

FAIR data management and Artificial Intelligence in research data infrastructures Croissant format for Machine Learning

Simon Hodson, Slava Tykhonov
CODATA

Workshop Re.Data 2026: Dados FAIR, Inteligência Artificial e serviços GDI em Portugal
Braga, February 9, 2026

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.

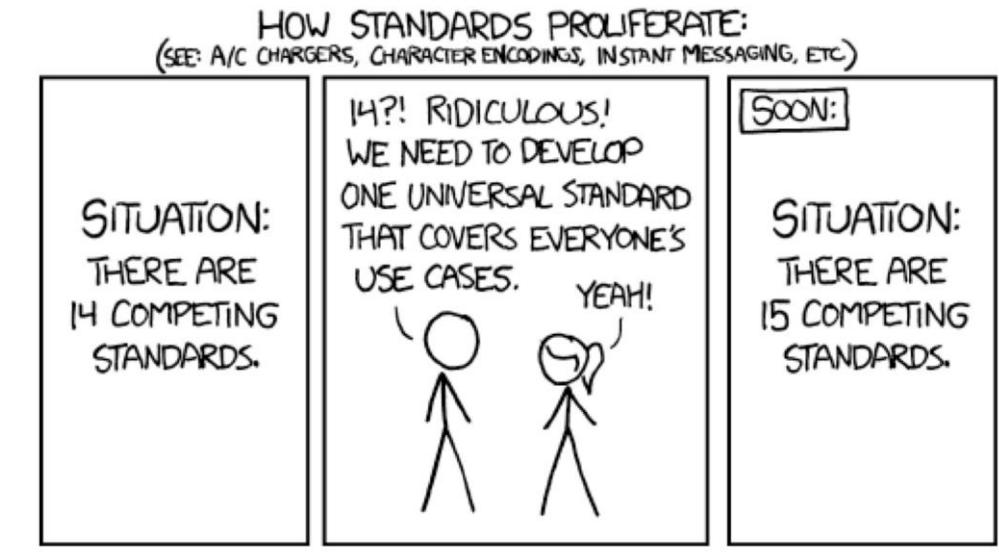
Needs to cover **Legal, Organisational, Technical, and Semantic Interoperability** (LOTS of Interoperability...)

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

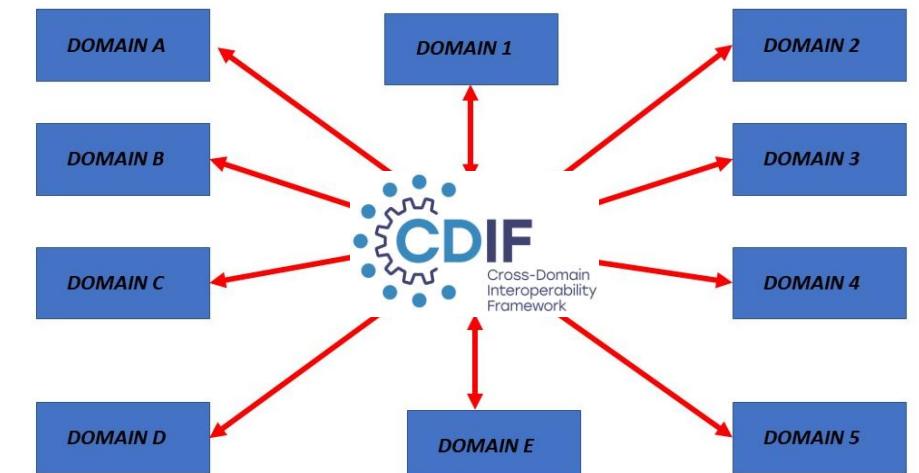


What is the CDIF (Cross-Domain Interoperability Framework)?

- 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*.**

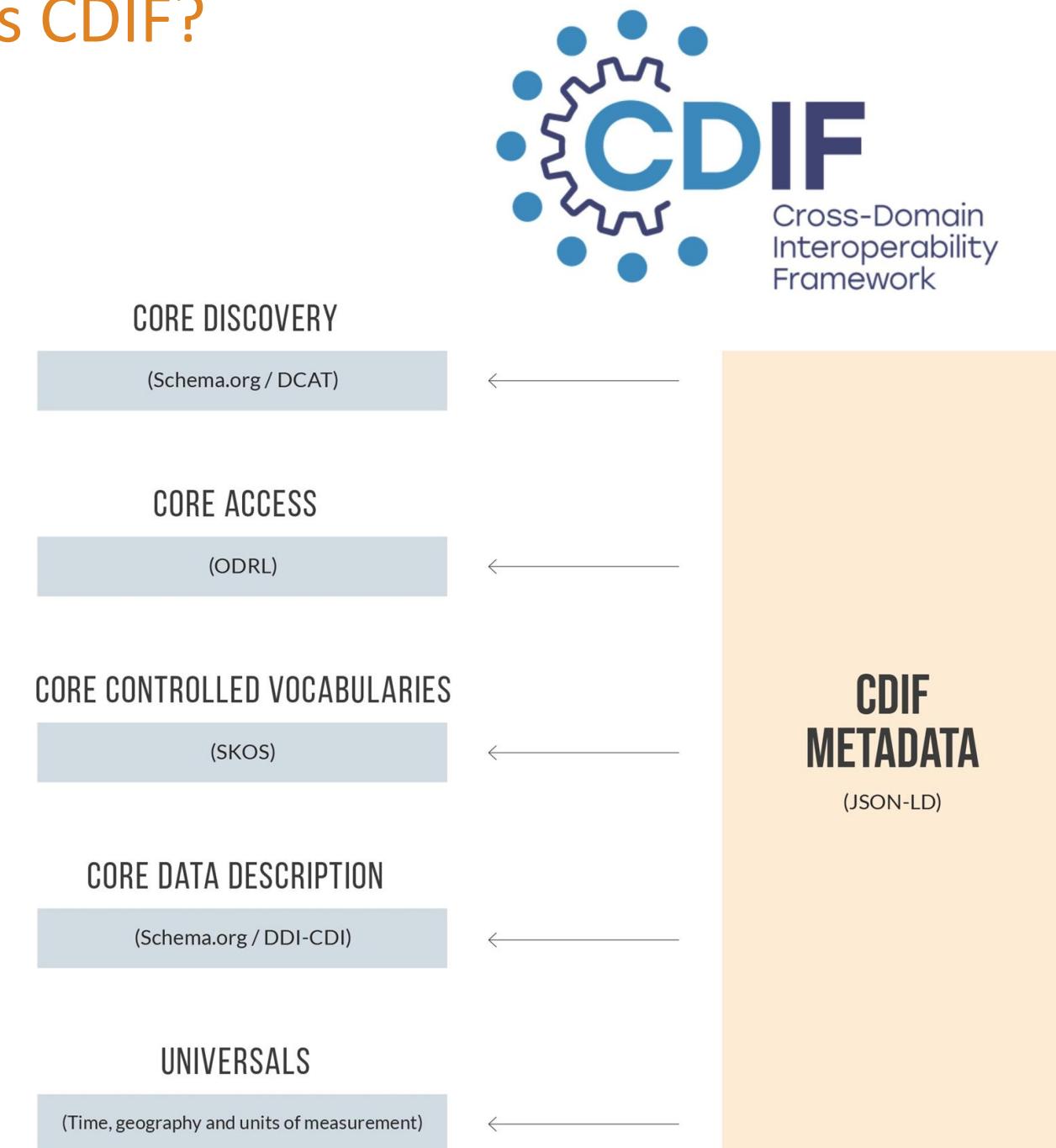


Source: xkcd.com

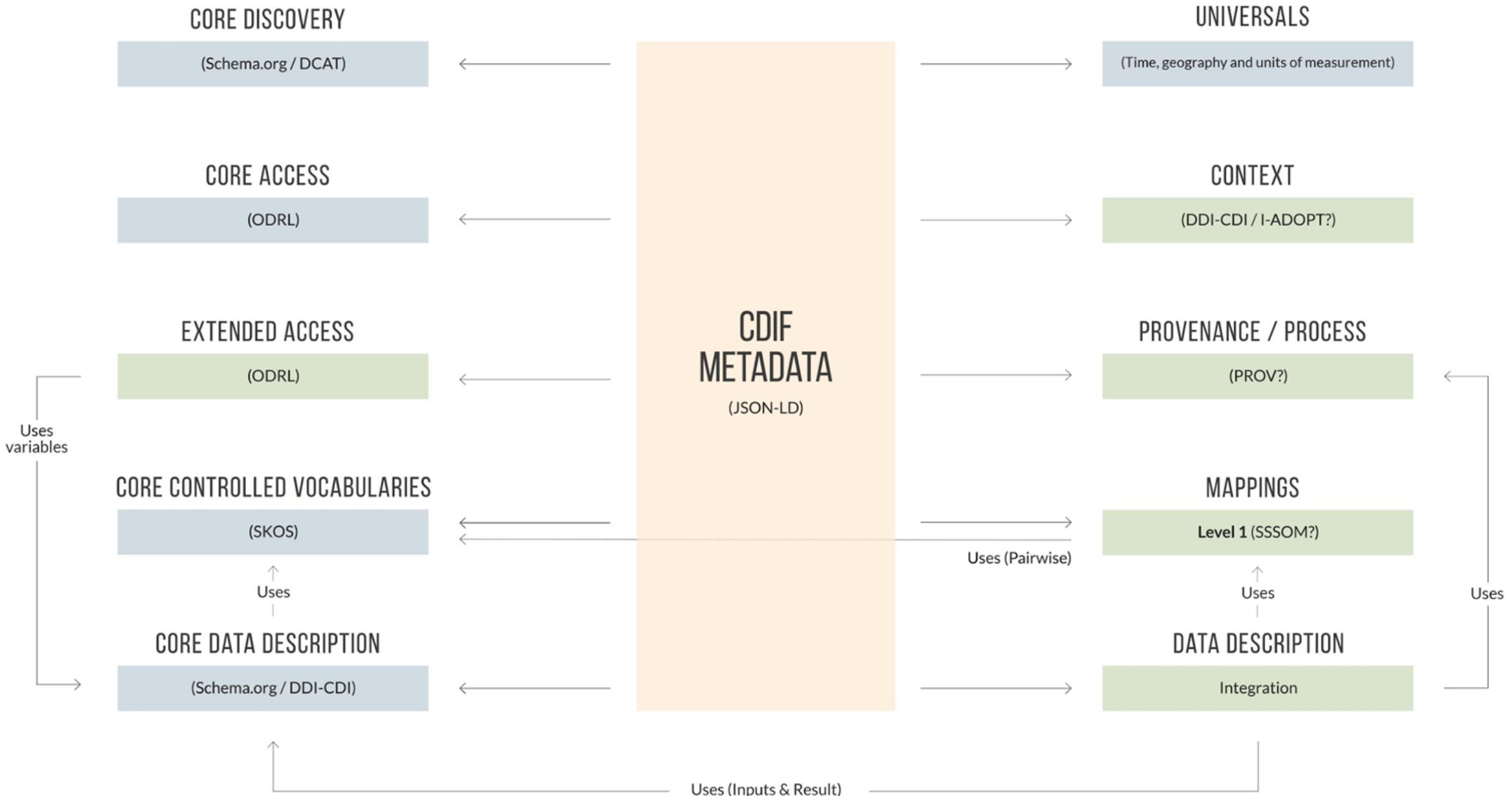


What is CDIF?

