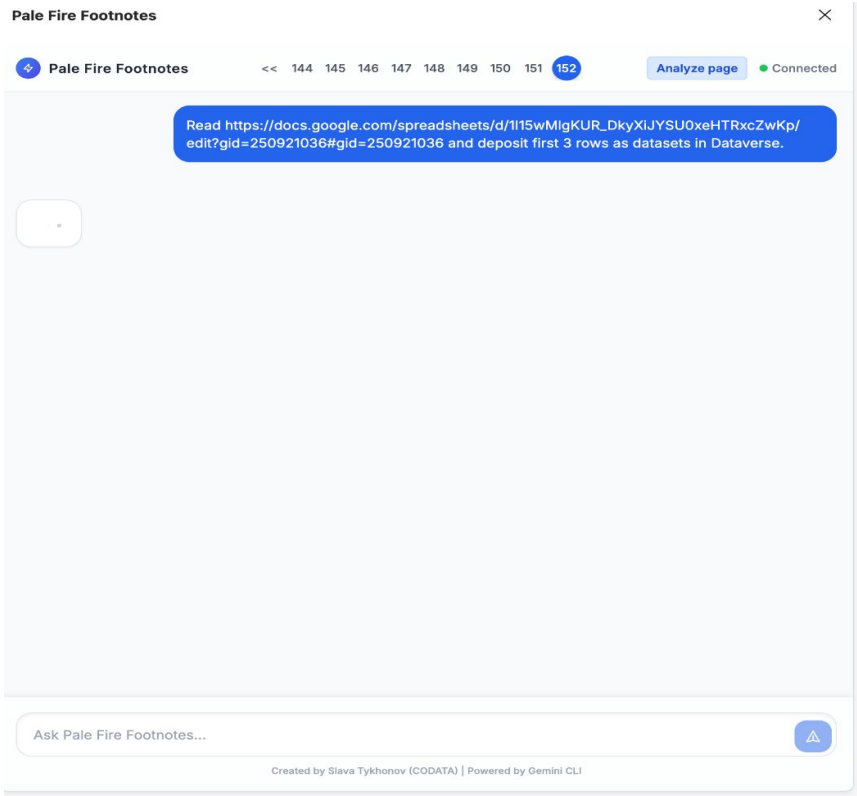
FAIR data management and AI in research data infrastructures:

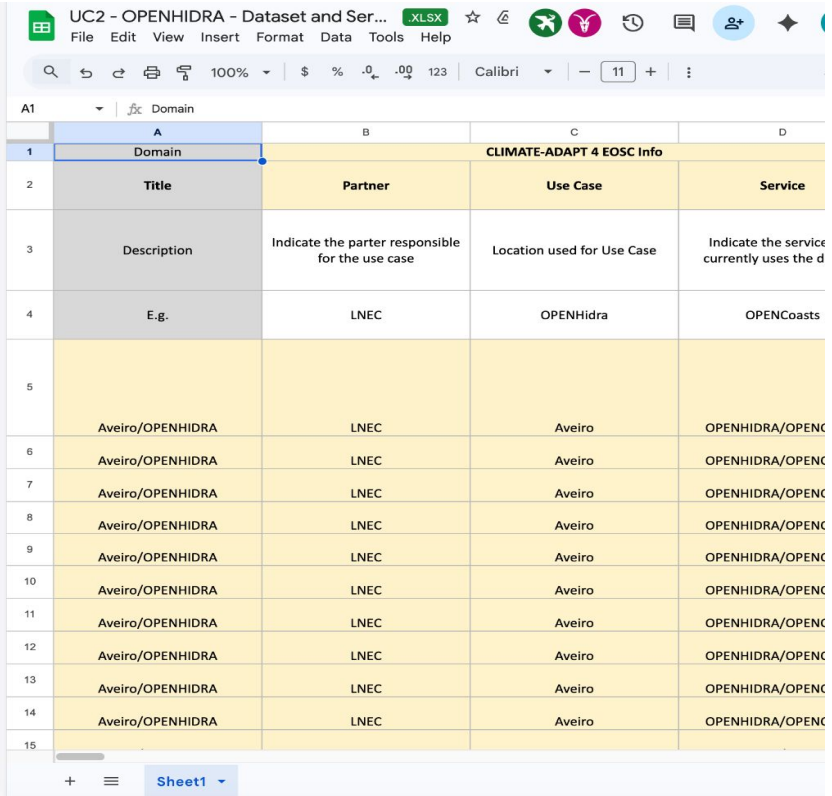
Semantic Croissant and CDIF

Slava Tykhonov
Head of AI and Interoperability at CODATA
Harvard Dataverse Ambassador

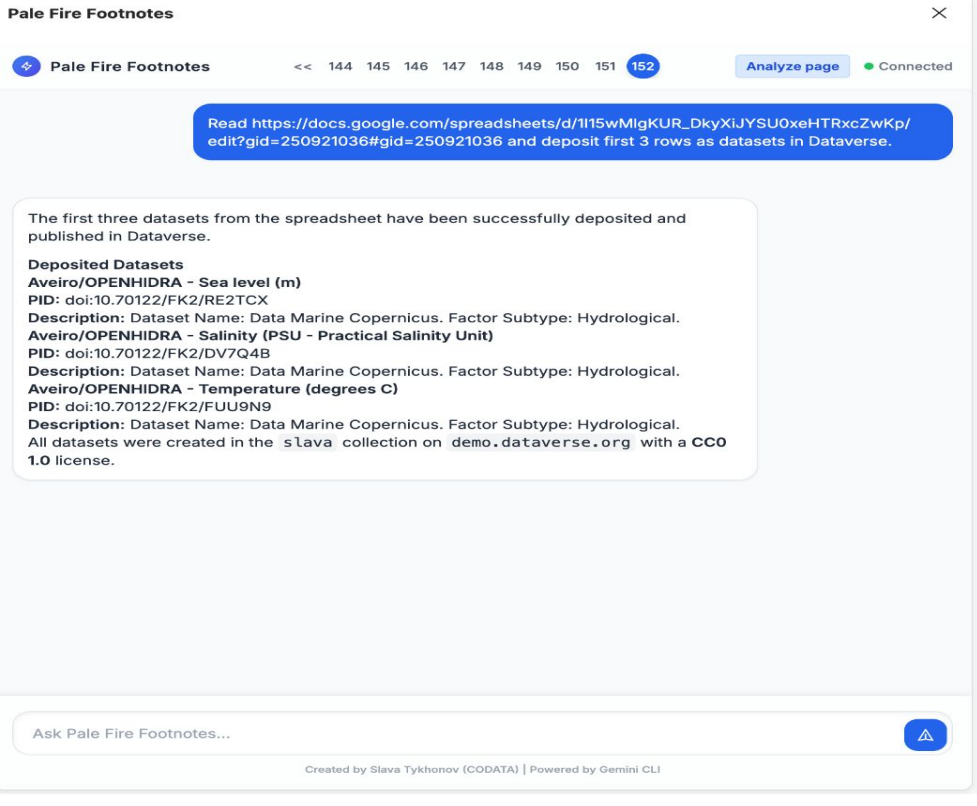
Keynote talk for the Nordic-Baltic Dataverse Hub (NAISH) project kick-off
Kaunas, Lithuania
21 April 2026

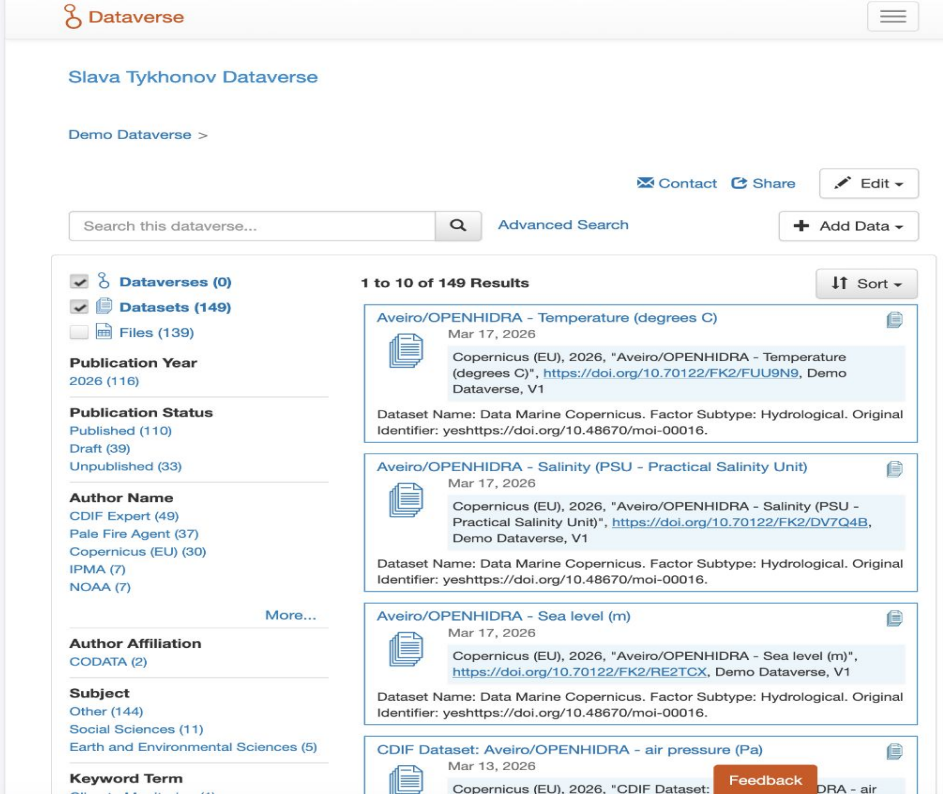
Intro: AI assistant to “collect” and deposit data



