in your household suffering from new diarrhea in the past week?	No Weight	

ID: v457599  
Name: Q2

Label  
Are you, or is anyone in your household currently suffering from a new cough in the past week?

Literal Question

Interviewer Instructions

Post Question

Universe

Notes  
UNF:6:M6R5ZcCIBShqNvwelw==

Group(s)  
Add to Group

No Weight Assigned

Is Weight

# DDI-CDI transformations in CDIF

## (Cross-domain interoperability framework)

Input

Frequency	Year	Age Cohort	Sex	Status	Median Income (USD)
A	2003	C	M	ACT	5500
A	2003	G	F	ACT	7500
A	2004	E	M	EST	10000
A	2005	B	F	ACT	14000
A	2004	B	M	EST	2000



Output

```
Code Blame 145 lines (145 loc) · 3.38 KB
1  {
2    "@context": "https://ddi-alliance.bitbucket.io/DDI-CDI/DDI-CDI_v1.0-rc1/encoding/json-ld/ddi-cdi.jsonld",
3    "@graph": [
4      {
5        "@id": "#Frequency",
6        "@type": "InstanceVariable",
7        "name": "Frequency",
8        "hasIntendedDataType": "http://rdf-vocabulary.ddialliance.org/cv/DataType/1.1.2/#String"
9      },
10     {
11       "@id": "#Year",
12       "@type": "InstanceVariable",
13       "name": "Year",
14       "hasIntendedDataType": "http://rdf-vocabulary.ddialliance.org/cv/DataType/1.1.2/#Integer"
15     },
16     {
17       "@id": "#Age_Cohort",
18       "@type": "InstanceVariable",
```

Graph representation serves as  
“Navigation layer” for AI models:  
“The system of cells interlinked”

# Dataset in Dataverse represented in CDIF

Root >

## X-ray absorption spectra for K edge, selenium in sodium selenate

Version 3.0



Richard, Stephen M., 2025, "X-ray absorption spectra for K edge, selenium in sodium selenate", <https://doi.org/10.5072/FK2/4ZSKVJ>, Root, V3

[Cite Dataset](#)

Learn about [Data Citation Standards](#).

[Access Dataset](#)  
[Contact Owner](#) [Share](#)

[Dataset Metrics](#)

129 Downloads

### Description

comment from xdi file: room temperature; measured at beamline 13-BM-D. need a better description of what's in this dataset-- maybe the DCAT metadata has what we need? (2025-11-18)

### Subject

Computer and Information Science

### Keyword

Se, K

### License/Data Use Agreement

 [CC0 1.0](#)

[Files](#) [Metadata](#) [Terms](#) [Versions](#)

Search this dataset... 

[Files](#) [Metadata](#) [Terms](#) [Versions](#)

Search this dataset... 

Filter by

File Type: All Access: All

 Sort

1 to 2 of 2 Files

 Download



Se\_Na2SeO4\_rt\_01.xdi

Unknown - 18.3 KB

Published Nov 19, 2025

76 Downloads

MD5: 012...757

[https://github.com/CDIF-4-XAS/XAS-CDIF/blob/main/se\\_na2so4\\_rt.xdi](https://github.com/CDIF-4-XAS/XAS-CDIF/blob/main/se_na2so4_rt.xdi)