- 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>

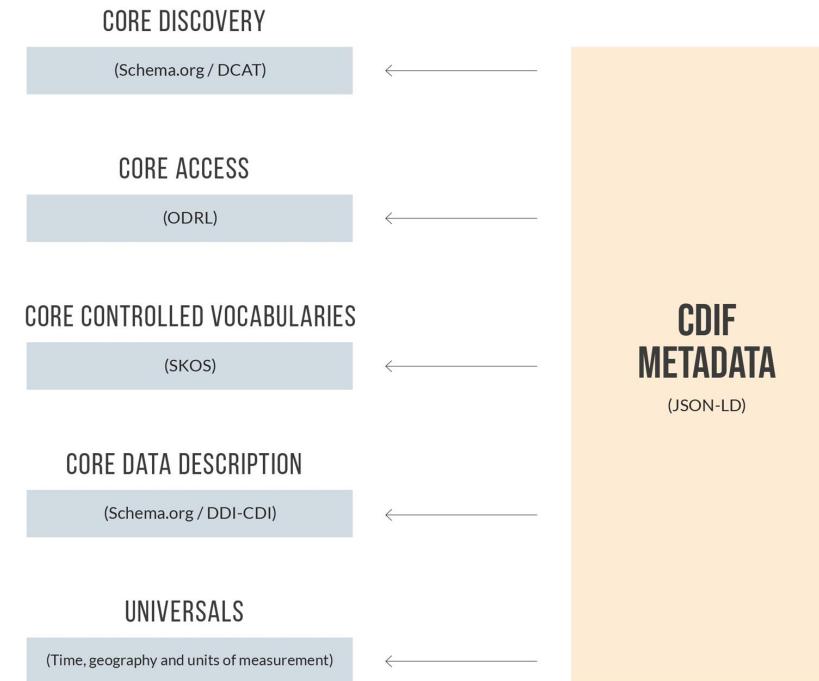


CDIF, Next Steps



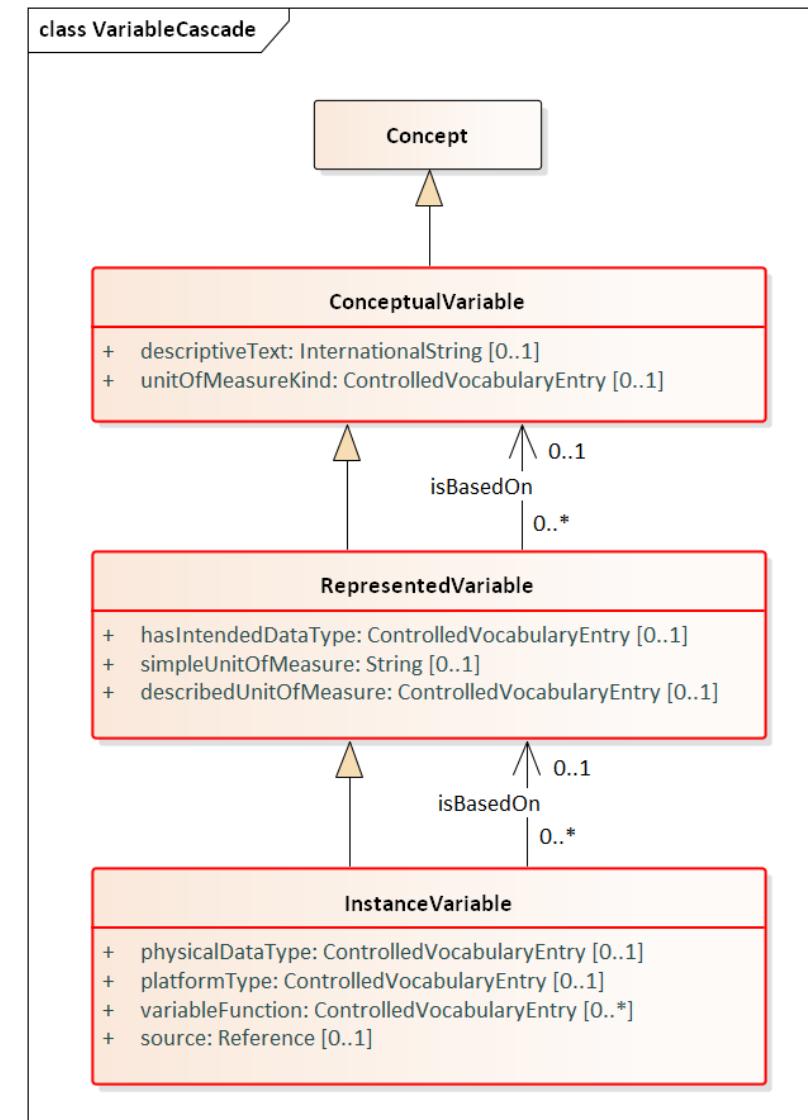
Discovery Profile

- Discovery profile: <https://bit.ly/cdif-discovery>
 - A [Content model](#) that specifies the information expected to be included in any metadata record, with required, recommended and optional content items.
 - A [JSON-LD serialization](#) for that content using the [Schema.org](#) vocabulary to define the fields in a metadata record, and [an implementation using the DCAT rdf vocabulary](#)
 - [Workflows to publish CDIF metadata](#) so that is can be found and indexed by search providers using standard web technology
- **Variable description in the discovery metadata**
 - Name of the variable as it appears in the dataset.
 - Uses schema.org variableMeasured.
 - Text description.
 - propertyID with URI for the represented concept.



Description Profile: DDI CDI for Data Structure, Variable Cascade, Provenance...

- Important to think about how we combine data for cross-domain research.
- Data Documentation Initiative (DDI) Cross-Domain Integration (CDI) specification contains three modules to assist with this:
 - **Structural Description:** assists processing of data structure transformations across four data structures.
 - **Data Description / Variable Cascade** describes data at an atomic level, describes relationships between concepts, representations and instances (assists with combining data and documenting information loss).
 - **Provenance and Processing:** module uses PROV-O and SDTL to provide and relay provenance and processing information.
 - Now officially released: <https://ddialliance.org/ddi-cdi>



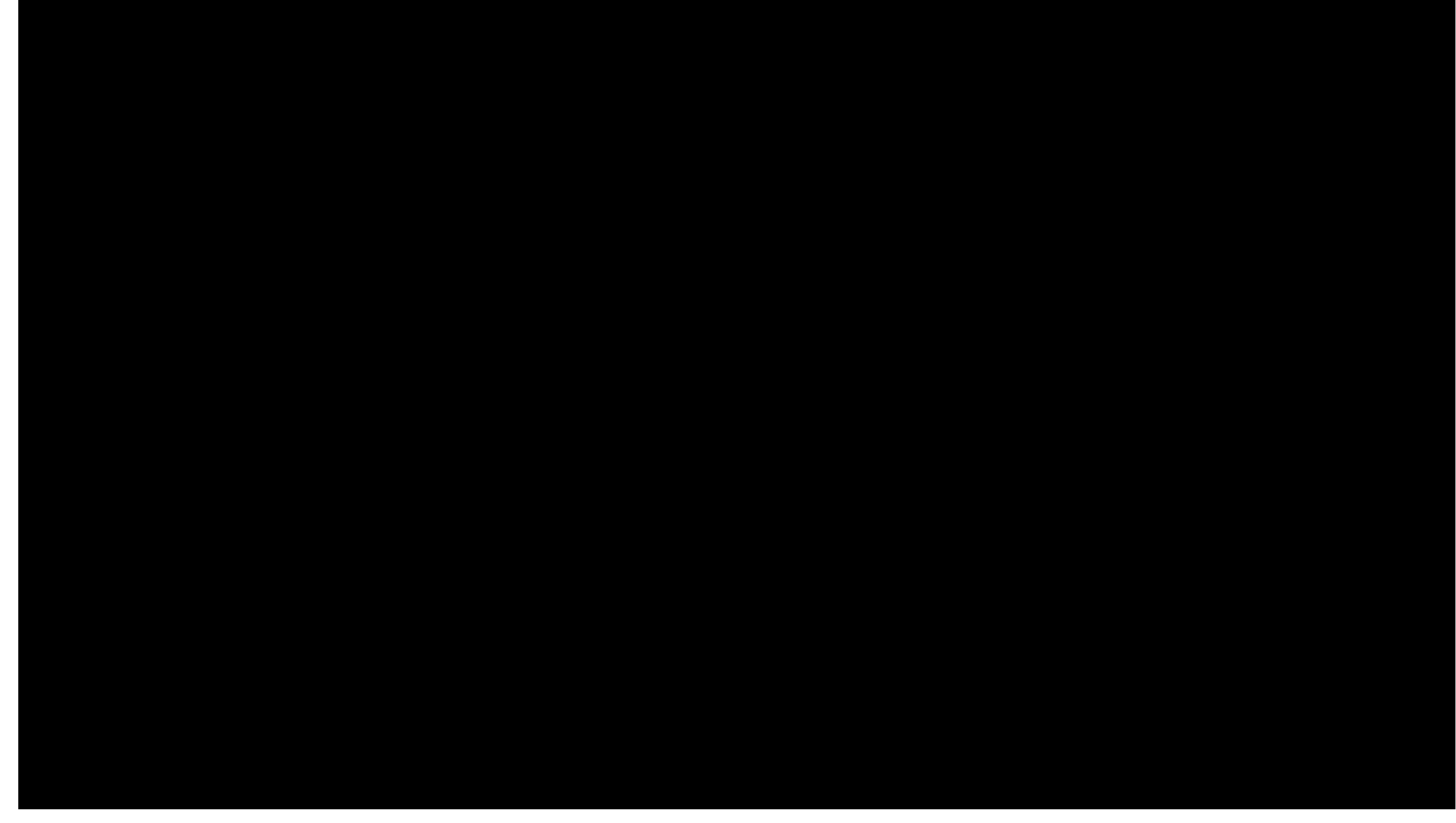
How AI can help with interoperability?

Human resources are very **expensive** and **deficit**, it's usually difficult to find appropriate expertise in-house.

Possible solutions:

- Building and fine-tuning AI/ML pipelines for the automatic metadata enrichment and linkage prediction, and verification with human experts.
- applying NLP for NER, data mining, topic classification etc - with human verification.
- building multidisciplinary knowledge graphs should facilitate the development of new projects with various research communities to facilitate and motivate them taking ownership of their own data as soon as they see the added value in AI-powered solutions.

But first, we need to start production of AI-ready data...



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:

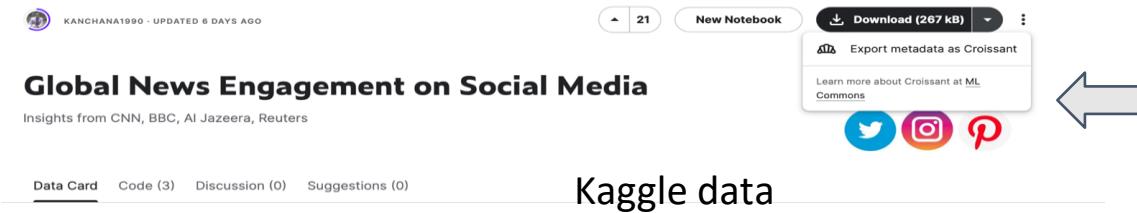
- Omar Benjelloun (Google),
- Elena Simperl (King's College London & ODI),
- Pierre Marcenac (Google),
- Pierre Ruyssen (Google),
- Costanza Conforti (Google),
- Michael Kuchnik (Meta),
- Jos van der Velde (Open ML),
- Luis Oala (Dotphoton),
- Steffen Vogler (Bayer),
- Mubashara Akthar (King's College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



Global News Engagement on Social Media

Insights from CNN, BBC, Al Jazeera, Reuters

Data Card Code (3) Discussion (0) Suggestions (0)

Kaggle data

About Dataset

This comprehensive dataset offers a deep dive into the social media engagement metrics of nearly 4,000 posts from four of the world's leading news channels: CNN, BBC, Al Jazeera, and Reuters. Curated to provide a holistic view of global news interaction on social media, the collection stands out for its meticulous assembly and broad spectrum of content.

Dataset Overview:

Spanning various global events, topics, and narratives, this dataset is a snapshot of how news is consumed and interacted with on social media platforms. It serves as a rich resource for analyzing trends, engagement patterns, and the dissemination of information across international borders.

Data Science Applications:

Ideal for researchers and enthusiasts in the fields of data science, media studies, and social analytics, this dataset opens doors to numerous explorations such as engagement analysis, trend forecasting, content strategy optimization, and the study of information flow in digital spaces. It also holds potential for machine learning projects aiming to predict engagement or classify content based on interaction metrics.

Column Descriptors:

Each record in the dataset is detailed with the following columns:

Usability ⓘ
10.00

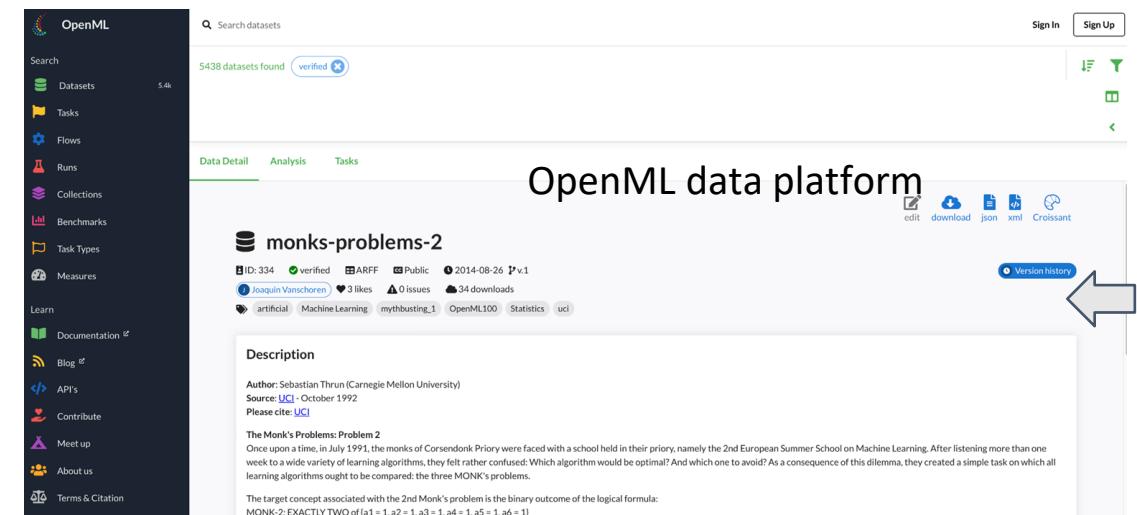
License
ODC Attribution License (ODC-...)

Expected update frequency
Never

Tags
Online Communities
News Social Networks
People and Society
Data Storytelling

Sign In Sign Up

OpenML



Search datasets

5438 datasets found verified

monks-problems-2

ID: 334 verified ARFF Public 2014-08-26 v.1

joaquin Vanschoren 3 likes 0 issues 34 downloads

artificial Machine Learning mythbusting_1 OpenML100 Statistics uci

Description

Author: Sebastian Thrun (Carnegie Mellon University)
Source: UCI - October 1992
Please cite: UCI

The Monks' Problems: Problem 2
Once upon a time, in July 1991, the monks of Corsondon Priory were faced with a school held in their priory, namely the 2nd European Summer School on Machine Learning. After listening more than one week to a wide variety of learning algorithms, they felt rather confused: Which algorithm would be optimal? And which one to avoid? As a consequence of this dilemma, they created a simple task on which all learning algorithms ought to be compared: the three MONK's problems.