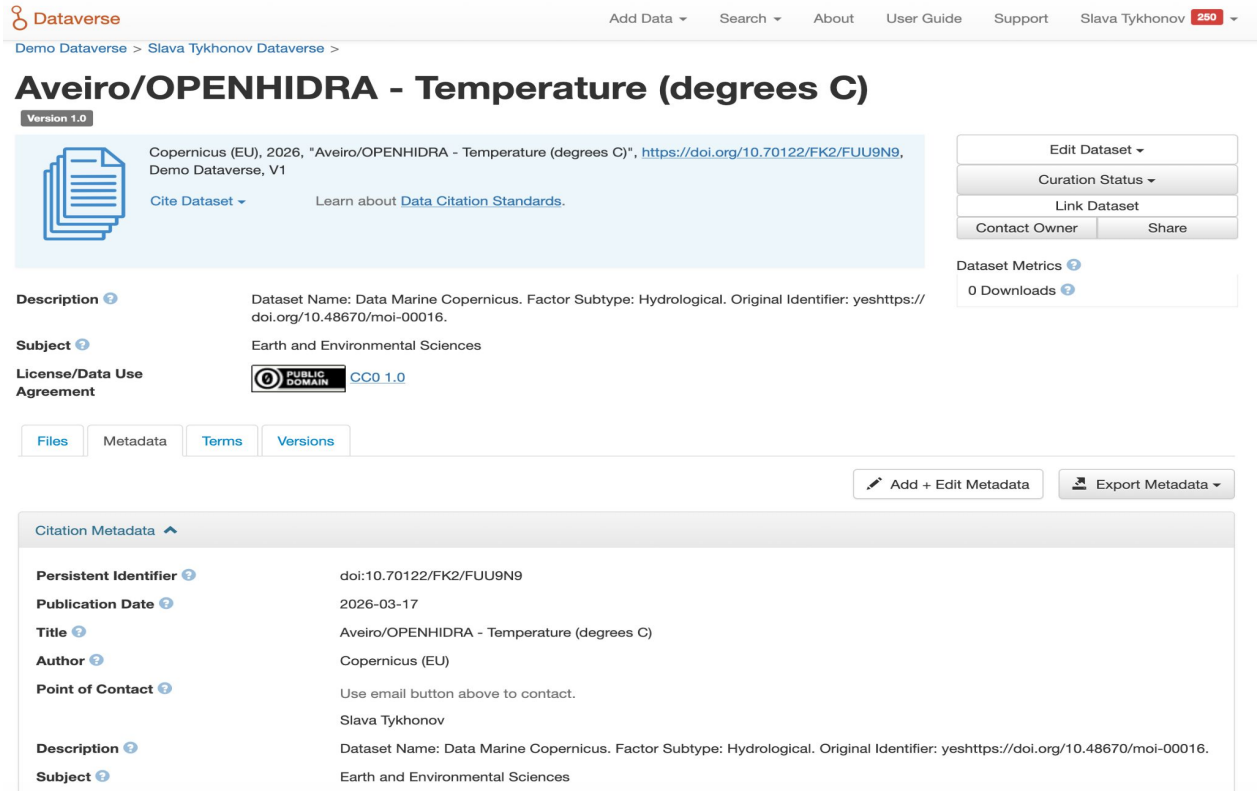


Intro: AI Assistant job as data depositor





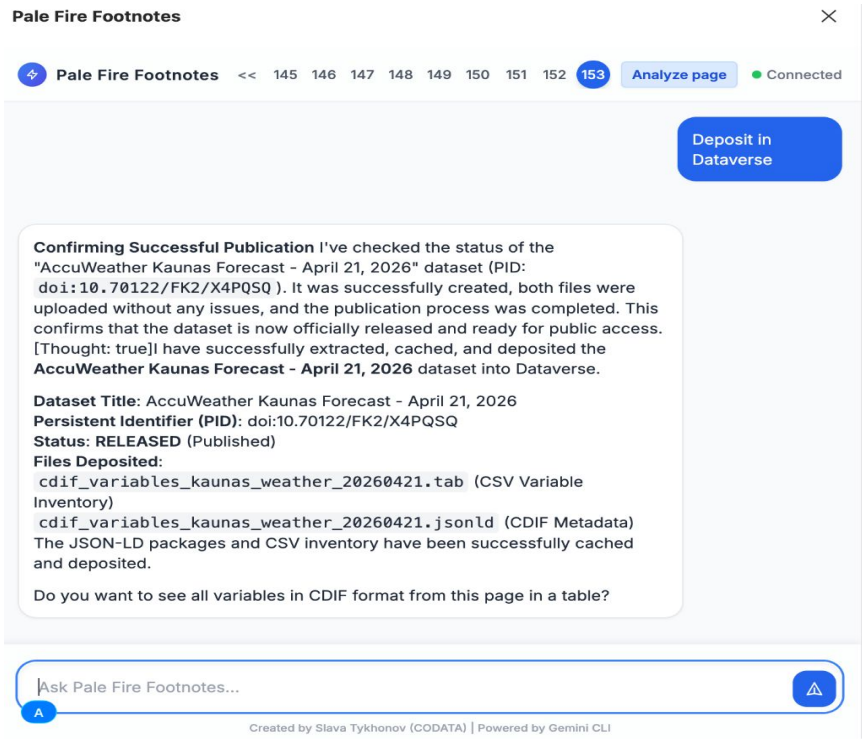
Intro: AI assistant is saving data in Dataverse

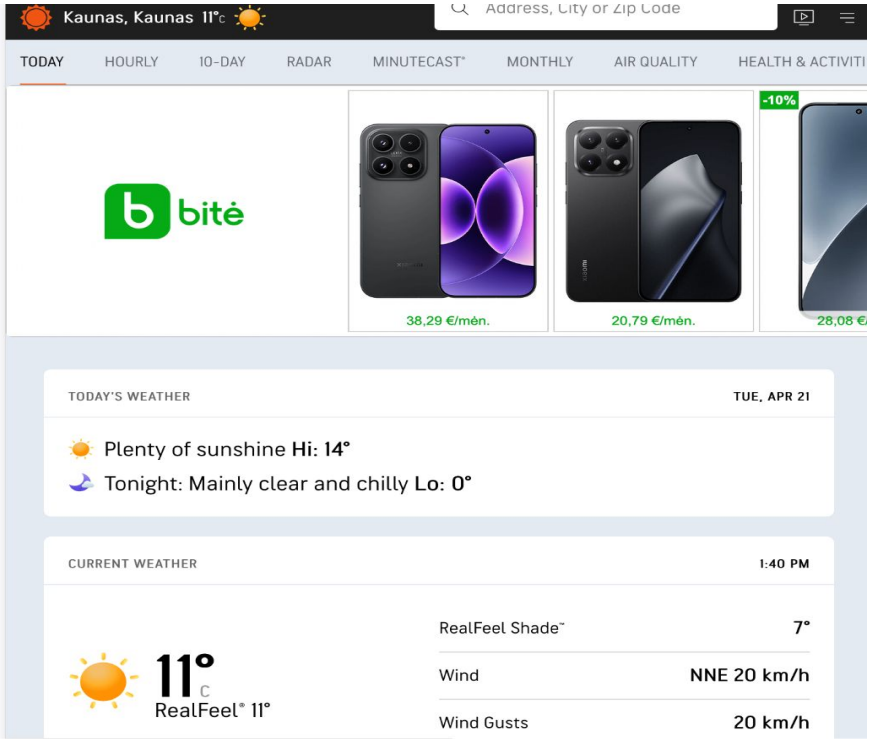


AI Agent job

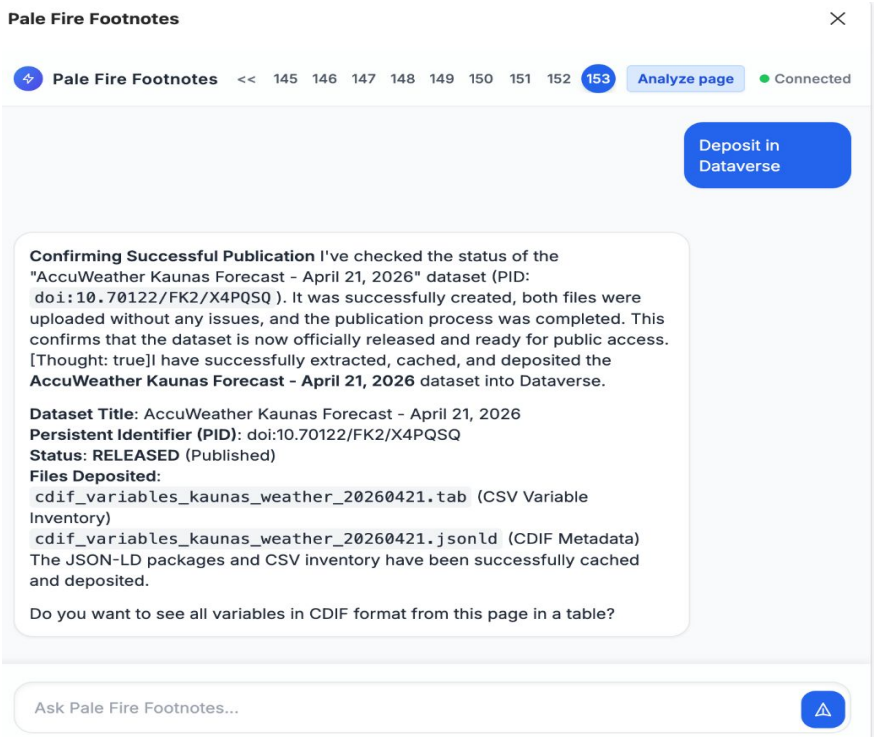
1. Read input from any source
2. Analyze Web pages and find data
3. Extract dataset descriptions and form metadata record
4. Create Croissant export for Machine Learning tasks
5. Deposit metadata as dataset in Dataverse data platform
6. Share Croissant back to give AI Agent pointers to metadata and data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

