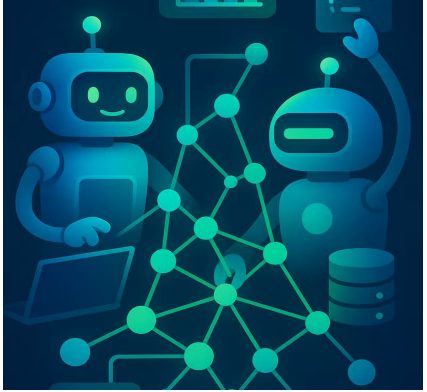


Pale Fire: Technical Deep Dive

Graph Exchange: Context Graph, Annotated Knowledge, and Vector+GraphRAG for More Accurate Q&A

Intelligent Knowledge Graph Research & Search Ecosystem

Slava Tykhonov / Alexy Khrabrov

San Francisco, California
Tuesday, March 17, 2026

Pale Fire big vision

- *Transforming disparate data silos into interconnected knowledge.*

- Knowledge Operating System: Seamlessly converts PDFs, URLs, and Spreadsheets into a high-fidelity entity relationship graph.
- Semantic Structure: Maps content based on conceptual weight and relationships rather than simple keyword matches.
- Machine Learning: Use Croissant for ML as a common Semantic Web “language” between AI models, data and interfaces.
- Narrative Synthesis: Employs LLMs to explain the 'WHY' behind entity connections in broader research contexts.
- Enterprise Readiness: Production-grade deployment with Docker and a persistent AI Agent Daemon.

Croissant 1.0 for Machine Learning

Croissant Format Specification

Version 1.0

Published: 2024/03/01

http://mlcommons.org/croissant/1.0

Authors:

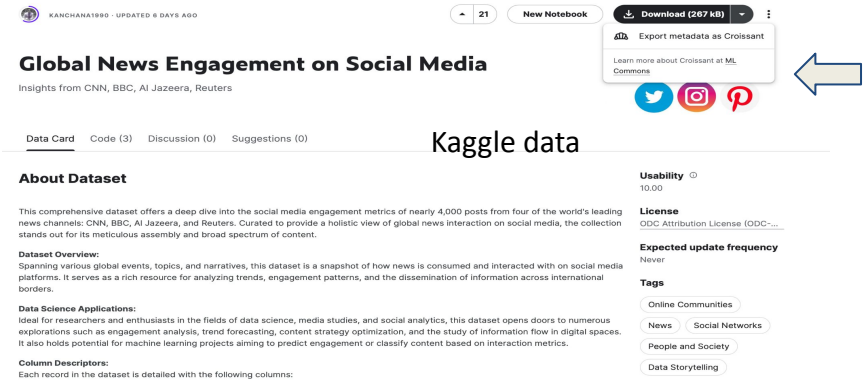
- Omar Benjelloun (Google),
- Elena Simperl (King's College London & ODI),
- Pierre Marcenac (Google),
- Pierre Ruyssen (Google),
- Costanza Conforti (Google),
- Michael Kuchnik (Meta),
- Jos van der Velde (Open ML),
- Luis Oala (Dotphoton),
- Steffen Vogler (Bayer),
- Mubashara Akthar (King's College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

Introduction

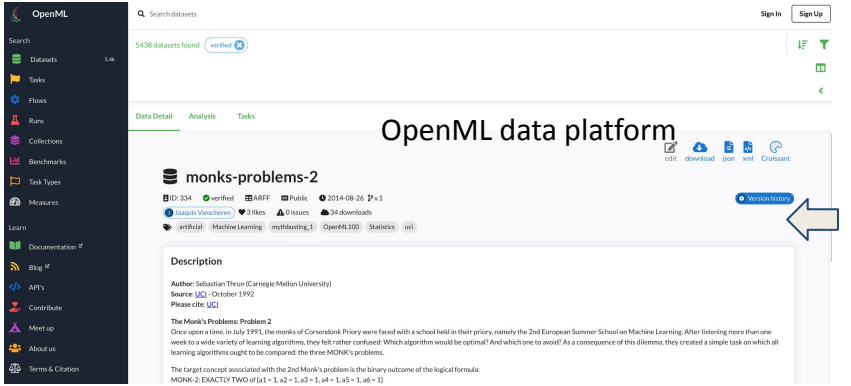
Datasets are the basis of machine learning (ML). However, a lack of standardization in the description and semantics of ML datasets has made it increasingly difficult for researchers and practitioners to explore, understand, and use all but a small fraction of popular datasets.