The target concept associated with the 2nd Monk's problem is the binary outcome of the logical formula:
MONK-2: EXACTLY TWO of $\{a_1 = 1, a_2 = 1, a_3 = 1, a_4 = 1, a_5 = 1, a_6 = 1\}$

edit download json xml Croissant Version history

OpenML data platform

←

10

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.

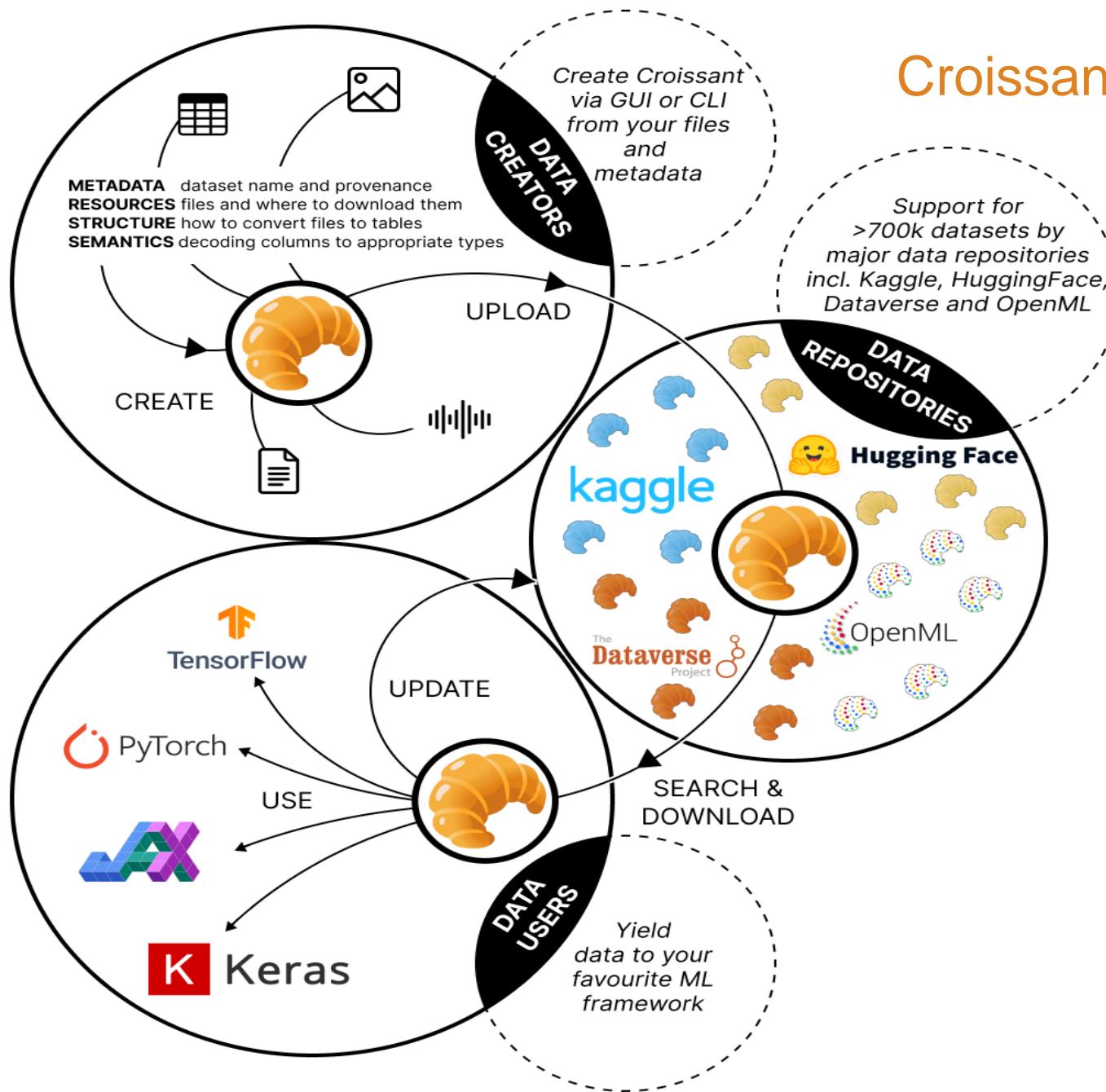
What makes Datasets used in Machine Learning special?

- 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



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>

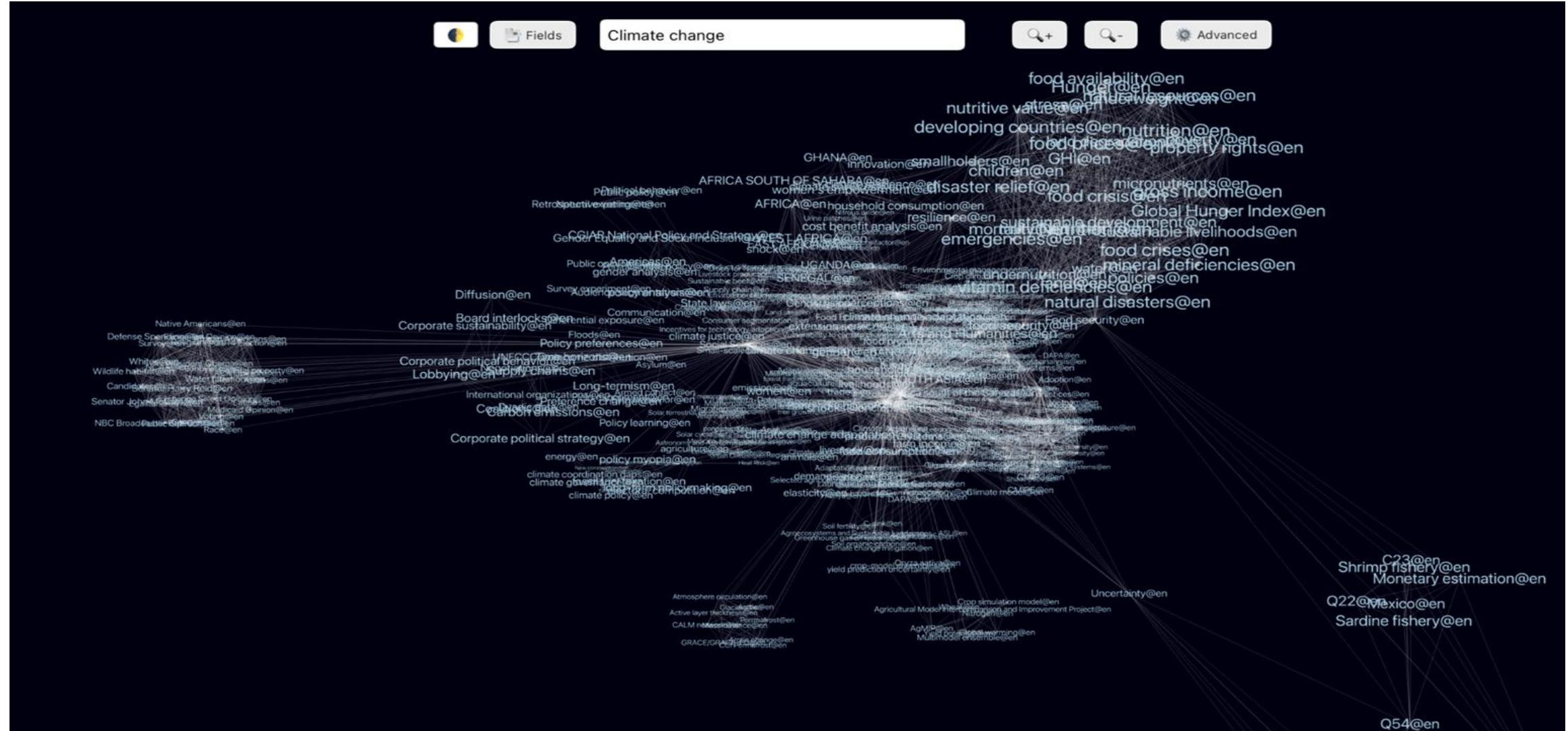
Discover and find

- Google Dataset Search
 - <https://datasetsearch.research.google.com/>
- Crawler in repo
 - https://github.com/mlcommons/croissant/blob/main/health/visualizer/report_huggingface.ipynb

Use it

- Colab
 - https://githubtocolab.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

Croissant Graph in Dataverse network



55 million Croissant triples in Qlever triple store. Want to try? <https://graph.codata.org>

Croissant layers in detail

- Dataset-level metadata
 - Based on 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": " | "}  
      }  
    }  
  ]  
}
```

```
{  
  "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": "#{splitEnums/name}"  
  }
```

```
name, url  
train, https://mlcommons.org/definitions/training_split  
val, https://mlcommons.org/definitions/validation_split  
test, https://mlcommons.org/definitions/test_split
```

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

Dataverse backbone is Application Program Interface (API)

Data Access API

The Data Access API provides programmatic download access to the files stored in a Dataverse installation. More advanced features of the Access API include format-specific transformations (thumbnail generation/resizing for images; converting tabular data into alternative file formats) and access to the data-level metadata that describes the contents of the tabular files.

Contents:

- [Downloading All Files in a Dataset](#)
 - [Basic Download By Dataset](#)
 - [Download By Dataset By Version](#)
- [Basic File Access](#)
 - [Parameters:](#)
 - [Headers:](#)
 - [Examples](#)
- [Multiple File \("bundle"\) download](#)
 - [Parameters:](#)
- ["All Formats" bundled download for Tabular Files.](#)
 - [Parameters:](#)
- [Data Variable Metadata Access](#)
 - [Parameters:](#)
- [Preprocessed Data](#)
- [Authentication and Authorization](#)
- [Access Requests and Processing](#)
 - [Allow Access Requests:](#)
 - [Request Access:](#)
 - [Grant File Access:](#)
 - [Reject File Access:](#)
 - [Revoke File Access:](#)
 - [List File Access Requests:](#)
 - [User Has Requested Access to a File:](#)
 - [Get User Permissions on a File:](#)

Downloading All Files in a Dataset

The "download by dataset" API downloads as many files as possible from a dataset as a zipped bundle.

By default, tabular files are downloaded in their "archival" form (tab-separated values). To download the original files (Stata, for example), add `format=original` as a query parameter.

There are a number of reasons why not all of the files can be downloaded:

- Some of the files are restricted and your API token doesn't have access (you will still get the unrestricted files).
- The Dataverse installation has limited how large the zip bundle can be.

Search API

Contents:

- [Parameters](#)
- [Basic Search Example](#)
- [Advanced Search Examples](#)
 - [Narrowed to Collection, Show Relevance and Facets](#)
 - [Retrieve Released Versions Only](#)
 - [Include Metadata Blocks and/or Metadata Fields](#)
 - [Include Specific Fields Only](#)
- [Date Range Search Example](#)
- [Iteration](#)
- [Search Services](#)
 - [Discovering Available Search Services](#)
 - [Using Different Search Services](#)
 - [Developing Search Services](#)

The Search API supports the same searching, sorting, and faceting operations as the Dataverse Software's web interface.

To search unpublished content, you must pass in an API token as described in the [API Tokens and Authentication](#) section.

The parameters and JSON response are partly inspired by the [GitHub Search API](#).

Note

The search API can be used from scripts running in web browsers, as it allows cross-origin resource sharing (CORS).

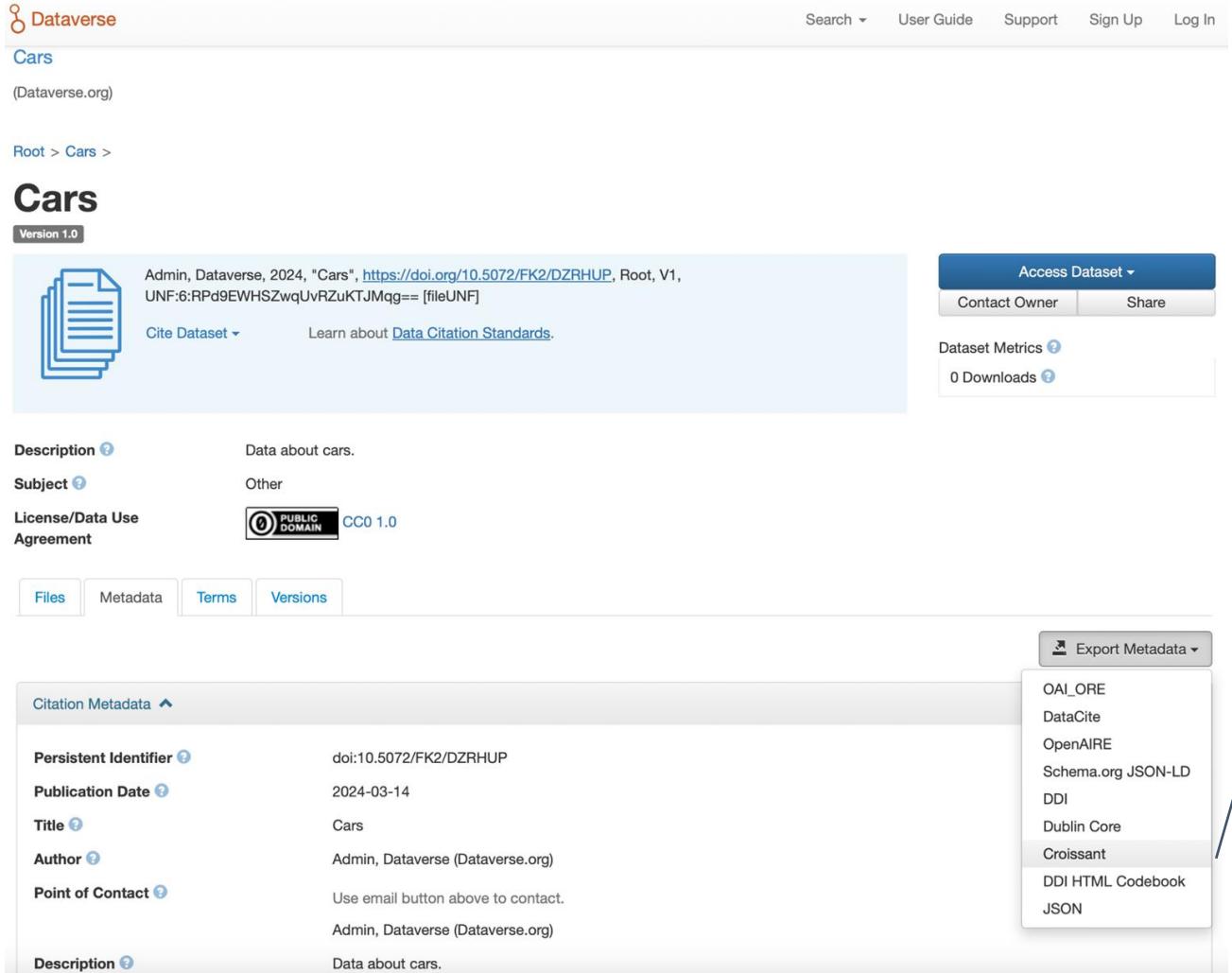
CORS

Please note that in Dataverse Software 4.3 and older the "citation" field wrapped the persistent ID URL in an `<a>` tag but this has been changed to plaintext. If you want the old value with HTML in it, a new field called "citationHtml" can be used.

Parameters

Name	Type	Description
<code>q</code>	string	The search term or terms. Using "title:data" will search only the "title" field. "*" can be used as a wildcard either alone or adjacent to a term (i.e. "bird*"). For example, https://demo.dataverse.org/api/search?q=title:data . For a list of fields to search, please see https://github.com/QSS/dataverse/issues/2558 (for now).
<code>type</code>	string	Can be either "dataverse", "dataset", or "file". Multiple "type" parameters can be used to include multiple types (i.e. <code>type=dataset&type=file</code>). If omitted, all types will be returned. For example, https://demo.dataverse.org/api/search?q=&type=dataset

Croissant ML export in 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](#).

Description Data about cars.

Subject Other

License/Data Use Agreement  CC0 1.0

Files Metadata Terms Versions

Citation Metadata ▾

Persistent Identifier doi:10.5072/FK2/DZRHUP

Publication Date 2024-03-14

Title Cars

Author Admin, Dataverse (Dataverse.org)

Point of Contact Use email button above to contact.
Admin, Dataverse (Dataverse.org)

Description Data about cars.

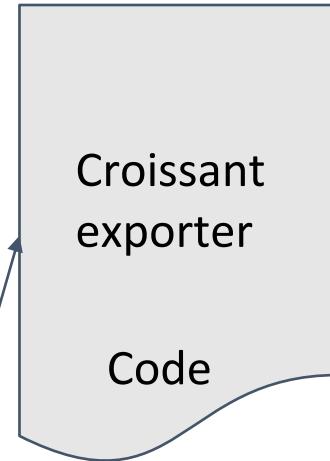
Access Dataset ▾ Contact Owner Share

Dataset Metrics 0 Downloads

Export Metadata ▾

- OAI_ORE
- DataCite
- OpenAIRE
- Schema.org JSON-LD
- DDI
- Dublin Core
- Croissant**
- DDI HTML Codebook
- JSON

Mappings