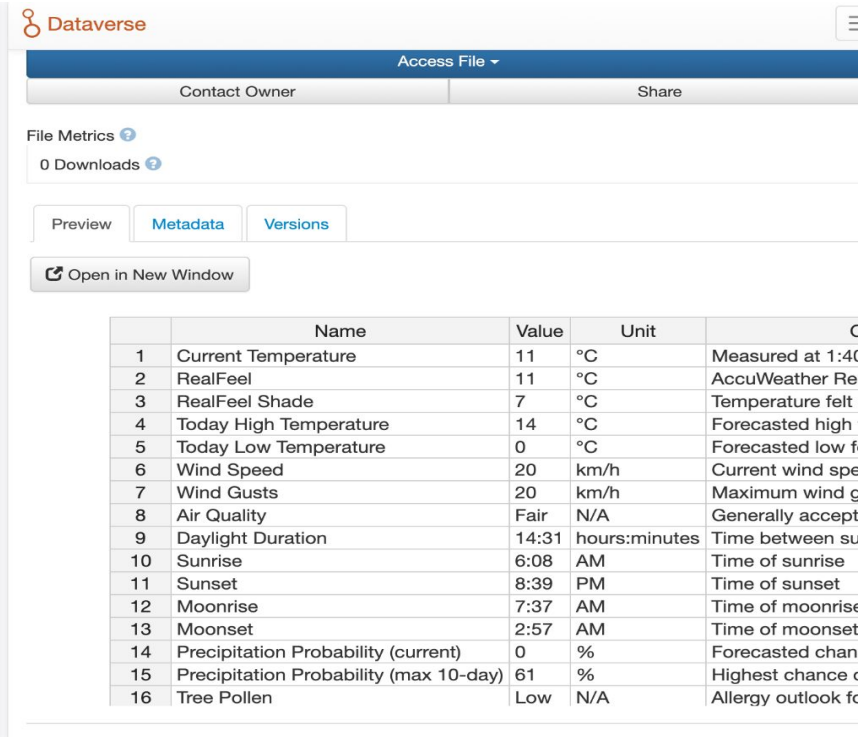
Intro: AI assistant for Kaunas weather





Intro: AI assistant for Kaunas weather in Dataverse



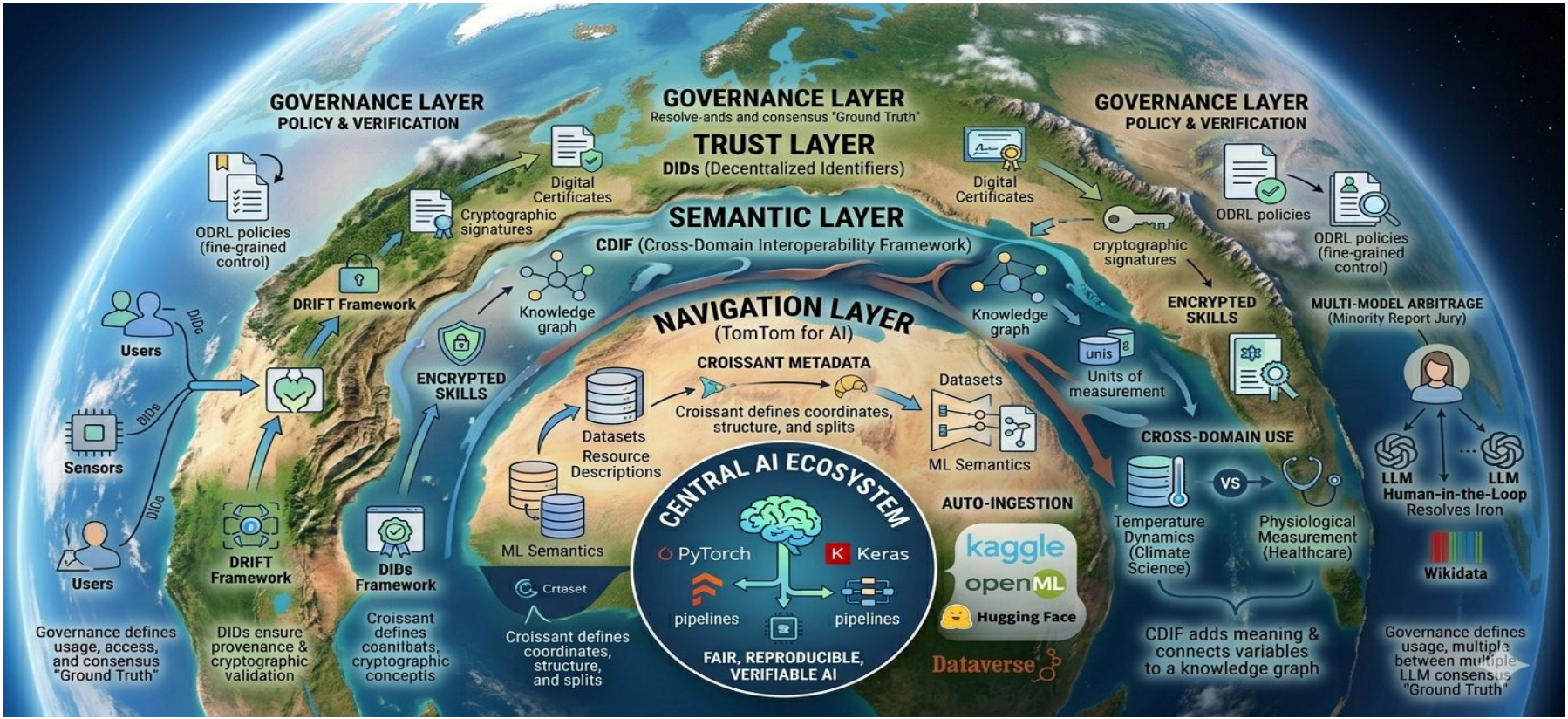


	Name	Value	Unit	C
1	Current Temperature	11	°C	Measured at 1:40
2	RealFeel	11	°C	AccuWeather Re
3	RealFeel Shade	7	°C	Temperature felt
4	Today High Temperature	14	°C	Forecasted high
5	Today Low Temperature	0	°C	Forecasted low f
6	Wind Speed	20	km/h	Current wind spe
7	Wind Gusts	20	km/h	Maximum wind g
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 fo

[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:

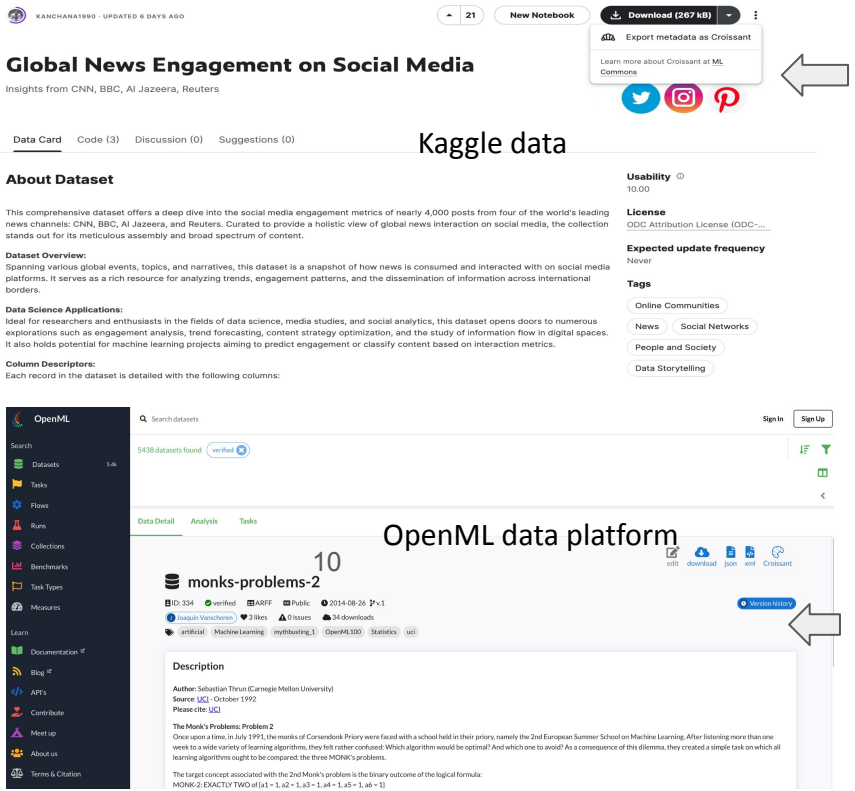
- Omar Benjelloun (Google),
- Elena Simperl (King's College London & ODI),
- Pierre Marcenac (Google),
- Pierre Ruyssen (Google),
- Costanza Conforti (Google),
- Michael Kuchnik (Meta),
- Jos van der Velde (Open ML),
- Luis Oala (Dotphoton),
- Steffen Vogler (Bayer),
- Mubashara Akthar (King's College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



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.