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

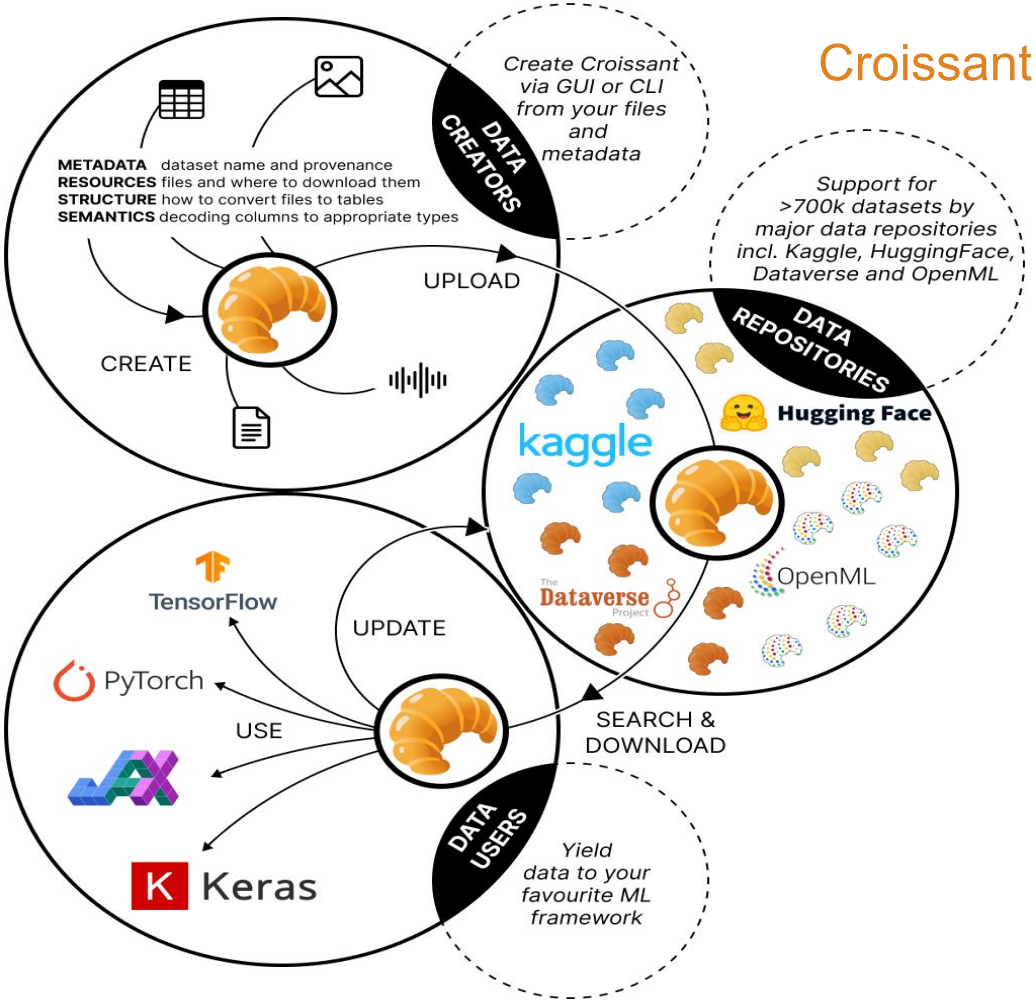


Kaggle data



OpenML data platform

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

Multilingual Semantic Croissant

croissant > {} 4_mysterious_discoveries_-_final_repeat.jsonld > ...


```
{
  "@context": {
    "@language": "en",
    "@vocab": "https://schema.org/",
    "cr": "http://mlcommons.org/croissant/",
    "sc": "https://schema.org/",
    "dct": "http://purl.org/dc/terms/",
    "conformsTo": "dct:conformsTo",
    "recordSet": "cr:recordSet",
    "field": "cr:field",
    "source": "cr:source",
    "fileObject": "cr:fileObject",
    "fileSet": "cr:fileSet",
    "dataType": {
      "@id": "cr:dataType",
      "@type": "@id"
    }
  },
  "@type": "sc:Dataset",
  "name": {
    "@language": "en",
    "@value": "4 Mysterious Discoveries - Final Repeat"
  },
  "description": {
    "@language": "en",
    "@value": "Listen closely. If even one of the artifacts in this video is real, then [music] our entire history
  },
  "url": "https://www.youtube.com/watch?v=y9RJRwYtiTI",
  "license": "CC-BY-4.0",
  "dct:conformsTo": "http://mlcommons.org/croissant/1.0",
  "distribution": [
    {
      "@type": "cr:FileObject",
      "name": "source-content",
      "contentUrl": "/Users/vyacheslavtykhonov/projects/croissant-toolkit/data/transcripts/y9RJRwYtiTI.txt",
      "encodingFormat": "text/plain"
    }
  ],
  "recordSet": [],
  "spatialCoverage": [
    {
      "@language": "en",
      "@value": "Nampa"
    },
    {
      "@language": "en",
      "@value": "Idaho"
    },
    {
      "@language": "en",
      "@value": "New Zealand"
    },
    {
      "@language": "en",
      "@value": "Moeraki Boulders"
    },
```

```
  ],
  "temporalCoverage": [
    {
      "@language": "en",
      "@value": "summer of 1889"
    },
    {
      "@language": "en",
      "@value": "2015"
    }
  ],
  "keywords": [