```
  @context:
    @language: "en"
    @vocab: "https://schema.org/"
    citeAs: "cr:citeAs"
    column: "cr:column"
    conformsTo: "dct:conformsTo"
    cr:
    data:
      @id: "cr:dataset"
      @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: "cr:type"
      conformsTo: "http://mlcommons.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: "7b1201ce6b469796837a835377338c5a"
        contentUrl: "stata13-auto.dta"
      recordSet:
        @ 0:
          @type: "cr:recordSet"
          fields:
            @ 0:
              @type: "cr:field"
              name: "make"
              description: "Make and Model"
              dataType: "sc:text"
              source:
                @id: "6"
                fileObject:
                  @id: "stata13-auto.dta"
```

Semantic Croissant alignment with Retrieval Augmented Generation (RAG)

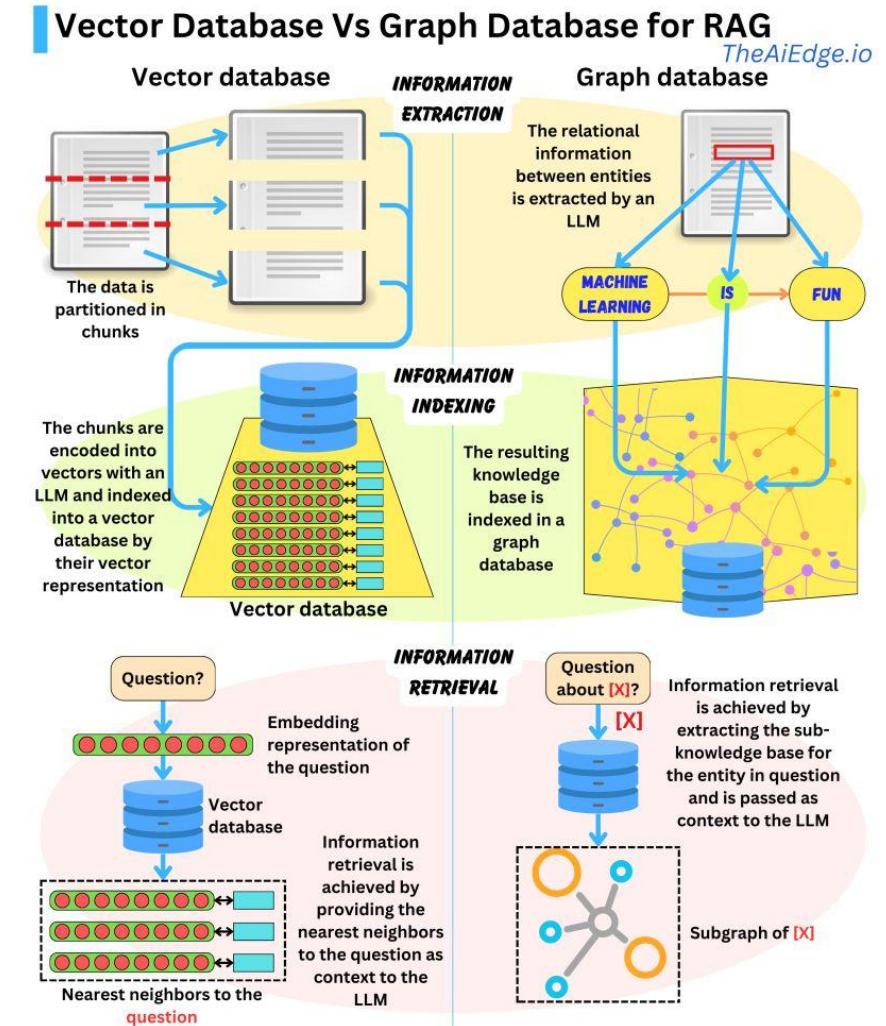
Semantic Croissant ML is knowledge graph (*metadata layer*) integrated with domain-specific **ontologies** and **controlled vocabularies** including responsible AI. This graph is dynamic and sustainable (FAIR).

Vector is the *content* of data files encoded in embeddings, with properties and their attributes. Embeddings computed by various ML algorithms and use different models.

Distributed AI: Croissant-powered knowledge graph is “navigation layers” and new model (vectors) are trained on distributed resources.

MCP is “interface” between human and AI, and works as a “reasoning engine” to provide both answers and new resources.

Implementation: LLM connected to the “**RAG library**” (graph), navigate through datasets and consuming embeddings (vectors) as a context.





Why Semantic Croissant?

RAG: LLM + KG

Response body

```
{  
  "analysis": {  
    "model": "llama3.2-vision",  
    "created_at": "2025-04-14T13:47:01.124268785Z",  
    "done": true,  
    "done_reason": "stop",  
    "total_duration": 4325784046,  
    "load_duration": 282833423,  
    "prompt_eval_count": 25,  
    "prompt_eval_duration": 1599593264,  
    "eval_count": 129,  
    "eval_duration": 2225709129,  
    "message": {  
      "role": "assistant",  
      "content": "The image shows the members of Queen, a British rock band, in a recording studio setting. \n\n* The drummer, Roger Taylor, is wearing a white shirt and black pants, while the guitarist, Brian May, is wearing a black jacket and jeans.\n* Freddie Mercury, the lead singer, is wearing a red jacket and black pants, and John Deacon, the bassist, is wearing a white shirt and black pants.  
The image suggests that the band is in the process of recording an album or single. The formal attire and professional setting indicate that they are taking their music seriously and are committed to producing high-quality recordings.",  
      "images": null,  
      "tool_calls": null  
    }  
  }  
}
```



Download

LLM concepts as cells interlinked with knowledge graph entities

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.



ODISSEI



AI assistants for human experts - ontology alignment

Use case 1: AI-powered mappings for ontologies in Semantic Croissant

Various metadata schemes:

- Dataverse
- DataCite
- HuggingFace
- Kaggle
- OpenML
- CodeMeta
- Many more

Can we build automated AI service to make predictions on how metadata fields should be mapped in Semantic Croissant with SKOS relationships?

Use case: Codemeta to DataCite crosswalks



<https://codemeta.github.io/crosswalk/datacite/>

Property	DataCite
programmingLanguage	Format
fileSize	size
author	creators
contributor	contributor
dateCreated	date
dateModified	date
datePublished	publicationYear

Ontologies prediction for Semantic Croissant

Semantic crosswalks: publicationYear in DataCite, datePublished in CodeMeta

Input: "publicationYear: The year when the data was or will be made publicly available."

*Based on the provided facts, the most relevant property is ****datePublished****.*

Here's why:

* *****datePublished**** is explicitly defined as "Date of first broadcast/publication." This aligns directly with the request for the date when data becomes publicly available.*

Let's look at why the other properties are less suitable:

- * *****dateCreated:**** This refers to the creation date of the work, not its public availability.*
- * *****dateModified:**** This is about changes ***after*** the initial publication, not the initial release date.*

Use case 2: automated linkage of values with controlled vocabularies

- Wikidata
- OntoPortal ontologies or other semantic artefacts
- SKOSMOS vocabularies
- Getty
- [SPDX licenses](#)
- Linked Open vocabularies (<https://lov.linkeddata.es/dataset/lov>)
- Geospatial Ontologies for GeoCroissant
- [GeoNames](#)
- [SWEET Ontology](#)
 - Interoperability with advanced semantic web tools such as SHACL or [R2RML](#)
- Biological / Healthcare
 - MeSH Terms
 - BioPortal (OntoPortal) / OLS / OntoBee / OBO
 - OHDSI OMOP / HL7 FHIR
 - SNOMED-CT, ICD, RxNorm, CDISC, FMA, LOINC, HPO, MONDO, UMLS, GO
- Governance related
 - GA4GH DUO
 - [Open Digital Rights Language](#) (ODRL)
 - [Data Privacy Vocabulary](#) (DPV)
- Software Supply Chain
 - [System Data Package Exchange](#) (SPDX)

KG+LLM integrator: automated concept prediction/1

Goal: find wikidata concept for song Hotel California
Query: Hotel California **Context:** Song by Eagles

GET /wikilink/ Get Wikilink

Parameters

Name	Description
term * required	Hotel California
context * required	song by Eagles
property	property
language	en
source	source
format	json

Cancel

Execute Clear

Responses

Curl

```
curl -X 'GET' \
  'https://sparqlmuse.now.museum/wikilink/?term=Hotel%20California&context=song%20by%20Eagles&language=en&format=json' \
  -H 'accept: application/json'
```

Request URL

<https://sparqlmuse.now.museum/wikilink/?term=Hotel%20California&context=song%20by%20Eagles&language=en&format=json>

Server response

Code Details

200 Response body

```
{
  "title": "Q780394",
  "label": "Hotel California",
  "description": "song written and composed by Don Felder, Don Henley and Glenn Frey; originally recorded by Eagles and released 1976",
  "url": "https://www.wikidata.org/wiki/Q780394"
}
```

Download

```
{
  "title": "Q780394",
  "label": "Hotel California",
  "description": "song written and composed by Don Felder, Don Henley and Glenn Frey; originally recorded by Eagles and released 1976",
  "url": "https://www.wikidata.org/wiki/Q780394"
}
```

↓

Hotel California (Q780394)

song written and composed by Don Felder, Don Henley and Glenn Frey; originally recorded by Eagles and released 1976 [edit](#)

▼ In more languages [Configure](#)

Language	Label	Description	Also known as
English	Hotel California	song written and composed by Don Felder, Don Henley and Glenn Frey; originally recorded by Eagles and released 1976	
Dutch	Hotel California	single van Eagles	Hotel California (single)
German	Hotel California	Lied der US-amerikanischen Band Eagles	
French	Hotel California	chanson des Eagles, sortie en 1977	

KG+LLM integrator: automated concept prediction/2

Goal: find wikidata concept for movie Hotel California about criminal

Query: Hotel California **Context:** criminal movie

GET /wikilink/ Get Wikilink

Parameters

Name	Description
term * required	Hotel California
context * required	criminal movie
property	property
language	en
source	source
format	json

Execute Clear

Responses

Curl