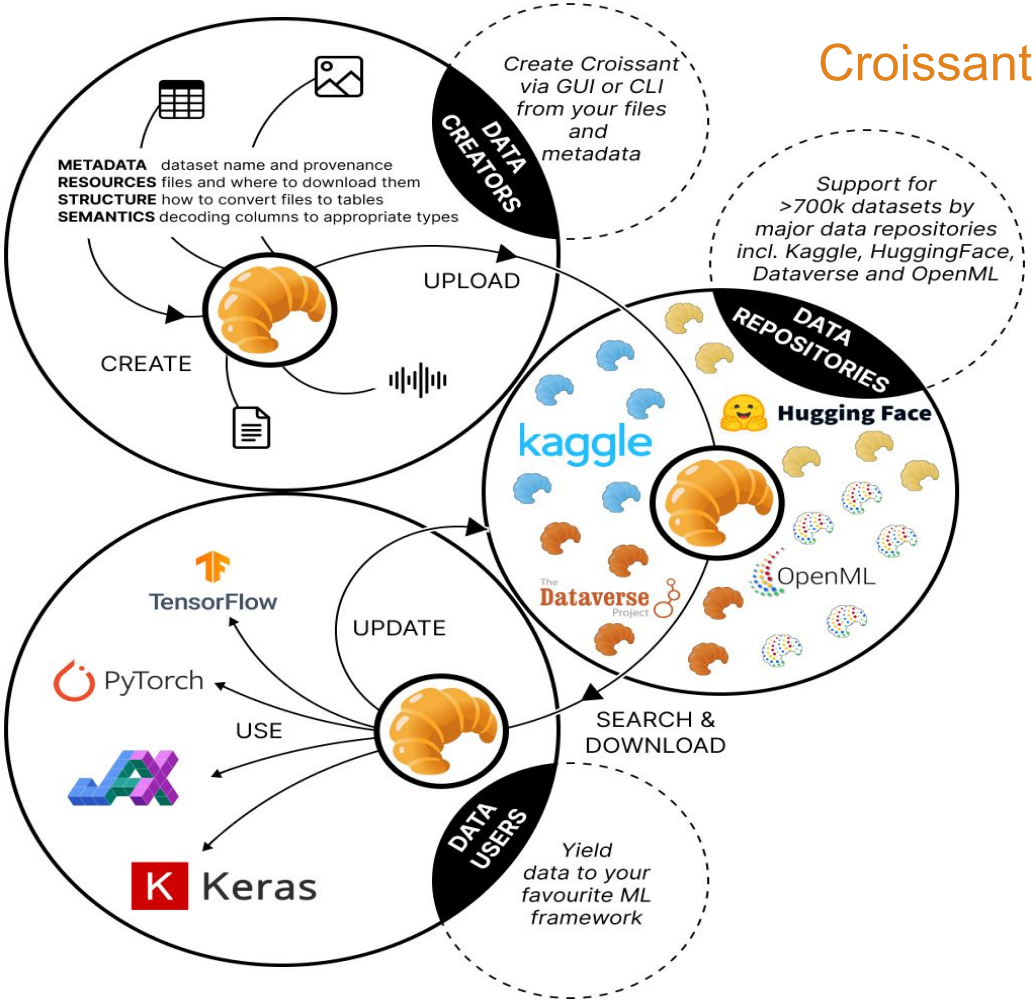
What makes Datasets used in Machine Learning special?

- 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

scan to access slides and links



Create

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

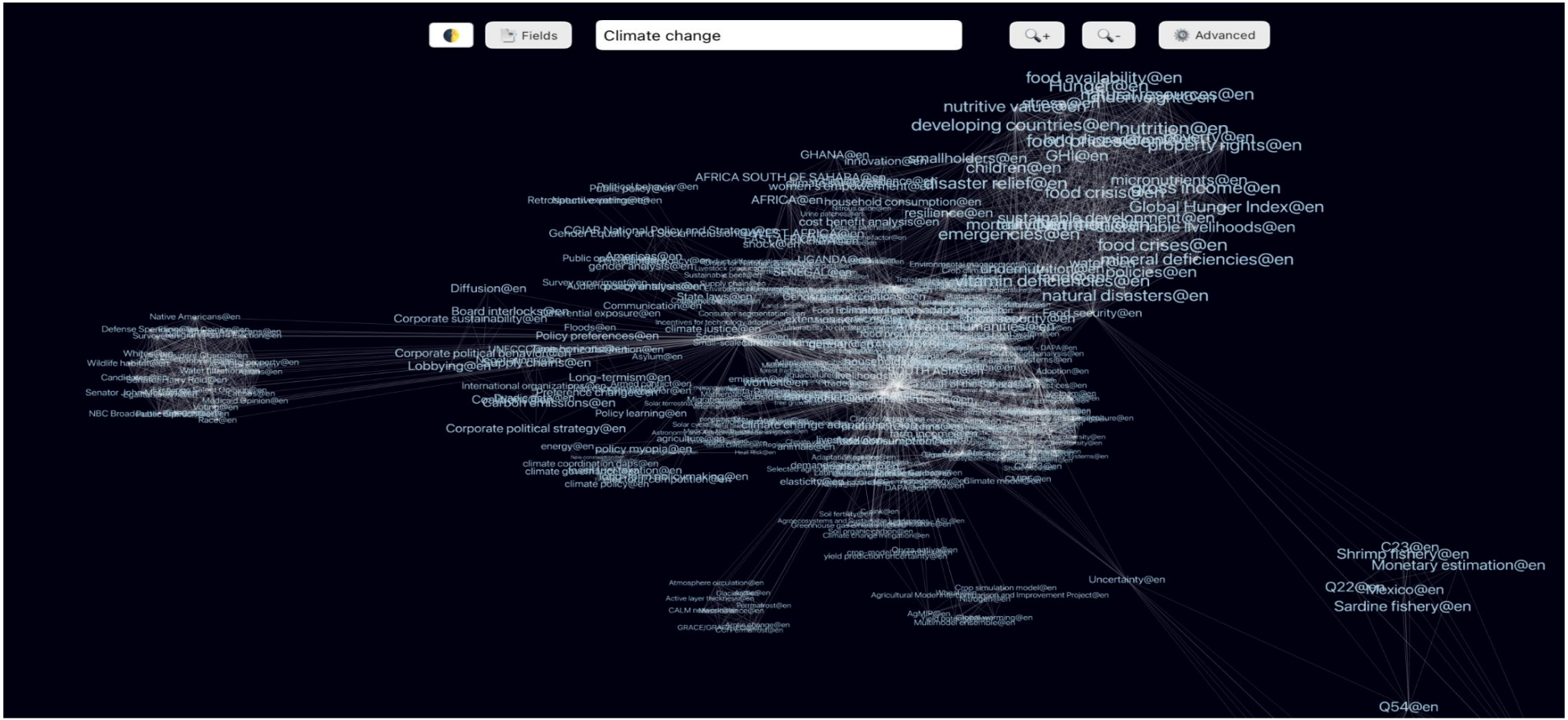
Discover and find

- 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.