XDI-CDIF-Mapping.xlsx

MS Excel Spreadsheet - 15.0 KB

Published Nov 21, 2025

53 Downloads

MD5: 53c...5c3



## Mappings

1	xdi dictionary	CDIF implementation		
Beamline	"prov:used": [ { "@type": [ "schema:Thing", "prov:Entity", "xas:Beamline" ],			
2	Beamline.collimation	"schema:additionalProperty": [ { "@type": "schema:PropertyValue", "schema:propertyID": { "@id": "xas:collimation" }, "schema:value": "(value)" },	Beamline	xdi
3	Beamline.detector	"schema:additionalProperty": [ { "@type": "schema:PropertyValue", "schema:propertyID": { "@id": "xas:detector" }, "schema:value": "(value)" },	Beamline	dat
4	Beamline.energy_range	"schema:additionalProperty": [ { "@type": "schema:PropertyValue", "schema:propertyID": { "@id": "xas:energy_range" }, "schema:value": "(value)" },	Beamline	dat
5	Beamline.energy_resolution	"schema:additionalProperty": [ { "@type": "schema:PropertyValue", "schema:propertyID": { "@id": "xas:energy_resolution" }, "schema:value": "(value)" } ]	Beamline	dat
6	Beamline.flux	"schema:additionalProperty": [ { "@type": "schema:PropertyValue", "schema:propertyID": { "@id": "xas:flux" }, "schema:value": "(value)" } ]	Beamline	dat
7	Beamline.focusing	"schema:additionalProperty": [ { "@type": "schema:PropertyValue", "schema:propertyID": { "@id": "xas:focusing" }, "schema:value": "(value)" } ]	Beamline	xdi
8	Beamline.harmonic_rejection	"schema:additionalProperty": [ { "@type": "schema:PropertyValue", "schema:propertyID": { "@id": "xas:harmonic_rejection" }, "schema:value": "(value)" } ]	Beamline	xdi
9	Beamline.name	"schema:additionalProperty": [	Beamline	xdi

# “Interlink”: Semantic Croissant with WikiData prediction

Variable	Linked Wikidata ID	Label (English)	Description (excerpt)	Relevance Note
energy	Q11379	<i>energy</i>	“quantitative property of a physical system, recognizable in the performance of work and in the form of heat and light”	Relevant – matches the definition of a monochromatic energy value.
angle	Q11352	<i>angle</i>	“something that is formed when two rays meet at a single or same point”	Relevant – matches the definition of a monochromatic angle measurement.
iO	Q54897085	<i>I3-iPSC</i> (cell line)	No direct match; the best available entity is unrelated to intensity monitoring.	<b>Not relevant</b> – no appropriate Wikidata entry for “monitor intensity”.
itrans	Q109607869	<i>iTransmission, Inc.</i>	Company named iTransmission.	<b>Not relevant</b> – no suitable Wikidata entry for transmission intensity.
ifluor	—	—	No matching Wikidata entity found.	<b>Not relevant</b> – needs a proper link to “fluorescence intensity”.
irefer	Q58595688	<i>iReference: Using Apple’s iPhone as a Reference Tool</i>	Article by David C. Murray (2008).	<b>Not relevant</b> – no suitable Wikidata entry for reference intensity.

# “Interlink”: AI-powered CDIF variables description

Variable	Typical Physical Quantity	Common Units (SI / Conventional)	Notes
<b>energy</b>	Photon or electron energy	eV, keV, MeV, GeV	In X-ray diffraction data sets the unit is often <i>keV</i> ; for detector pixel calibration “pixel” can be used as a relative unit.
<b>angle</b>	Scattering angle ( $2\theta$ )	degrees, radians, steps	Degrees are standard in crystallography tables; “steps” may refer to detector step size when the angle is sampled discretely.
<b>i0</b>	Incident beam intensity (monitor count)	counts/s, photons/s, arbitrary units (a.u.)	If a dedicated monitor is used, the unit could be <i>counts</i> per second or simply “arbitrary” if no calibration exists.
<b>itrans</b>	Transmitted beam intensity	counts/s, photons/s, arbitrary units (a.u.)	Often expressed relative to <i>i0</i> (i.e., $\text{transmission} = \text{itrans} / i0$ ).
<b>ifluor</b>	Fluorescence emission intensity	counts/s, photons/s, arbitrary units (a.u.)	If a fluorescence detector is calibrated, it may be reported in <i>photons</i> per second.
<b>irefer</b>	Reference beam intensity (e.g., from a standard sample)	counts/s, photons/s, arbitrary units (a.u.)	Used for normalisation; typically expressed relative to <i>i0</i> or as an absolute count rate if the reference is calibrated.

# “Interlink”: Multilingual properties in Semantic Croissant: “energy”

Language	Translation
English	energy
Dutch	energie
German	Energie
French	énergie
Spanish	energía
Ukrainian	енергія (enerhiya)
Russian	энергия (energiya)
Italian	energia
Portuguese	energia
Polish	energia
Czech	energie
Slovak	energia
Greek	ενέργεια (energeia)
Swedish	energi
Norwegian	energi
Danish	energi

## Short Description

*Energy is the capacity to do work or perform tasks. It is a fundamental concept in physics and is often measured in units such as joules or kilowatt-hours. Energy can be transferred from one object to another, and can be transformed from one form to another. It is essential for powering machines, lighting homes, and powering transportation systems.*