```
curl -X "GET" \
  "https://sparqlmuse.now.museum/wikilink/?term=Hotel%20California&context=criminal%20movie%20&language=en&format=json" \
  -H "Accept: application/json"
```

Request URL

```
https://sparqlmuse.now.museum/wikilink/?term=Hotel%20California&context=criminal%20movie%20&language=en&format=json
```

Server response

Code Details

200 Response body

```
{ "title": "Q4509386", "label": "Hotel California", "description": "2008 film by Geo Santini", "url": "https://www.wikidata.org/wiki/Q4509386" }
```

Download

```
{ "title": "Q4509386", "label": "Hotel California", "description": "2008 film by Geo Santini", "url": "https://www.wikidata.org/wiki/Q4509386" }
```



Hotel California (Q4509386)

2008 film by Geo Santini

edit

▼ In more languages

Configure

Language	Label	Description	Also known as
English	Hotel California	2008 film by Geo Santini	
Dutch	Hotel California	film uit 2008	Hotel California (film)
German	Hotel California	Film von 2008	
French	Hotel California	film de Geo Santini, sorti en 2008	

KG+LLM integrator: automated concept prediction/3

Goal: find wikidata concept for 2013 movie Hotel California
Query: Hotel California **Context:** movie from 2013

GET /wikilink/ Get Wikilink

Parameters

Name	Description
term * required string (query)	Hotel California
context * required string (query)	movie from 2013
property string (query)	property
language string (query)	en
source string (query)	SOURCE
format string (query)	json

Cancel

Execute Clear

Responses

Curl

```
curl -X 'GET' \
  'https://sparqlmuse.now.museum/wikilink/?term=Hotel%20California&context=movie%20from%202013%20&language=en&format=json' \
  -H 'Accept: application/json'
```

Request URL

```
https://sparqlmuse.now.museum/wikilink/?term=Hotel%20California&context=movie%20from%202013%20&language=en&format=json
```

Server response

Code	Details
200	Response body

```
{
  "title": "Q5911306",
  "label": "Hotel California",
  "description": "2013 film by Aji John",
  "url": "https://www.wikidata.org/wiki/Q5911306"
}
```

Download

{
"title": "Q5911306",
"label": "Hotel California",
"description": "2013 film by Aji John",
"url":
"https://www.wikidata.org/wiki/Q5911306 "
}



Hotel California (Q5911306)

2013 film by Aji John

edit

▼ In more languages

Configure

Language	Label	Description	Also known as
English	Hotel California	2013 film by Aji John	
Dutch	No label defined	film uit 2013	
German	Hotel California	Film von 2013	
French	Hotel California	film de Aji John, sorti en 2013	

How to control Artificial Intelligence



Problem:

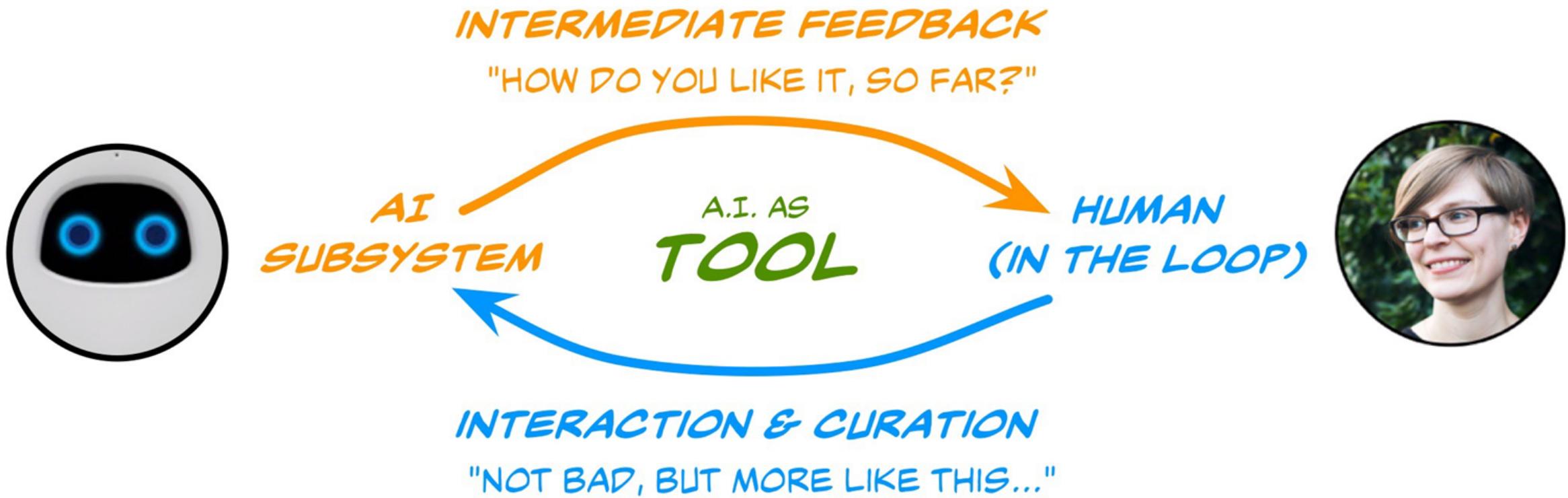
It's naive to fully trust Machine Learning and AI, we need to support a **"human in the loop"** processes to take a control over automatic workflows. Ethics is also important, fake detection problem.

Solution:

A lot of "human in the loop" tools already developed in research projects, we need to support the best for the different use cases, add the appropriate maturity, for example, with CI/CD and introduce them to research communities.

Annotations as “Human in the loop” - domain specific

“how do we build a smarter system?” to “how do we incorporate useful, meaningful human interaction into the system?”



General blueprint for a human-in-the-loop interactive AI system. Credits: [Stanford University HAI](#)

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

 Dataverse

Add Data ▾ Search ▾ User Guide Support Sign Up Log In

Root >

Earthquake dataset

Version 1.0

User, Test, 2025, "Earthquake dataset", <https://doi.org/10.5072/FK2/BXSHPO>, Root, V1, UNF:6:bkDZ89gr9xKH!GoW5hkZw== [fileUNF]

 Cite Dataset ▾ Learn about Data Citation Standards.

Description: Earthquake data from Kaggle

Subject: Astronomy and Astrophysics

License/Data Use Agreement:  CC0 1.0

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    

Dataset Information Variables Export documentation Mappings Endpoint Variable cascade

Name	Label	Definition	Type	Coded	Details
0 time	time	Time is the continuous, unidirectional flow of events	String / Text		
1 latitude	latitude	The angular coordinate that specifies the position of a point in a two-dimensional coordinate system	Decimal		
2 longitude	longitude	A geographic coordinate that specifies the position of a point in a two-dimensional coordinate system	Decimal		
3 depth	depth	Depth is the vertical distance from a horizontal reference point	Decimal		
4 mag	mag	Magnitude (mag) is a quantitative measure of the size of an earthquake	Decimal		
5 magType	magType	The magnitude type is a categorical variable	String / Text		
6 nst	nst	The variable 'nst' represents the **N	Decimal		
7 gap	gap	A gap is the measurable space or separation between two points	Decimal		
8 dmin	dmin	The minimum spatial or temporal separation between two events	Decimal		
9 rms	rms	Root Mean Square (RMS) is a statistical measure of the magnitude of a varying quantity	Decimal		
10 net	net	A derived numeric quantity that represents the net effect of a series of events	String / Text		
11 id	id	The measure of the warmth or coldness of an object	String / Text		
12 updated	updated	The measure of the average kinetic energy of the particles in a system	String / Text		
13 place	place	A variable that captures the geographical location of an event	String / Text		
14 type	type	The ambient air temperature measured at a specific location	String / Text		

Demo

32

Semantic Croissant in the Cross-domain interoperability framework (CDIF)



Contents

Introducing CDIF
What is CDIF
Who can use the CDIF?
The CDIF workgroup
How to contribute

Introducing CDIF

Many important research questions demand a multi-disciplinary approach in which data and resources are used across domain and infrastructure boundaries. In such scenarios, domain-specific community standards fall short of the requirements for FAIR exchange of the critical metadata and other information needed. The **Cross-Domain Interoperability Framework (CDIF)** is designed to support FAIR implementation for these projects by establishing a 'lingua franca' for this information, based on existing standards and technology to support interoperability, in both human- and machine-actionable fashion. CDIF is a set of implementation recommendations, based on profiles of common, domain-neutral metadata standards which are aligned to work together to support core functions required by FAIR.

The idea for CDIF first emerged from workshops and discussions at conferences prior to the WorldFAIR project, beginning in 2018. The WorldFAIR project provided an opportunity to advance that vision, through a set of 11 case studies across many domains, allowing the needs and practices around FAIR within such domains to be summarised in the form of FAIR Implementation Profiles (FIPs). Based on the FIPs and focused meetings, the requirements for CDIF were established. A group of 30 invited experts from different FAIR initiatives and standards bodies made up a Working Group and an Advisory Group to synthesise the findings from WorldFAIR and to produce the current CDIF draft.

The framework is based on a set of five core profiles that address the most important functions for cross-domain FAIR implementation.

1. Discovery (patterns for metadata content, serialization and publication)
2. Data access (documentation of access conditions and permitted use)
3. Controlled vocabularies (practices for the publication of controlled vocabularies and semantic artefacts)
4. Data integration (documentation of the structural and semantic aspects of data to make it integration-ready)

Responsible AI: Croissant and DDI

(Data Documentation Initiative)



CESSDA DATA CATALOGUE (CDC) DDI3.3 PROFILE

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 [Croissant spec v1.0](#). 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).“

DDI_XPath	Required	UI Label	Type	Repeatable	Usage note
/ddi:DDIInstance/@xml: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@Xml: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"). (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@Xml: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:CreatorRef	Recommended	Creator	Container element	Yes	Reference to Principal investigator Person OR Institution
/ddi:DDIInstance/s:StudyUnit/r:Citation/r:Creator/r:CreatorRef/r:TypeOfObject	Mandatory if '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:Publisher	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="v63820" name="gender" intrvl="contin">
  <location fileid="f101579"/>
  <><catgr>
    <catValue1><catValue>
      <lbl level="category">male, or</lbl>
    </catValue>
    <catValue2><catValue>
      <lbl level="category">RF</lbl>
    </catValue>
    <catValue3><catValue>
      <lbl level="category">female</lbl>
    </catValue>
  </catgr>
  <notes subject="Universal Numeric Fingerprint" level="variable" type="Dataverse:UNF">UNF:3:IMDNaBjEgQqogNVuPI5KwA==</notes>
</var>
<var ID="v63998" name="marit_a" intrvl="contin">
  <location fileid="f101579"/>
  <><catgr>
    <catValue1><catValue>
      <lbl level="category">no</lbl>
    </catValue>
    <catValue2><catValue>
      <lbl level="category">RF</lbl>
    </catValue>
    <catValue3><catValue>
      <lbl level="category">K</lbl>
    </catValue>
    <catValue4><catValue>
      <lbl level="category">yes</lbl>
    </catValue>
  </catgr>
  <notes subject="Universal Numeric Fingerprint" level="variable" type="Dataverse:UNF">UNF:3:oueQkfY5y0dpXf9yYmIGM0==</notes>
</var>
<var ID="v63876" name="marit_b" intrvl="contin">
  <location fileid="f101579"/>
  <><catgr>
    <catValue1><catValue>
      <lbl level="category">divorced, or</lbl>
    </catValue>
    <catValue2><catValue>
      <lbl level="category">never married</lbl>
    </catValue>
    <catValue3><catValue>
      <lbl level="category">widowed</lbl>
    </catValue>
  </catgr>
```

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.

	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	

Items Per Page: 5 1 - 7 of 21

Edit Variable

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:M6BSZcTBShsqNvwvlefw==

Group(s)
Add to Group

No Weight Assigned Is Weight

Responsible AI in Croissant?

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

Raw

```
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/Data Type/1.1.2/#String"
9      },
10     {
11       "@id": "#Year",
12       "@type": "InstanceVariable",
13       "name": "Year",
14       "hasIntendedDataType": "http://rdf-vocabulary.ddialliance.org/cv/Data Type/1.1.2/#Integer"
15     },
16     {
17       "@id": "#Age_Cohort",
18       "@type": "InstanceVariable",
19     }
20   ]
21 }
```

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

Dataset in Dataverse represented in CDIF

 **Dataverse**

Add Data ▾ Search ▾ User Guide Support Sign Up Log In

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/FR2/4ZSKVU>, Root, V3

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

 **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... 

Add Data ▾ Search ▾ User Guide Support Sign Up Log In

Dataverse

Files Metadata Terms Versions

Search this dataset... 