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": "|"}
        }
    }
  ]
}
```

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

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

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’.

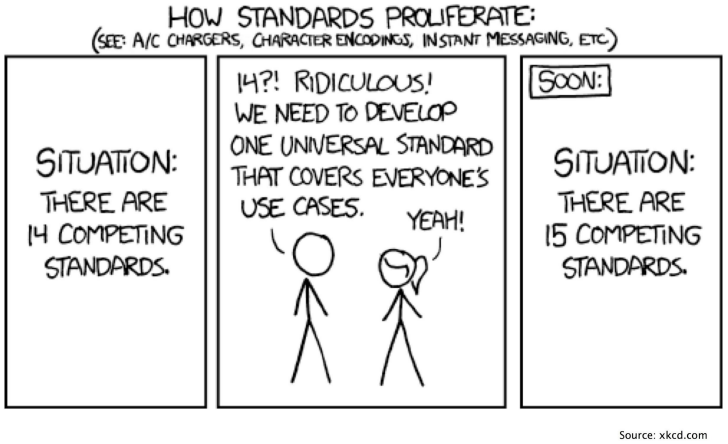
Led directly to the EOSC Interoperability Framework.

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

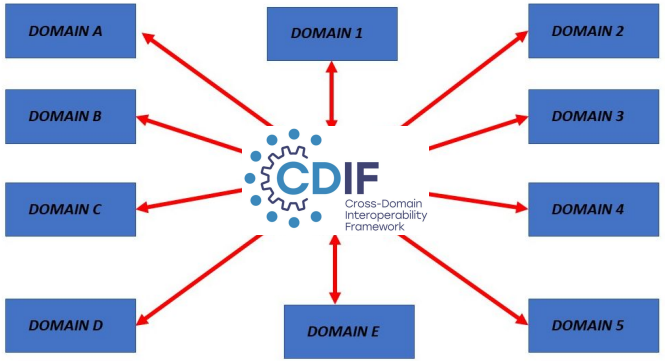
Core driver of CODATA work on the WorldFAIR project and WorldFAIR+ initiative, and on the Cross-Domain Interoperability Framework (CDIF).

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*.**



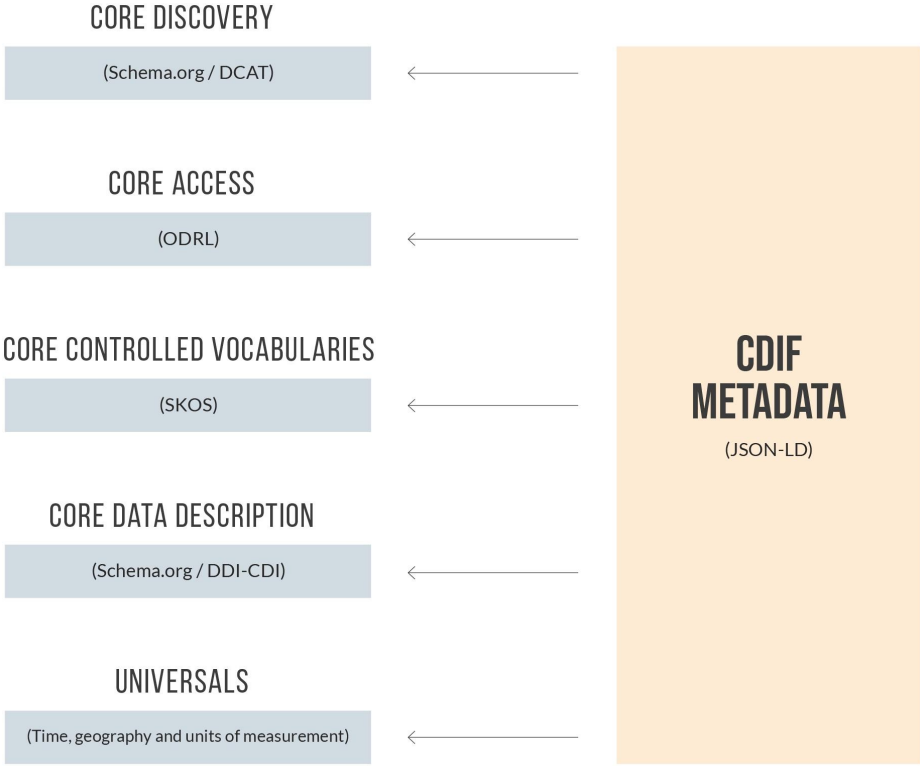
Source: xkcd.com



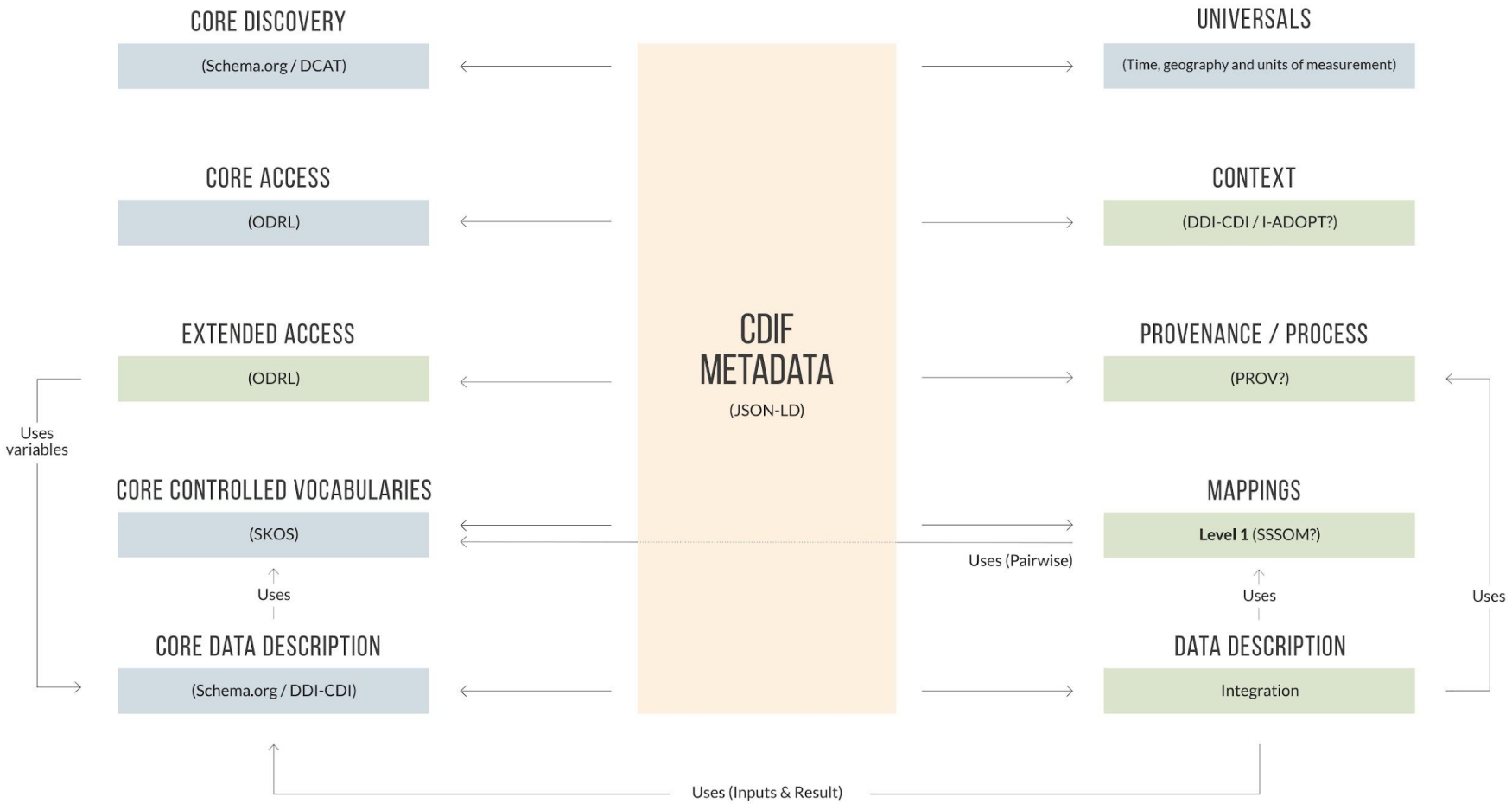
What is CDIF?

- 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
Cross-Domain Interoperability Framework



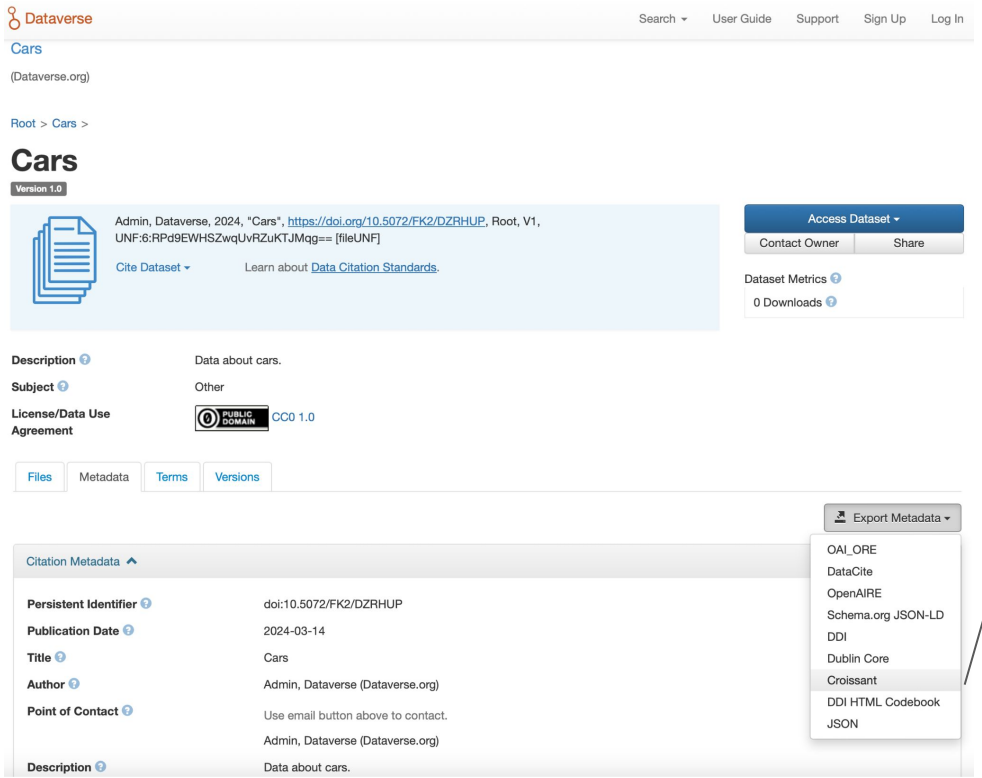
CDIF, Next Steps



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

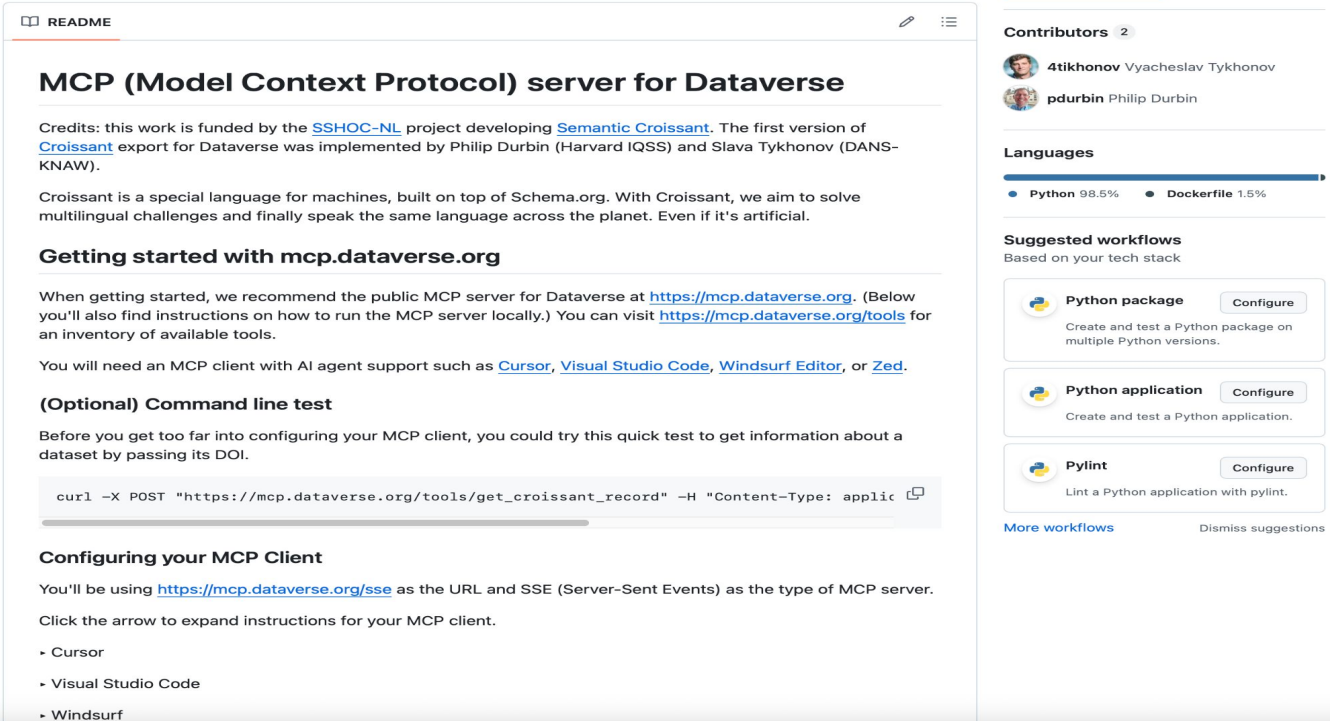


Mappings

Croissant exporter

Code

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

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

ODISSEI