    {
      "@language": "en",
      "@value": "a few thousand years ago"
    },
    {
      "@language": "en",
      "@value": "millions of years old"
    },
    {
      "@language": "en",
      "@value": "summer of 1889"
    },
    {
      "@language": "en",
      "@value": "Nampa"
    },
    {
      "@language": "en",
      "@value": "Idaho"
    },
    {
      "@language": "en",
      "@value": "300 ft"
    },
    {
      "@language": "en",
      "@value": "New Zealand"
    },
    {
      "@language": "en",
      "@value": "Moeraki Boulders"
    },
    {
      "@language": "en",
      "@value": "7 ft across"
    },
    {
      "@language": "en",
      "@value": "several tons"
    },
    {
      "@language": "en",
      "@value": "China"
```

```
  "keywords": [
    {
      "@language": "en",
      "@value": "Gate of the Sun"
    },
    {
      "@language": "en",
      "@value": "10 tons"
    },
    {
      "@language": "en",
      "@value": "48 square symbols"
    },
    {
      "@language": "en",
      "@value": "365.2422 days"
    },
    {
      "@language": "en",
      "@value": "14,000 years ago"
    },
    {
      "@language": "en",
      "@value": "the last ice age"
    },
    {
      "@language": "en",
      "@value": "the glacial melt"
    },
    {
      "@language": "en",
      "@value": "Glasgow"
    },
    {
      "@language": "en",
      "@value": "Scotland"
    },
    {
      "@language": "en",
      "@value": "40 ft wide"
    },
    {
      "@language": "en",
      "@value": "Cochno Stone"
    },
    {
      "@language": "en",
      "@value": "1800s"
    },
    {
      "@language": "en",
      "@value": "1960s"
    },
    {
      "@language": "en",
      "@value": "2015"
    }
```

Use case: spreadsheets to structured data

UC2 - OPENHIDRA - Dataset and Service Requirements .XLSX

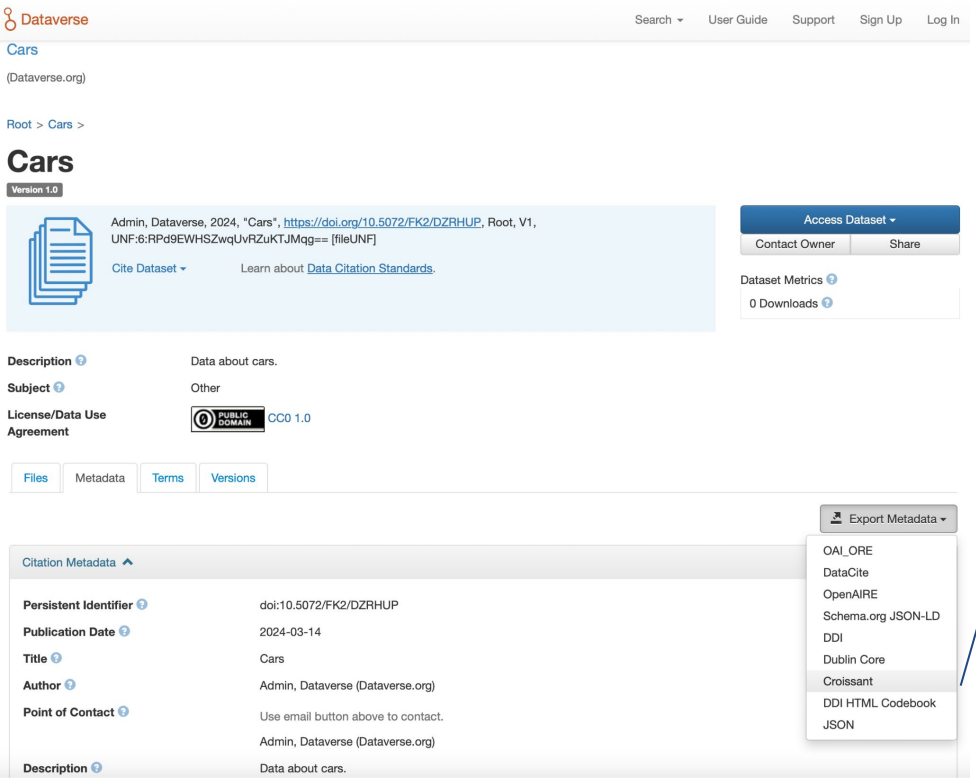
File Edit View Insert Format Data Tools Help