Filter by File Type: All ▾ Access: All ▾

1 to 2 of 2 Files

		Download
<input type="checkbox"/>	Se_Na2SeO4_rt_01.xdi Unknown - 18.3 KB Published Nov 19, 2025 76 Downloads MD5: 012...757 	
<input type="checkbox"/>	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	
2	Beamline	"prov:used": [{ "@type": "schema:Thing", "prov:Entity": "xas:Beamline" },	Beamline xdi
3	Beamline.collimation	"schema:additionalProperty": [{ "@type": "schema:PropertyValue", "schema:propertyID": { "@id": "xas:collimation" }, "schema:value": "(value)" }],	Beamline dat
4	Beamline.detector	"schema:additionalProperty": [{ "@type": "schema:PropertyValue", "schema:propertyID": { "@id": "xas:detector" }, "schema:value": "(value)" }],	Beamline dat
5	Beamline.energy_range	"schema:additionalProperty": [{ "@type": "schema:PropertyValue", "schema:propertyID": { "@id": "xas:energy_range" }, "schema:value": "(value)" }],	Beamline dat
6	Beamline.energy_resolution	"schema:additionalProperty": [{ "@type": "schema:PropertyValue", "schema:propertyID": { "@id": "xas:energy_resolution" }, "schema:value": "(value)" }],	Beamline dat
7	Beamline.flux	"schema:additionalProperty": [{ "@type": "schema:PropertyValue", "schema:propertyID": { "@id": "xas:flux" }, "schema:value": "(value)" }],	Beamline dat
8	Beamline.focusing	"schema:additionalProperty": [{ "@type": "schema:PropertyValue", "schema:propertyID": { "@id": "xas:focusing" }, "schema:value": "(value)" }],	Beamline xdi
9	Beamline.harmonic_rejection	"schema:additionalProperty": [{ "@type": "schema:PropertyValue", "schema:propertyID": { "@id": "xas:harmonic_rejection" }, "schema:value": "(value)" }],	Beamline xdi
	Beamline.name	"schema:additionalProperty": [Beamline xdi

Experimental CDIF as a service

GET	/ Read Root	⌄
GET	/cdi-intermidiate Read Cdi	⌄
GET	/data Read Data	⌄
GET	/data/example Read Data Example	⌄
GET	/data/dataset Read Data Dataset	⌄
GET	/datapoints Read Datapoints	⌄
GET	/data/serialize Read Data Serialize	⌄
GET	/data/type Read Data Type	⌄
GET	/data/properties Read Data Properties	⌄
GET	/data/triple_by_triple Read Data Triple By Triple	⌄
GET	/data/lookup Read Data Lookup	⌄
GET	/data/lookup/predicate Read Data Lookup Predicate	⌄
GET	/data/lookup/object Read Data Lookup Object	⌄
GET	/data/lookup/subject Read Data Lookup Subject	⌄

Demo here: <https://cdif-4-xas.dev.codata.org/docs>

Intermediate step with SKOS transformation

Curl

```
curl -X 'GET' \
  'https://cdif-4-xas.dev.codata.org/cdi-intermidiate?url=https%3A%2F%2Fraw.githubusercontent.com%2FXraySpectroscopy%2FXAS-Data-Interchange%2Frefs%2Fheads%2Fmaster%2Fdata%2Fnonxafs_2d.xdi&format=turtle' \
  -H 'accept: application/json'
```

Request URL

```
https://cdif-4-xas.dev.codata.org/cdi-intermidiate?url=https%3A%2F%2Fraw.githubusercontent.com%2FXraySpectroscopy%2FXAS-Data-Interchange%2Frefs%2Fheads%2Fmaster%2Fdata%2Fnonxafs_2d.xdi&format=turtle
```

Server response

Code	Details
200	Response body

```
@prefix label: <https://ddi-cdi.org/label/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix xas: <http://ddialliance.org/Specification/XAS/> .

xas:Beamline skos:definition "Tags related to the structure of the beamline and its photon delivery system" ;
  skos:prefLabel "Beamline" .

xas:Beamline.collimation xas:format "free-format string" ;
  xas:units "none" ;
  skos:definition "A concise statement of how beam collimation is provided" ;
  skos:prefLabel "Beamline.collimation" .

xas:Beamline.focusing xas:format "free-format string" ;
  xas:units "none" ;
  skos:definition "A concise statement about how beam focusing is provided" ;
  skos:prefLabel "Beamline.focusing" .

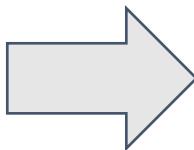
xas:Beamline.harmonic_rejection xas:format "free-format string" ;
  xas:units "none" ;
  skos:definition "A concise statement about how harmonic rejection is accomplished" ;
  skos:prefLabel "Beamline.harmonic_rejection" .

xas:Beamline.name xas:format "free-format string" ;
  xas:units "none" ;
```

  Download

CDIF transformations - json-ld output

```
{ "@context": [ "https://docs.ddialliance.org/DDI-CDI/1.0/model/encoding/json-ld/ddi-cdi.jsonld", { "skos": "http://www.w3.org/2004/02/skos/core#", "xdi": "http://www.w3.org/2004/02/skos/core#", "cdi": "https://docs.ddialliance.org/DDI-CDI/1.0/model/encoding/json-ld/ddi-cdi.jsonld" } ], "DDICDIModels": [ { "@id": "#measureComponent-energy", "cdi:DataStructureComponent_isDefinedBy_RepresentedVariable": [ { "@id": "#instanceVariable-energy" } ] }, { "@id": "#wideDataStructure", "@type": [ "cdi:WideDataStructure" ], "cdi:DataStructure_has_DataStructureComponent": [ { "@id": "#identifierComponent-Outer.value" }, { "@id": "#identifierComponent-line_no" }, { "@id": "#measureComponent-energy" } ] }, { "@id": "#identifierComponent-line_no", "cdi:DataStructureComponent_isDefinedBy_RepresentedVariable": [ { "@id": "#instanceVariable-line_no" } ] }, { "@id": "#logicalRecord-wds", "cdi:LogicalRecord_has_InstanceVariable": [ { "@id": "#instanceVariable-line_no" }, { "@id": "#instanceVariable-Outer.value" } ] } ] }
```



Nectar Publisher Pre-alpha 0.0.2

import tabular data import metadata import data via service NEW: import external dataset about

Dataset Information Variables Export documentation

MarkDown csv ddi-cdi (json-ld) <> ddi-l (xml) <> ddi-c (xml) <> ddi-40-l (json)

save copy

```
<?xml version="1.0" encoding="UTF-8"?>
<ddi:FragmentInstance xmlns:ddi="ddi:instance:3_3" xmlns:s="ddi:studyunit:3_3" xmlns:r="ddi:reusable:3_3" xmlns:pi="ddi:physicalinstance:3_3" xmlns:=">
  <ddi:TopLevelReference>
    <r:Agency xmlns:r="ddi:reusable:3_3">int.example</r:Agency>
    <r:ID xmlns:r="ddi:reusable:3_3">dfbeb171-c83a-4c09-875f-bb3f583b6c80</r:ID>
    <r:Version xmlns:r="ddi:reusable:3_3">1.0.0</r:Version>
    <r>TypeOfObject xmlns:r="ddi:reusable:3_3">StudyUnit</r>TypeOfObject>
  </ddi:TopLevelReference>
  <ddi:Fragment>
    <s:StudyUnit xmlns:s="ddi:physicalinstance:3_3">
      <r:URN xmlns:r="ddi:reusable:3_3">urn:ddi:int.example:dfbeb171-c83a-4c09-875f-bb3f583b6c80:1.0.0</r:URN>
      <r:Agency xmlns:r="ddi:reusable:3_3">int.example</r:Agency>
      <r:ID xmlns:r="ddi:reusable:3_3">dfbeb171-c83a-4c09-875f-bb3f583b6c80</r:ID>
      <r:Version xmlns:r="ddi:reusable:3_3">1.0.0</r:Version>
      <r:Abstract xmlns:r="ddi:reusable:3_3">Please describe the content and Method of this study.</r:Abstract>
      <r:PhysicalInstanceReference xmlns:r="ddi:reusable:3_3">
        <r:Agency>int.example</r:Agency>
        <r:ID>2a60dd57-082b-4870-a3d1-3351919ac0f8</r:ID>
        <r:Version>1.0.0</r:Version>
        <r>TypeOfObject>PhysicalInstance</r>TypeOfObject>
      </r:PhysicalInstanceReference>
      <r:LogicalProductReference xmlns:r="ddi:reusable:3_3">
        <r:Agency>int.example</r:Agency>
        <r:ID>ba7708c9-c942-406b-a3c2-e133c8d486d1</r:ID>
        <r:Version>1.0.0</r:Version>
        <r>TypeOfObject>LogicalProduct</r>TypeOfObject>
      </r:LogicalProductReference>
      </s:StudyUnit>
    </ddi:Fragment>
    <ddi:Fragment>
      <pi:PhysicalInstance xmlns:pi="ddi:physicalinstance:3_3">
        <r:URN xmlns:r="ddi:reusable:3_3">urn:ddi:int.example:2a60dd57-082b-4870-a3d1-3351919ac0f8:1.0.0</r:URN>
        <r:Agency xmlns:r="ddi:reusable:3_3">int.example</r:Agency>
        <r:ID xmlns:r="ddi:reusable:3_3">2a60dd57-082b-4870-a3d1-3351919ac0f8</r:ID>
      </pi:PhysicalInstance>
    </ddi:Fragment>
  </ddi:FragmentInstance>
```

Nectar Publisher

“Interlink”: Semantic Croissant with WikiData

...

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 (cell line)</i>	No direct match; the best available entity is unrelated to intensity monitoring.	Not relevant – no appropriate Wikidata entry for “monitor intensity”.
itrans	Q109607869	<i>iTransmission, Inc.</i>	Company named iTransmission.	Not relevant – no suitable Wikidata entry for transmission intensity.
ifluor	—	—	No matching Wikidata entity found.	Not relevant – 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).	Not relevant – no suitable Wikidata entry for reference intensity.

“Interlink”: AI-powered CDIF variables description

Variable	Typical Physical Quantity	Common Units (SI / Conventional)	Notes
energy	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.
angle	Scattering angle (2θ)	degrees, radians, steps	Degrees are standard in crystallography tables; “steps” may refer to detector step size when the angle is sampled discretely.
i0	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.
itrans	Transmitted beam intensity	counts/s, photons/s, arbitrary units (a.u.)	Often expressed relative to i0 (i.e., transmission = itrans / i0).
ifluor	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.
irefer	Reference beam intensity (e.g., from a standard sample)	counts/s, photons/s, arbitrary units (a.u.)	Used for normalisation; typically expressed relative to i0 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

SKOS Relationships (Wikidata Links)

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

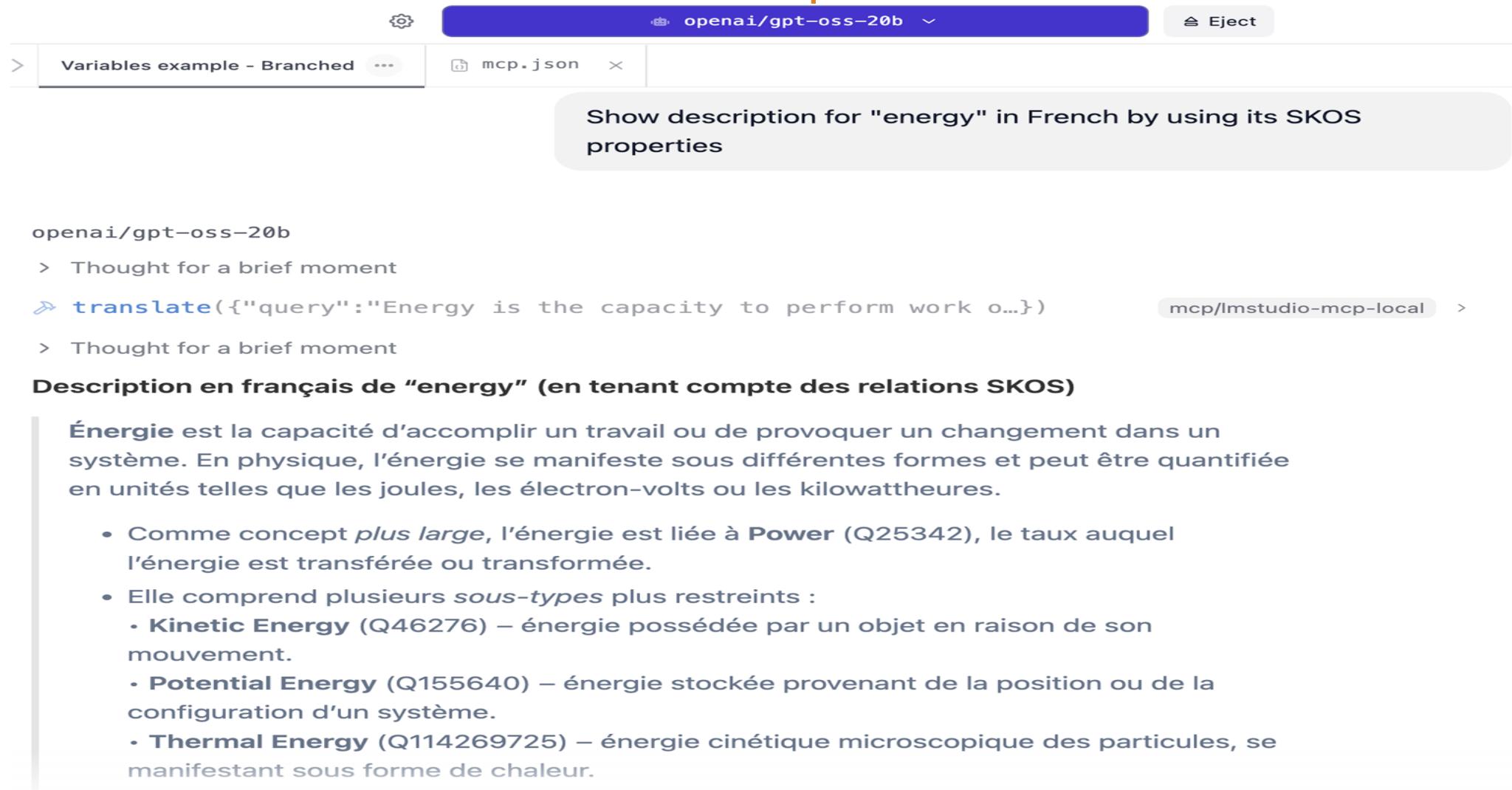
“Interlink”: Semantic Croissant with CDIF as “navigation” system for AI



The screenshot shows a web-based AI interface with the following elements:

- Header: "openai/gpt-oss-20b" in a blue bar with a gear icon and "Eject" button.
- Left sidebar: "Variables example" and "mcp.json".
- Input field: "Make new description for 'energy' by using its SKOS properties".
- Output area:
 - Text: "openai/gpt-oss-20b" and "Thought for a brief moment".
 - Section header: "Revised Description of 'energy' (incorporating SKOS relationships)".
 - Text: "Energy is the capacity to perform work or produce change in a system. In physical science it manifests in various forms and can be quantified in units such as joules, electron-volts, or kilowatt-hours.".
 - List:
 - As a *broader* concept, energy is linked to **Power** (Q25342), the rate at which energy is transferred or transformed.
 - It includes several *narrower* subtypes:
 - Kinetic Energy** (Q46276) – energy possessed by an object due to its motion.
 - Potential Energy** (Q155640) – stored energy arising from position or configuration.
 - Thermal Energy** (Q114269725) – microscopic kinetic energy of particles, manifesting as heat.
 - Related concepts provide additional context: **Work** (Q192581), the transfer of energy through force over distance; **Heat** (Q60377217), a specific form of energy flow driven by temperature gradients; and **Radiation** (Q18335), the emission of electromagnetic waves carrying energy across space or media.
 - Text: "This integrated view highlights energy as both an overarching property and a nexus connecting multiple physical phenomena."