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:

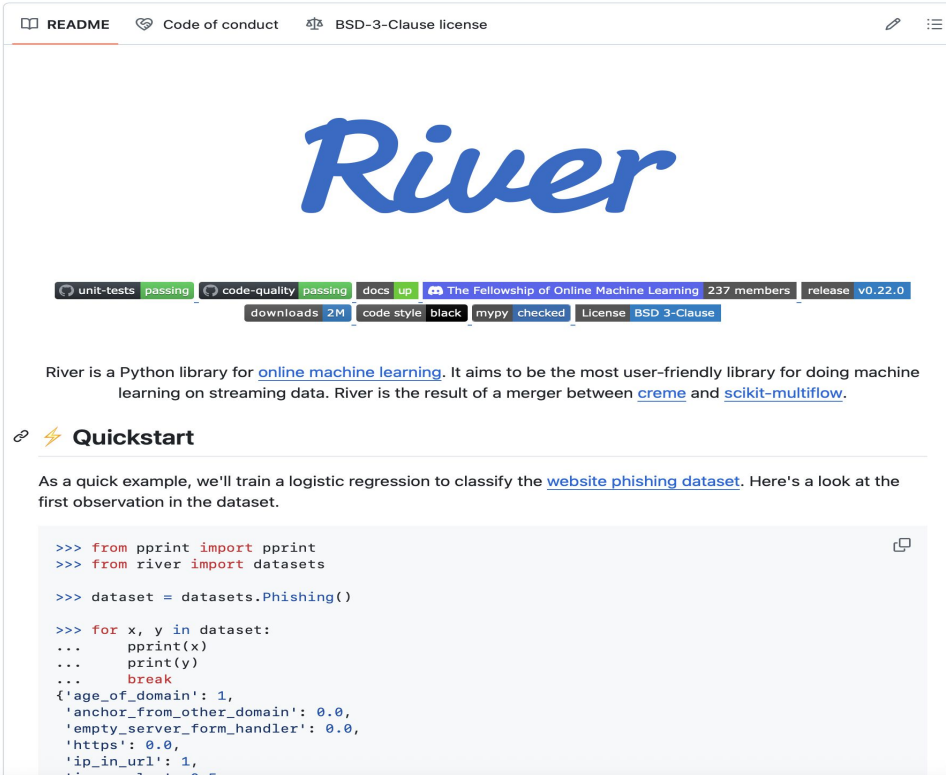
- **Infrastructure Drift** – Changes in the software environment that can break or invalidate infrastructure configurations.
- **Structural Drift** – Changes to the data schema that may invalidate databases or data contracts.
- **Semantic Drift** – Changes in the meaning of data without altering its 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



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.[1] 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).[3][2]

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 the 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 archived with global decentralized identifier - DID.

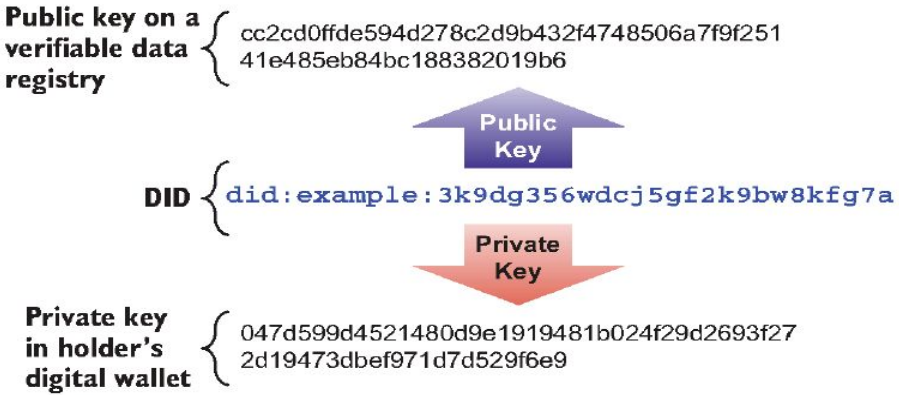
The resolution of the decentralized identifier (DID) is **cryptographically verifiable to prove the identity and the ownership** of that identifier and can support Model Context Protocol with sustainable infrastructure to keep provenance and origins of prompts, resources and tools.

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

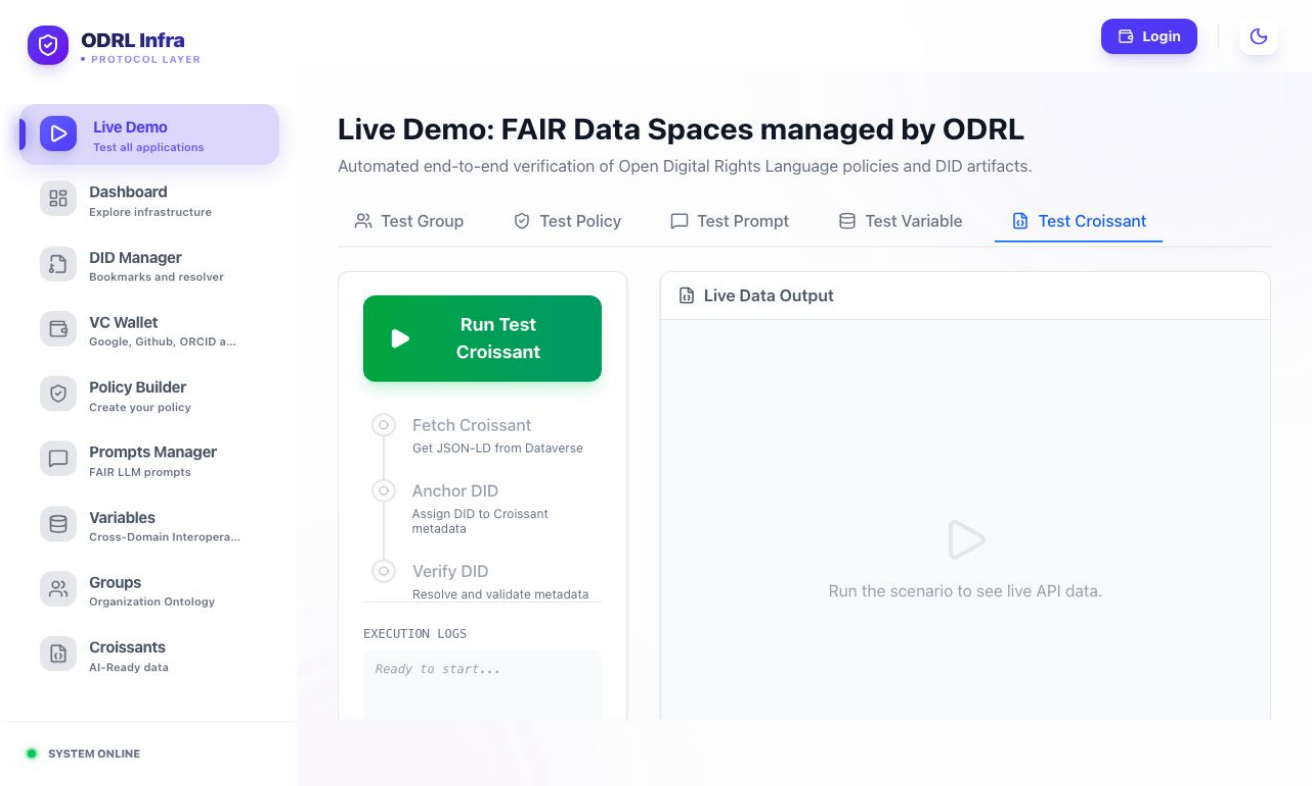


EXAMPLE 20: Usage of the service property