## SKOS Relationships (Wikidata Links)

SKOS Property	Concept	Wikidata ID	Link
skos:broader	Power	Q25342	<a href="https://www.wikidata.org/wiki/Q25342">https://www.wikidata.org/wiki/Q25342</a>
skos:narrower	Kinetic energy	Q46276	<a href="https://www.wikidata.org/wiki/Q46276">https://www.wikidata.org/wiki/Q46276</a>
	Potential energy	Q155640	<a href="https://www.wikidata.org/wiki/Q155640">https://www.wikidata.org/wiki/Q155640</a>
	Thermal energy	Q114269725	<a href="https://www.wikidata.org/wiki/Q114269725">https://www.wikidata.org/wiki/Q114269725</a>
skos:related	Work	Q192581	<a href="https://www.wikidata.org/wiki/Q192581">https://www.wikidata.org/wiki/Q192581</a>
	Heat	Q60377217	<a href="https://www.wikidata.org/wiki/Q60377217">https://www.wikidata.org/wiki/Q60377217</a>
	Radiation	Q18335	<a href="https://www.wikidata.org/wiki/Q18335">https://www.wikidata.org/wiki/Q18335</a>

AI-generated concept description powered by CDIF and based on factual data (MCP)

# “Interlink”: AI “understands” CDIF variables cascade

cdi-xas / docs / CDIF\_overview\_missing.md ↑ Top

Preview Code Blame 118 lines (102 loc) · 12.8 KB Raw Copy Download Edit Menu

Property	Value
Name	X-ray fluorescence spectra for K edge, Iron c3d
Description	Metadata example based on NEXUS NXxas file.
Identifier	<a href="https://doi.org/10.xxxxx/fe_c3d001">https://doi.org/10.xxxxx/fe_c3d001</a>
Contributor	APS (Facility) – ROR: <a href="https://ror.org/aps">https://ror.org/aps</a>
License	Placeholder – “To be FAIR must include license/usage constraint information”
Distribution	One HDF5 file ( application/x-hdf5 ) at <a href="https://millenia.cars.aps.anl.gov/nxxas/MultiSpectrumFiles/FeXAS.nxs">https://millenia.cars.aps.anl.gov/nxxas/MultiSpectrumFiles/FeXAS.nxs</a>
Measurement Technique	X-ray Absorption Spectroscopy (XAS) – <a href="https://w3id.org/geochem/1.0/analyticalmethod/xrayabsorptionspectrometry">https://w3id.org/geochem/1.0/analyticalmethod/xrayabsorptionspectrometry</a>
Keywords	<ul style="list-style-type: none"><li>• K-edge (ID: <a href="https://xas.org/vocab/absorptionedge/k">https://xas.org/vocab/absorptionedge/k</a> )</li><li>• Iron (URI placeholder)</li></ul>
Variables Measured	<ol style="list-style-type: none"><li>1. nx:monochromatorEnergy (energy, eV, 7052–7380.83 eV)</li><li>2. nx:incidentIntensity (i0, counts)</li><li>3. nx:fluorescenceIntensity (ifluor, counts)</li><li>4. nx:mufluorescence (mufluor, unitless)</li></ol>
Provenance – Event	<ul style="list-style-type: none"><li>• Event ID: <i>not defined</i></li><li>• Start: 2020-08-12T04:34:49</li><li>• Instrument:<ul style="list-style-type: none"><li>- Beamline “13-ID-E” (identifier placeholder)</li><li>- Monochromator “Si311” (d-spacing 1.6375 Å)</li></ul></li><li>• Edge Energy: 7112 eV (unit placeholder)</li></ul>

AI expert system is able to recreate documentation: [CDIF reports](#)



# CDIF in Climate-Adapt4EOSC

- Collaborating on a number of projects to **implement, test, refine and extend CDIF**.
- Cluster of EC-funded projects, including **Climate-Adapt4EOSC**, looking at various case studies, including urban heat, coastal management and shrink-swell of soils.
  - CDIF for semantic and technical interoperability.
    - CDIF metadata will assist with the process of combining and integrating data for the specific use cases.
  - RO-Crates for packaging and orchestration.





# Questions

Slava Tykhonov, CODATA

[slava@codata.org](mailto:slava@codata.org)

[Listen Croissant song!](#)

Demo - Dataverse MCP and CDIF

<https://mcp.dataverse.org/>