“Interlink”: Multilingual support powered by CDIF



Variables example - Branched

mcp.json

openai/gpt-oss-20b

Eject

Show description for "energy" in French by using its SKOS properties

openai/gpt-oss-20b

> Thought for a brief moment

translate({"query":"Energy is the capacity to perform work o..."})

mcp/lmstudio-mcp-local >

> Thought for a brief moment

Description en français de "energy" (en tenant compte des relations SKOS)

Énergie est la capacité d'accomplir un travail ou de provoquer un changement dans un système. En physique, l'énergie se manifeste sous différentes formes et peut être quantifiée en unités telles que les joules, les électron-volts ou les kilowattheures.

- Comme concept *plus large*, l'énergie est liée à **Power** (Q25342), le taux auquel l'énergie est transférée ou transformée.
- Elle comprend plusieurs *sous-types* plus restreints :
 - **Kinetic Energy** (Q46276) – énergie possédée par un objet en raison de son mouvement.
 - **Potential Energy** (Q155640) – énergie stockée provenant de la position ou de la configuration d'un système.
 - **Thermal Energy** (Q114269725) – énergie cinétique microscopique des particules, se manifestant sous forme de chaleur.

“Interlink”: AI “understands” CDIF variables

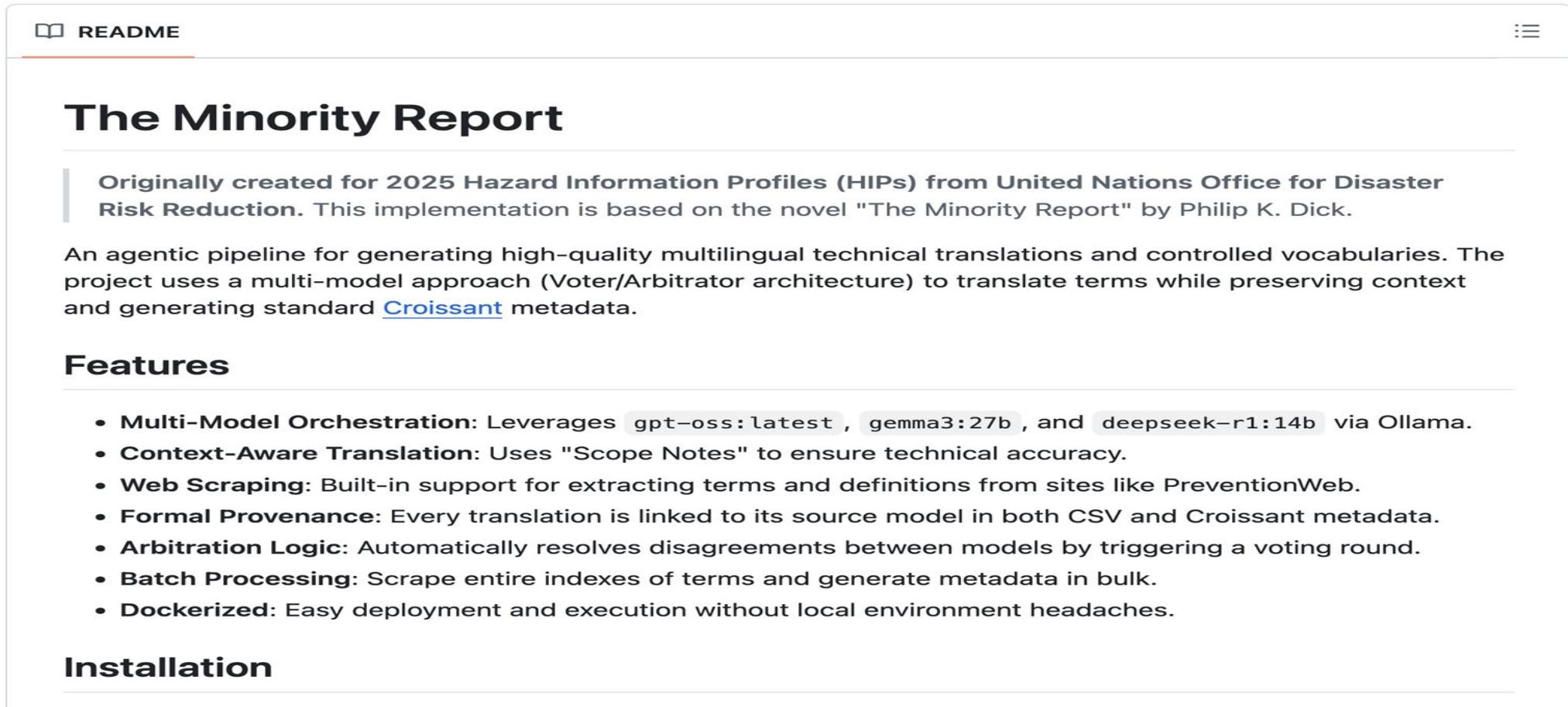
cdi-xas / docs / CDIF_overview_missing.md ↑ Top

Preview Code Blame 118 lines (102 loc) · 12.8 KB Raw     

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

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

The Minority Report



The screenshot shows a GitHub README page for a project named "The Minority Report". The page has a header with a "README" button and a three-dot menu. The main content starts with a section titled "The Minority Report" which includes a description of the project's origin and purpose. Below this is a "Features" section with a bulleted list of seven items. The "Features" section is followed by an "Installation" section.

README ⋮

The Minority Report

Originally created for 2025 Hazard Information Profiles (HIPs) from United Nations Office for Disaster Risk Reduction. This implementation is based on the novel "The Minority Report" by Philip K. Dick.

An agentic pipeline for generating high-quality multilingual technical translations and controlled vocabularies. The project uses a multi-model approach (Voter/Arbitrator architecture) to translate terms while preserving context and generating standard [Croissant](#) metadata.

Features

- **Multi-Model Orchestration:** Leverages `gpt-oss:latest`, `gemma3:27b`, and `deepseek-r1:14b` via Ollama.
- **Context-Aware Translation:** Uses "Scope Notes" to ensure technical accuracy.
- **Web Scraping:** Built-in support for extracting terms and definitions from sites like PreventionWeb.
- **Formal Provenance:** Every translation is linked to its source model in both CSV and Croissant metadata.
- **Arbitration Logic:** Automatically resolves disagreements between models by triggering a voting round.
- **Batch Processing:** Scrape entire indexes of terms and generate metadata in bulk.
- **Dockerized:** Easy deployment and execution without local environment headaches.

Installation

Reference: [Github](#)

UN Hazard Information Profiles (HIPS): 2025

On this page

Meteorological and Hydrological

Extraterrestrial

Geological

Environmental

Chemical

Biological

Technological

Societal

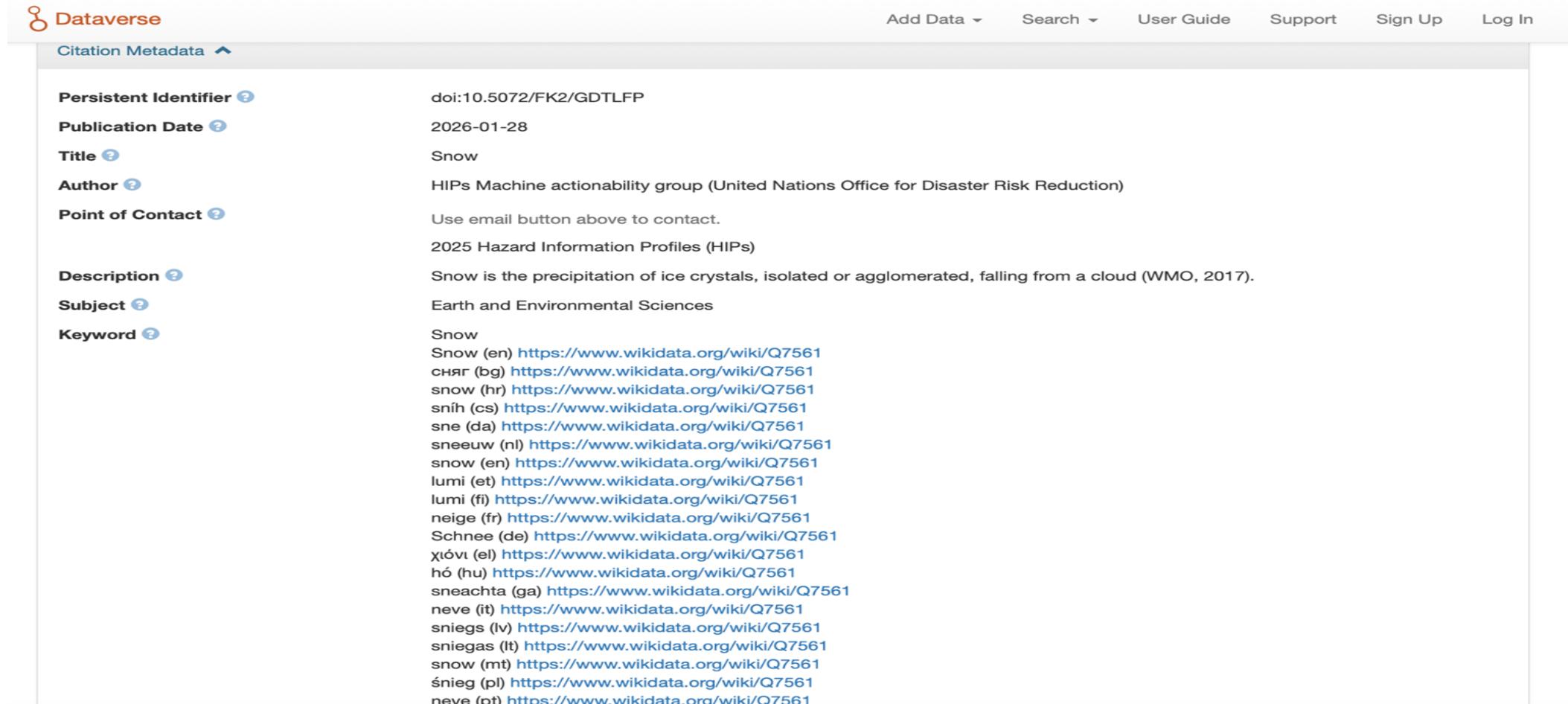
Meteorological and Hydrological

8 hazards clusters and 53 specific hazards

Specific hazard	Identifier	Cluster
Avalanche	MH0801	Terrestrial
Blizzard	MH0403	Precipitation-related
Coastal Flooding	MH0601	Water-related
Cold Wave	MH0502	Temperature-related
Depression or Cyclone (Low Pressure Area)	MH0306	Wind- & Pressure-related
Derecho	MH0302	Wind- & Pressure-related
Downburst	MH0101	Convective-related
Drought	MH0401	Precipitation-related
Dust storm or Sandstorm	MH0201	Particle-related
Dzud	MH0503	Temperature-related
Estuarine (Coastal) Flooding	MH0602	Water-related
Extra-tropical Cyclone	MH0307	Wind- & Pressure-related

Reference: United Nations [HIPS](#)

Automated translations in all European languages with LLM models



The screenshot shows a Dataverse citation metadata page. The top navigation bar includes links for Add Data, Search, User Guide, Support, Sign Up, and Log In. The main content area is titled 'Citation Metadata' and displays the following data:

Field	Value
Persistent Identifier	doi:10.5072/FK2/GDTLFP
Publication Date	2026-01-28
Title	Snow
Author	HIPs Machine actionability group (United Nations Office for Disaster Risk Reduction)
Point of Contact	Use email button above to contact. 2025 Hazard Information Profiles (HIPs)
Description	Snow is the precipitation of ice crystals, isolated or aggregated, falling from a cloud (WMO, 2017).
Subject	Earth and Environmental Sciences
Keyword	Snow Snow (en) https://www.wikidata.org/wiki/Q7561 сняг (bg) https://www.wikidata.org/wiki/Q7561 snow (hr) https://www.wikidata.org/wiki/Q7561 sníh (cs) https://www.wikidata.org/wiki/Q7561 sne (da) https://www.wikidata.org/wiki/Q7561 sneeuw (nl) https://www.wikidata.org/wiki/Q7561 snow (en) https://www.wikidata.org/wiki/Q7561 lumi (et) https://www.wikidata.org/wiki/Q7561 lumi (fi) https://www.wikidata.org/wiki/Q7561 neige (fr) https://www.wikidata.org/wiki/Q7561 Schnee (de) https://www.wikidata.org/wiki/Q7561 χιόνι (el) https://www.wikidata.org/wiki/Q7561 hó (hu) https://www.wikidata.org/wiki/Q7561 sneachta (ga) https://www.wikidata.org/wiki/Q7561 neve (it) https://www.wikidata.org/wiki/Q7561 sniegs (lv) https://www.wikidata.org/wiki/Q7561 sniegas (lt) https://www.wikidata.org/wiki/Q7561 snow (mt) https://www.wikidata.org/wiki/Q7561 śnieg (pl) https://www.wikidata.org/wiki/Q7561 neve (pt) https://www.wikidata.org/wiki/Q7561

Reference: "Snow" translations in Dataverse

MCP server for HIPs running on custom model

Analyze: Avalanche risk is high in the mountains.