```
{
  "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

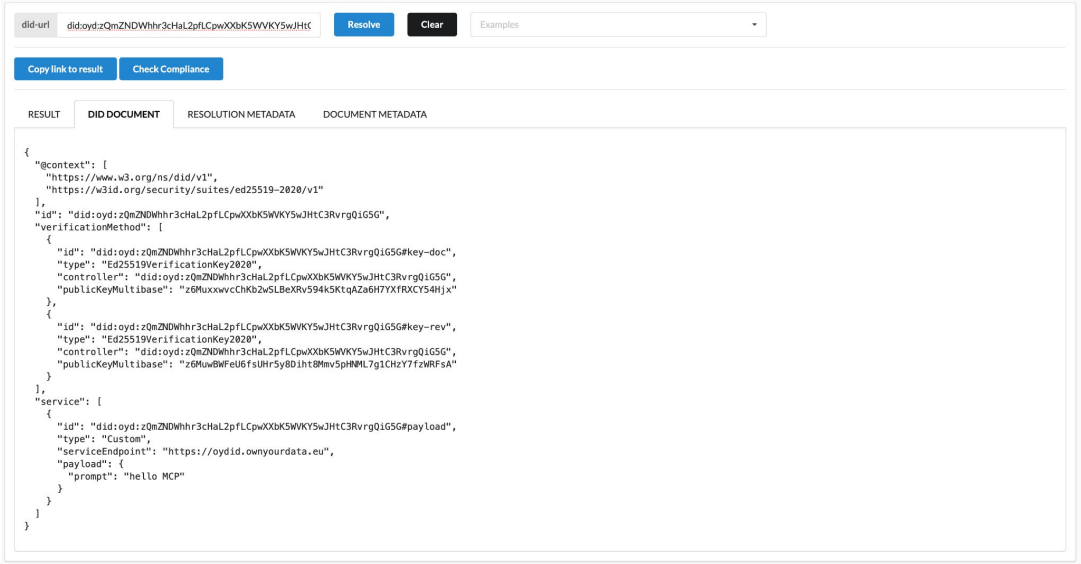


https://odrl.dev.codata.org/de

Universal Resolver for AI primitives

Registrar output:


```
{"didState":{"did":"did:oyd:zQmZNDWhhr3cHaL2pfLCpwXXbK5WVKY5wJHtC3RvrgQiG5G","state":"finished","secret":{"documentKey":"z1S5Z4B25YCwiGKYuFwAppS31ioGK3BcDc2HkR2ckXVwQae4","revocationKey":"z1S5YBrmeShGQwGPNyTEGXfGNLiEL2HN9zy1Y4TXmtBjiAfC","revocationLog":{"ts":1749461892,"op":1,"doc":"zQmP9FPNn8HWTHWMJExcXvGhBzGjGsDjSMfCxirgNzdKcCV","sig":"z4pmbitSijGHkvqYT1qGkqYemkrXDBMSpS37cvS6banJ2L9Lc4wUaADcjwsBca8VJG7VmRezmmVHVgnLcJcGZq6rR"}},"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":"z6MuxxwvcChKb2wSLBeXRv594k5KtqAZa6H7YXfRXCY54Hjx"},{"id":"did:oyd:zQmZNDWhhr3cHaL2pfLCpwXXbK5WVKY5wJHtC3RvrgQiG5G#key-rev","type":"Ed25519VerificationKey2020","controller":"did:oyd:zQmZNDWhhr3cHaL2pfLCpwXXbK5WVKY5wJHtC3RvrgQiG5G","publicKeyMultibase":"z6MuwBWFeU6fsUHr5y8Diht8Mmv5pHNML7g1CHzY7fzWRFsA"}],"service":[{"id":"did:oyd:zQmZNDWhhr3cHaL2pfLCpwXXbK5WVKY5wJHtC3RvrgQiG5G#payload","type":"Custom","serviceEndpoint":"https://oydid.ownyourdata.eu","payload":{"prompt":"hello MCP"}}]}},"didRegistrationMetadata":{},"didDocumentMetadata":{"did":"did:oyd:zQmZNDWhhr3cHaL2pfLCpwXXbK5WVKY5wJHtC3RvrgQiG5G","registry":"https://oydid.ownyourdata.eu","log_hash":"zQmT6wtciv8n1d4maP6BcyLrXFZykK6D7LZjVCxjfNQ7hZH","log":[{"ts":1749461892,"op":2,"doc":"zQmZNDWhhr3cHaL2pfLCpwXXbK5WVKY5wJHtC3RvrgQiG5G","sig":"z3av6sHSPzZM2Loqq3GuGwMAPoHpkPgYLYZXNzbuKbGckRTRa4WwX4oM6LNCvAZxeD8ueuWCFsDUySpceo7hmKs9S","previous":[]},{"ts":1749461892,"op":0,"doc":"zQmSUFr1jKnWCGga5eDmEsbUW5sfTyngkTcyPGQVwM45hyD","sig":"z2kyjucLcNuvGJEkBCx3isfkvfvNpyiDanTSbzxHHXbBxha5y8KbQ1knziWaPMevUgCA1PPSXwptyHPWyAEw3kR6h","previous":[]}]}}
```

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

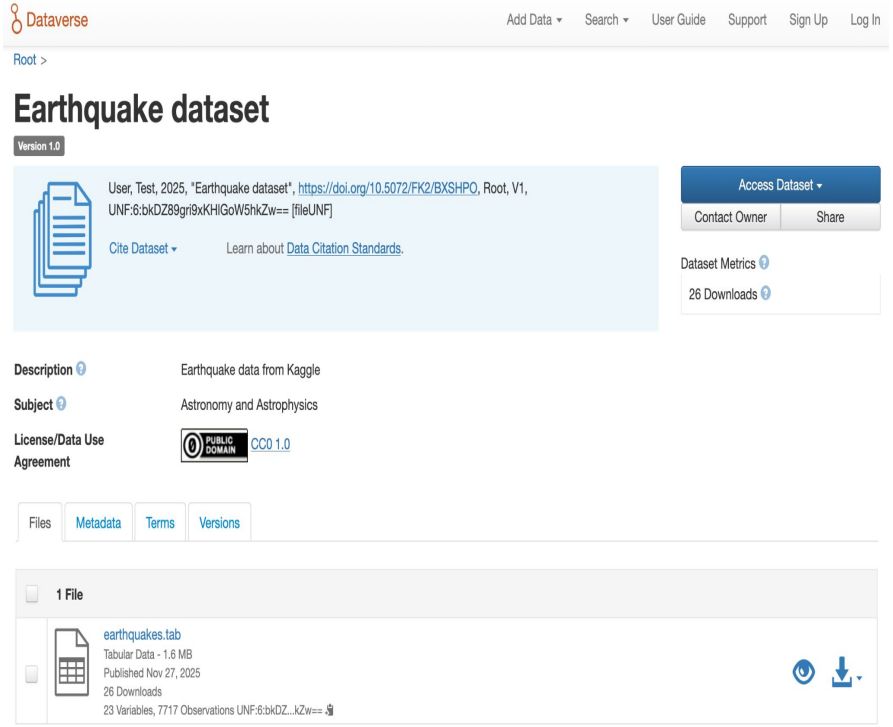
Decentralized identifier (DID):

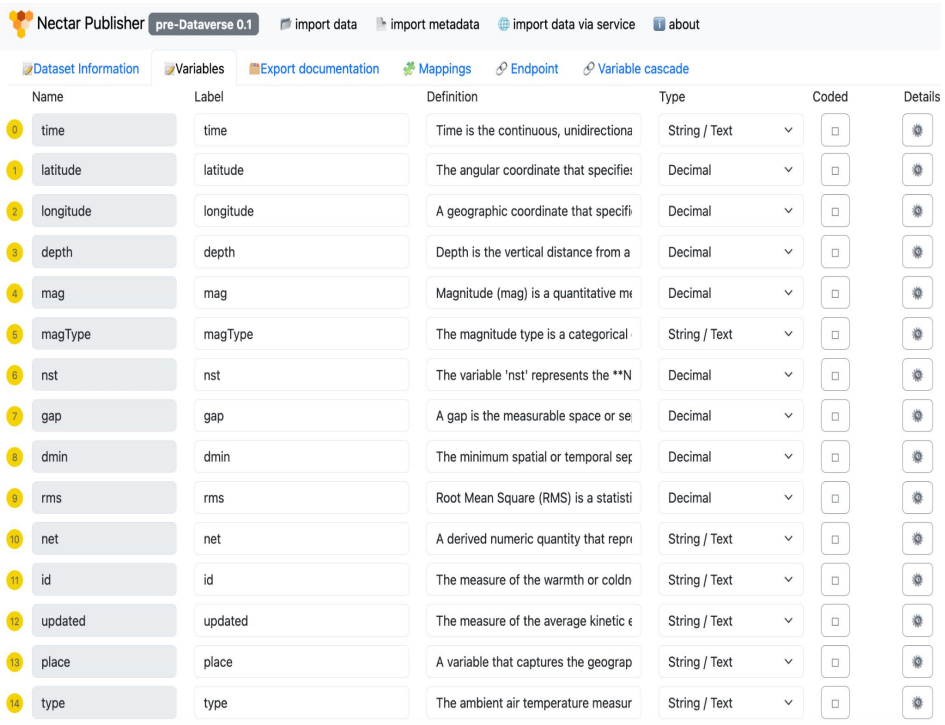
did:oyd:zQmZNDWhhr3cHaL2pfLCpwXXbK5WVKY5wJHtC3RvrgQiG5G

curl https://dev.uniresolver.io/1.0/identifiers/did:oyd:zQmZNDWhhr3cHaL2pfLCpwXXbK5WVKY5wJHtC3RvrgQiG5G

```
 "service": [
   {
     "id":
"did:oyd:zQmZNDWhhr3cHaL2pfLCpwXXbK5WVKY5
wJHtC3RvrgQiG5G#payload",
     "type": "Custom",
     "serviceEndpoint":
"https://oydid.ownyourdata.eu",
     "payload": {
       "prompt": "hello MCP"
     }
   }
 ]
```

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





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 CESSDA DATA CATALOGUE (CDC) DDI3.3 PROFILE

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"). 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 specifed 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 specifed 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: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

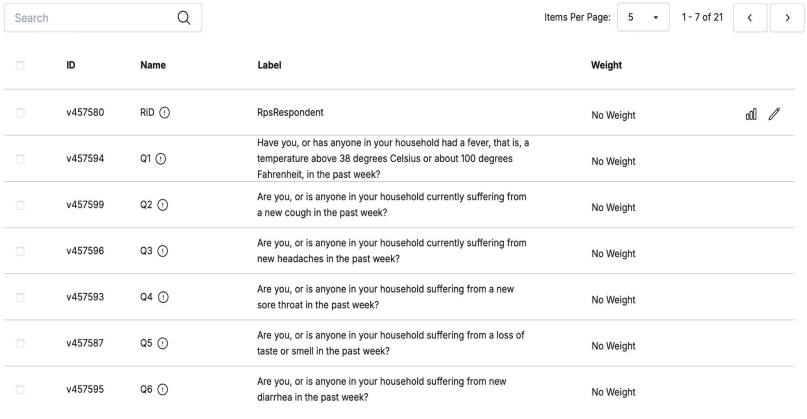
(attributes, categories, units of measurements, …)

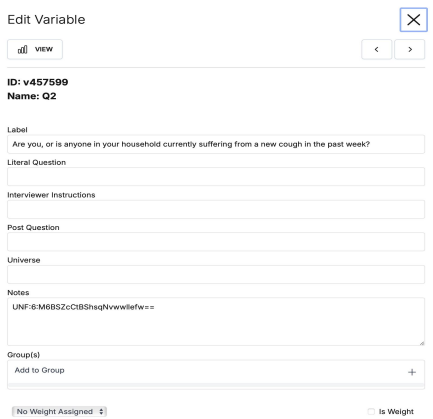

```
<var ID="v63820" name="gender" intrvl="contin">
  <location fileid="f101579"/>
  <catgry>
    <catValu>1</catValu>
    <labl level="category">male, or</labl>
  </catgry>
  <catgry>
    <catValu>9</catValu>
    <labl level="category">RF</labl>
  </catgry>
  <catgry>
    <catValu>2</catValu>
    <labl level="category">female</labl>
  </catgry>
  <varFormat type="numeric"/>
  <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"/>
  <catgry>
    <catValu>2</catValu>
    <labl level="category">no</labl>
  </catgry>
  <catgry>
    <catValu>9</catValu>
    <labl level="category">RF</labl>
  </catgry>
  <catgry>
    <catValu>8</catValu>
    <labl level="category">DK</labl>
  </catgry>
  <catgry>
    <catValu>1</catValu>
    <labl level="category">yes</labl>
  </catgry>
  <varFormat type="numeric"/>
  <notes subject="Universal Numeric Fingerprint" level="variable" type="Dataverse:UNF">UNF:3:oueQkfY5y0dpXf9yYmI6MQ==</notes>
</var>
<var ID="v63876" name="marit_b" intrvl="contin">
  <location fileid="f101579"/>
  <catgry>
    <catValu>5</catValu>
    <labl level="category">divorced, or</labl>
  </catgry>
  <catgry>
    <catValu>6</catValu>
    <labl level="category">never married</labl>
  </catgry>
  <catgry>
    <catValu>3</catValu>
    <labl level="category">widowed</labl>
  </catgry>
```


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 into 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?