I'll analyze this English text for disaster risk terminology and hazards.

R find_hazards

Based on the hazard analysis, I found the following disaster risk terminology in the English text:

Identified Hazard:

- **Avalanche risk** - HIPS Code: **BI0601**

- This is classified as a biological/geological hazard (likely geological given the context)
- Position in text: characters 0-14
- Note: The label shows "TR_BG" which may indicate Bulgarian language tagging, though this appears to be a language classification artifact

Context: The text describes high avalanche risk in mountainous areas.

The NER model successfully identified "Avalanche risk" as hazard terminology. Avalanches are typically classified under geological or mass movement hazards in disaster risk reduction frameworks, as they involve the rapid movement of snow and ice down slopes, often triggered by weather conditions, terrain features, or human activity.



find hazards: Ein Schneesturm ist ein heftiges Wetterphänomen, bei dem große Mengen Schnee in kurzer Zeit vom Himmel fallen und die Sicht stark reduziert

I'll analyze this German text for disaster risk terminology and hazards.

R find_hazards

Based on the hazard analysis, I found the following disaster risk terminology in the German text:

Identified Hazard:

- **Schneesturm** (Snowstorm) - HIPS Code: **MH0406**

- This is classified as a meteorological/hydrological hazard
- Position in text: characters 4-15

Context: The text provides a definition of a snowstorm as "a severe weather phenomenon in which large amounts of snow fall from the sky in a short time and visibility is greatly reduced."

The NER model successfully identified "Schneesturm" as the standardized German term for snowstorm, matching it to the same HIPS classification code (MH0406) as the Finnish "Lumimyrsky" from the previous query, demonstrating the multilingual hazard identification capability.



CDIF demo: “random” spreadsheet file without metadata

aut_core

country	pageid	wikidataid	wikititle	name	sex	ethnicity	religion	birth	death	birthplace	deathplace
AUT	3024179	Q2301596	Sonja_Ablinger	Ablinger Sonja	female	white	NA	1966-05-12	NA	48.15,14.01667	NA
AUT	3958848	Q1511220	Gerhard_Abraham	Abraham Gerhard	male	NA	NA	1947-10-24	NA	46.84194,14.84083	NA
AUT	3041286	Q2287337	Simon_Abram	Abram Simon	male	white	NA	1871-04-03	1940-02-29	47.2569,11.3861	47.8,13.03333
AUT	3879475	Q560234	Anna_Elisabeth_Achazt	Achazt Anna Elisabeth	female	NA	NA	1950-05-07	NA	NA	NA
AUT	3016689	Q1331445	Elke_Achleitner	Achleitner Elke	female	NA	NA	1964-04-02	NA	48.21,13.48944	NA
AUT	4104949	Q1909660	Matthias_Achs	Achs Matthias	male	NA	NA	1939-12-06	2011-05-25	47.89778,16.90889	47.89778,16.90889
AUT	377116	Q93503	Friedrich_Adler_(Politiker)	Adler Friedrich	male	white	NA	1879-07-09	1960-01-02	48.20833,16.37306	47.37861,8.54
AUT	3521961	Q1908424	Mathias_Adlersflügel	Adlersflügel Mathias	male	NA	NA	1868-01-17	1933-09-01	48.20833,16.37306	48.20833,16.37306
AUT	1088182	Q94397	Josef_Afritsch	Afritsch Josef	male	white	NA	1901-03-13	1964-08-25	47.06667,15.43333	48.20833,16.37306
AUT	3961890	Q1730231	Karl_Aichhorn	Aichhorn Karl	male	NA	NA	1903-04-20	1967-02-20	48.20833,16.37306	48.20833,16.37306
AUT	1098388	Q1285924	Edmund_Aigner	Aigner Edmund	male	NA	NA	1900-03-29	1968-05-03	48.14444,13.73639	48.3,14.28333
AUT	4945354	Q1609919	Josef_Aigner_(Politiker,_1884)	Aigner Josef	male	NA	NA	1884-03-19	1947-08-05	48.3,14.28333	48.3,14.28333
AUT	3961920	Q2650136	Alois_Alberer	Alberer Alois	male	NA	NA	1922-01-07	2011-10-27	46.71361,14.20639	46.76667,14.36028
AUT	4896439	Q566129	Anneliese_Albrecht	Albrecht Anneliese	female	NA	NA	1921-01-29	2018-10-31	48.20833,16.37306	NA
AUT	880210	Q1596544	Heinrich_Allina	Allina Heinrich	male	NA	NA	1878-11-24	1953-12-10	48.86685,15.73466	48.20833,16.37306
AUT	6346640	Q964070	Niko_Alm	Alm Nikolaus	male	white	atheism	1975-08-30	NA	48.20833,16.37306	NA
AUT	3962418	Q1446019	Franz_Altenbacher	Altenbacher Franz	male	NA	NA	1862-02-07	1936-11-06	51.27606,8.58328	47.06667,15.43333
AUT	1065328	Q1362800	Erwin_Altenburger	Altenburger Erwin	male	NA	NA	1903-11-03	1984-02-07	47.4,14.83333	48.20833,16.37306
AUT	3962846	Q1594491	Sepp_Aman	Aman Sepp	male	NA	NA	1870-12-02	1957-04-17	47.22222,9.71667	47.06667,15.43333
AUT	7769254	Q15819226	Hannes_Amesbauer	Amesbauer Hannes	male	white	NA	1981-04-18	NA	47.41056,15.26861	NA
AUT	3047911	Q1926796	Michael_Amlacher	Amlacher Michael	male	NA	NA	1882-03-28	1964-01-12	47.4,15.21667	46.955,15.88833
AUT	466005	Q1771151	Werner_Amon	Amon Werner	male	NA	NA	1969-05-28	NA	47.06667,15.43333	NA
AUT	3962875	Q1251723	Heinrich_Amtmann	Amtmann Heinrich	male	NA	NA	1929-10-24	2014-08-12	46.5,11.35	46.61667,13.83333
AUT	312684	Q877350	Hannes_Androsch	Androsch Hannes	male	white	NA	1938-04-18	NA	48.20833,16.37306	NA
AUT	7404842	Q1911243	Maurice_Androsch	Androsch Maurice	male	white	NA	1970-04-25	NA	48.20833,16.37306	NA
AUT	8311185	Q17352678	Erwin_Angerer	Angerer Erwin	male	NA	NA	1964-12-30	NA	46.8575,13.35194	NA

AI pipeline was asked to “predict” possible concepts in columns (“cells”), and “interlink”, and contemplate

Demo: dataset registered in Dataverse as “resource”

 **Dataverse** Add Data ▾ Search ▾ User Guide Support Sign Up Log In

License/Data Use Agreement  CC0 1.0

Files Metadata Terms Versions

Export Metadata ▾

Citation Metadata ▾

Persistent Identifier ⓘ	doi:10.5072/FK2/RIB64Q
Publication Date ⓘ	2025-11-13
Title ⓘ	aut_core test
Author ⓘ	Admin, Dataverse (Dataverse.org)
Point of Contact ⓘ	Use email button above to contact. Admin, Dataverse (Dataverse.org)
Description ⓘ	aut_core test data
Subject ⓘ	Agricultural Sciences
Depositor ⓘ	Admin, Dataverse
Deposit Date ⓘ	2025-11-13

DataCite
Schema.org JSON-LD
DDI Codebook v2
OAI_ORE
RO-Crate
OpenAIRE
CDIF variables JSON_LD
Dublin Core
Croissant
DDI HTML Codebook
OAI-ORE ResourceMap
JSON
Semantic Croissant

AI pipeline: variable “birth” described by AI

```
▼ 17:
  name:                      "birth"
  ▼ ollama_remote:
    name:                      "birth"
  ▼ ollama:
    variable:                  "birth"
    definition:                "Birth refers to the event at which a baby or organism is brought into existence, marking the transition from a gestational state to independent life."
  ▼ units_of_measurement:
    ▼ 0:
      name:                      "date"
      unit:                      "calendar date (YYYY-MM-DD)"
      description:               "Date of birth"
    ▼ 1:
      name:                      "time"
      unit:                      "hh:mm:ss"
      description:               "Time of birth"
    ▼ 2:
      name:                      "birth_weight_kg"
      unit:                      "kilograms (kg)"
      description:               "Weight of the newborn at birth"
    ▼ 3:
      name:                      "birth_weight_g"
      unit:                      "grams (g)"
      description:               "Weight of the newborn at birth (alternate unit)"
    ▼ 4:
      name:                      "birth_height_cm"
      unit:                      "centimeters (cm)"
      description:               "Length/height of the newborn at birth"
```

Analyze this dataset: dataVERSE.dev.codata.org/api/datasets/13

Providing dataset analysis 13s

Overview

This resource is a **tabular biographical dataset of individuals linked to Wikipedia/Wikidata**, with rich variable-level metadata. It appears to contain ~1,923 records (from [page 1](#)) and 12 variables describing each person's identity, demographics, and basic life events, plus geographic information about birth and death locations [1](#).

Dataset structure

- **Format**

- One row per **individual**.
- Columns (variables) are listed under `variables` and `context`; their semantics are detailed under `results`.
- `variableFormatType` is mostly `CHARACTER`, with `pageid` as `NUMERIC`.
- `UNF` fields are **Universal Numeric Fingerprints** (hashes to verify data integrity), not analytically meaningful.

- **Core variables (columns)**

All 12 variables in `context`:

1. `country` (CHARACTER, categorical):

Political entity / sovereign state associated with the record, likely the country of the person (e.g., residence or nationality). Described as a nominal string with possible ISO2/ISO3/numeric codes, continent, subregion, region, and name variants [1](#).

2. `pageid` (NUMERIC, discrete):

Unique integer identifier of the person's **Wikipedia page**; used as primary key in MediaWiki. Summary stats show:

- `mean` ≈ 5.1M, `min` = 31,149, `max` ≈ 11,114,596, `stdev` ≈ 2.86M, `vald` = 1923.

3. `wikidataid` (CHARACTER):

The person's **Wikidata entity ID** (e.g., Q42), allowing linkage to Wikidata knowledge graph.

4. `wikititle` (CHARACTER):

Wikipedia article title for the person.

5. `name` (CHARACTER):

Person's human-readable name (may coincide with `wikititle` or be a cleaned version).

Analyze the provided dataset

+

Analyze this dataset: dataverse.dev.codata.org/api/datasets

...

Person's human-readable name (may coincide with `wikititle` or be a cleaned version).

6. `sex` (CHARACTER):

Sex / gender category (likely "male", "female", etc.); nominal categorical.

7. `ethnicity` (CHARACTER):

Ethnic identity; free-text or coded categories (not fully specified).

8. `religion` (CHARACTER):

Religious affiliation; also free-text or categorical codes.

9. `birth` (CHARACTER with date format `yyyy-MM-dd`):

Date of birth; a date string that can be parsed to compute age or lifespans.

10. `death` (CHARACTER):

Date of death; likely similarly formatted and used to derive age at death or survival analyses.

11. `birthplace` (CHARACTER, geographic):

Location where the person was born. Metadata specifies:

- Can be **human-readable string** ("New York City, New York, USA"), **standardized code** ("US-NY"), **latitude/longitude**, or **postal code**.
- Attributes include `city`, `state_province_region`, `country`, `postal_code`, `iso_country_code`, `latitude`, `longitude` [1](#).

12. `deathplace` (CHARACTER, geographic):

Location where the person died. Metadata includes:

- A single string field or coordinates, with sub-attributes for `type` (Hospital/Home/etc.), `address`, `city`, `state_province`, `country`, `postal_code`, `latitude`, `longitude`.
- Focused on **place-of-death classification** (hospital vs home, etc.) and potential geocoding [1](#).

Nature of the data

- **Subject:**

Biographical records for people who **have Wikipedia pages and Wikidata items**, with demographics and basic life-course information.

- **Granularity:**

One person per row, with links to external identifiers:

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

[Listen Croissant song!](#)

Demo - Dataverse MCP and CDIF
<https://mcp.dataverse.org/>