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

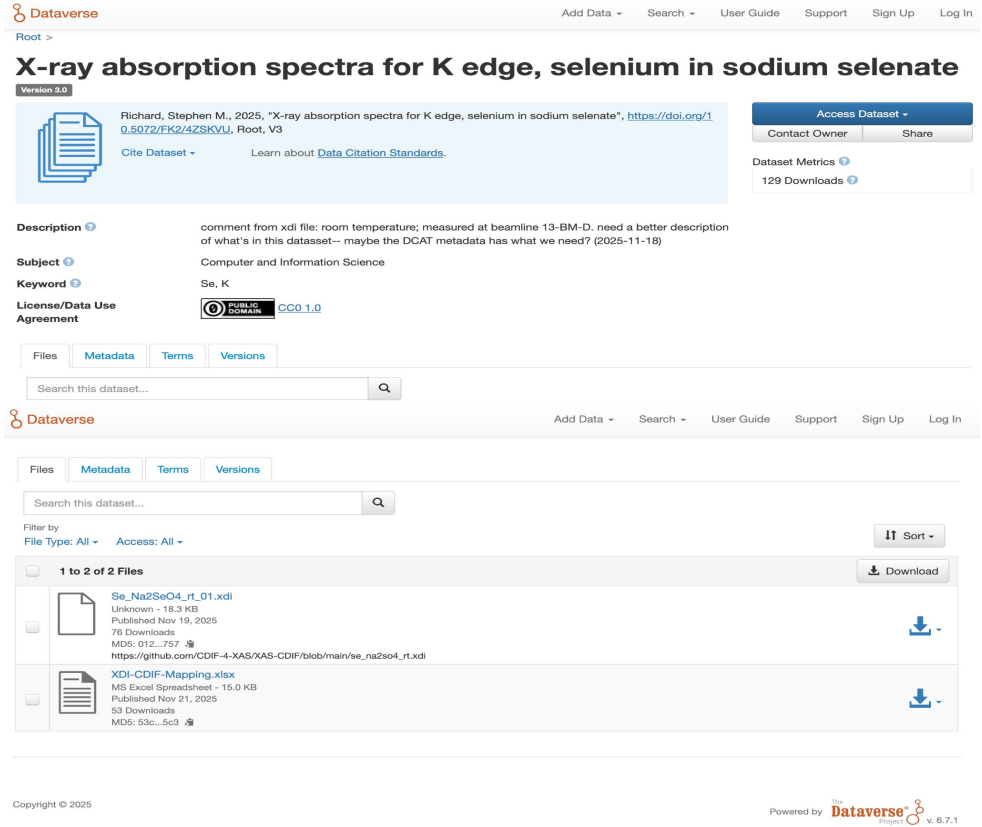
Output

Code | Blame | 145 lines (145 loc) · 3.38 KB | Raw

```
{
  "@context": "https://ddi-alliance.bitbucket.io/DDI-CDI/DDI-CDI_v1.0-rc1/encoding/json-ld/ddi-cdi.jsonld",
  "@graph": [
    {
      "@id": "#Frequency",
      "@type": "InstanceVariable",
      "name": "Frequency",
      "hasIntendedDataType": "http://rdf-vocabulary.ddialliance.org/cv/DataType/1.1.2/#String"
    },
    {
      "@id": "#Year",
      "@type": "InstanceVariable",
      "name": "Year",
      "hasIntendedDataType": "http://rdf-vocabulary.ddialliance.org/cv/DataType/1.1.2/#Integer"
    },
    {
      "@id": "#Age_Cohort",
      "@type": "InstanceVariable",
```

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

Dataset in Dataverse represented in CDIF



Mappings

	xdi dictionary	CDIF implementation		
2	Beamline	"prov:used":[{ "@type":["schema:Thing", "prov:Entity","xas:Beamline"],		
3	Beamline.collimation	"schema:additionalProperty": [{"@type": "schema:PropertyValue", "schema:propertyID": {"@id": "xas:collimation"}, "schema:value": "(value)" },	Beamline	xdi
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

“Interlink”: Semantic Croissant with WikiData prediction

Variable	Linked Wikidata ID	Label (English)	Description (excerpt)	Relevance Note
energy	Q11379	*energy*	“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	*angle*	“something that is formed when two rays meet at a single or same point”	Relevant – matches the definition of a monochromatic angle measurement.
i0	Q54897085	*I3-iPSC* (cell line)	No direct match; the best available entity is unrelated to intensity monitoring.	**Not relevant** – no appropriate Wikidata entry for “monitor intensity”.
itrans	Q109607869	*iTransmission, Inc.*	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	*iReference: Using Apple’s iPhone as a Reference Tool*	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 *keV*; 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 *counts* 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 *photons* 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

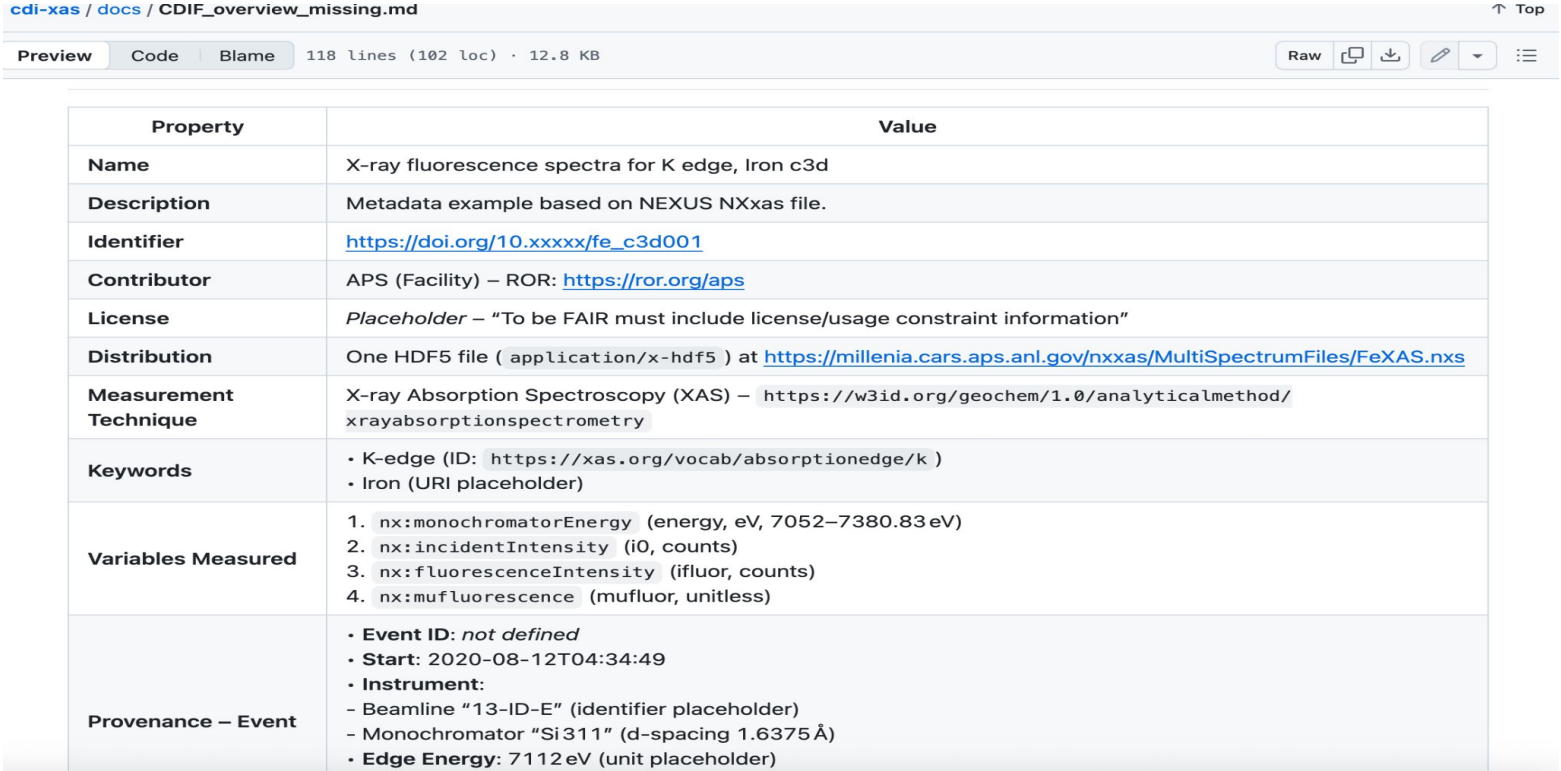
Energy is the capacity to do work or perform tasks. It is a fundamental concept in physics and is often measured in units such as joules or kilowatt-hours. Energy can be transferred from one object to another, and can be transformed from one form to another. It is essential for powering machines, lighting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

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

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

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	*Placeholder* – “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	• K-edge (ID: `https://xas.org/vocab/absorptionedge/k`) • Iron (URI placeholder)
Variables Measured	1. `nx:monochromatorEnergy` (energy, eV, 7052–7380.83 eV) 2. `nx:incidentIntensity` (i0, counts) 3. `nx:fluorescenceIntensity` (ifluor, counts) 4. `nx:mufluorescence` (mufluor, unitless)
Provenance – Event	• **Event ID**: *not defined* • **Start**: 2020-08-12T04:34:49 • **Instrument**: - Beamline “13-ID-E” (identifier placeholder) - Monochromator “Si 311” (d-spacing 1.6375 Å) • **Edge Energy**: 7112 eV (unit placeholder)

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.

CDIF
Cross-Domain Interoperability Framework

RO-Crate

ClimateAdapt | eosc

Questions

Slava Tykhonov, CODATA
slava@codata.org

Listen Croissant song!

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