A1 Domain

	A	B	C	D	E	F	G	H
1	Domain	CLIMATE-ADAPT 4 EOSC Info						
2	Title	Partner	Use Case	Service	Dataset Source Type	Dataset Name	Data Format	Data Standard
3	Description	Indicate the parter responsible for the use case	Location used for Use Case	Indicate the service that currently uses the dataset	Service/File/Database	Name or origin of data	Data Format	Data Standard
4	E.g.	LNEC	OPENHidra	OPENCoasts		Climate File	CSV	
5	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPENCoasts	Service	Data Marine Copernicus	NetCDF	
6	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPENCoasts	Service	Data Marine Copernicus	NetCDF	
7	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPENCoasts	Service	Data Marine Copernicus	NetCDF	
8	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPENCoasts	Service	Data Marine Copernicus	NetCDF	
9	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPENCoasts	Service	Data Marine Copernicus	NetCDF	
10	Aveiro/OPENHIDRA	LNEC	Aveiro	OPENHIDRA/OPENCoasts	Service	Data Marine Copernicus	NetCDF	

Starting point: Spreadsheet link or Web page

Harvard Dataverse data repository



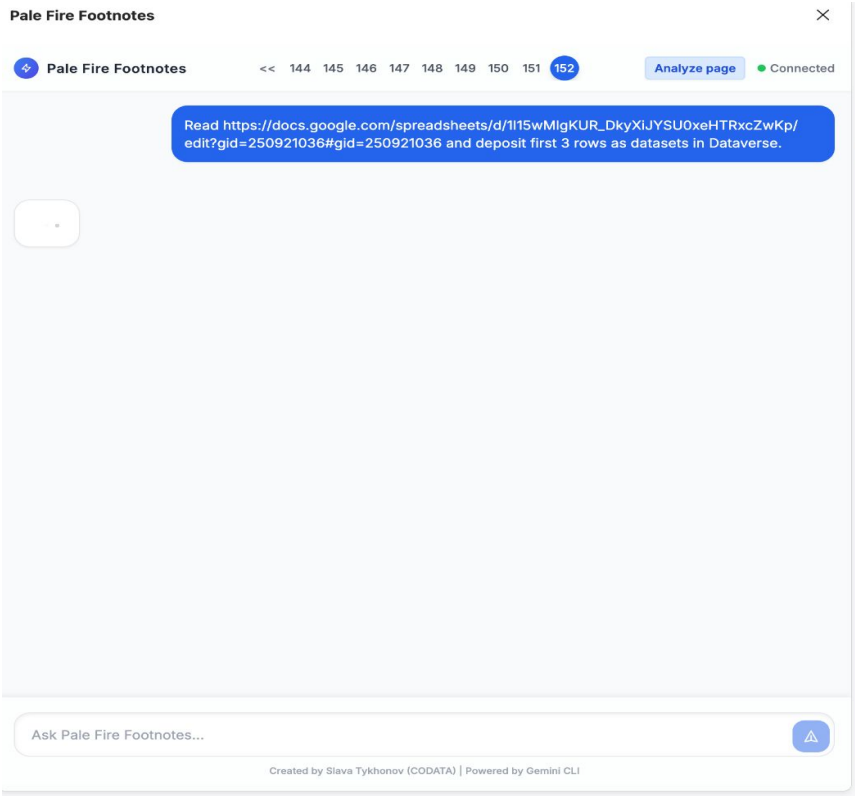
Mappings

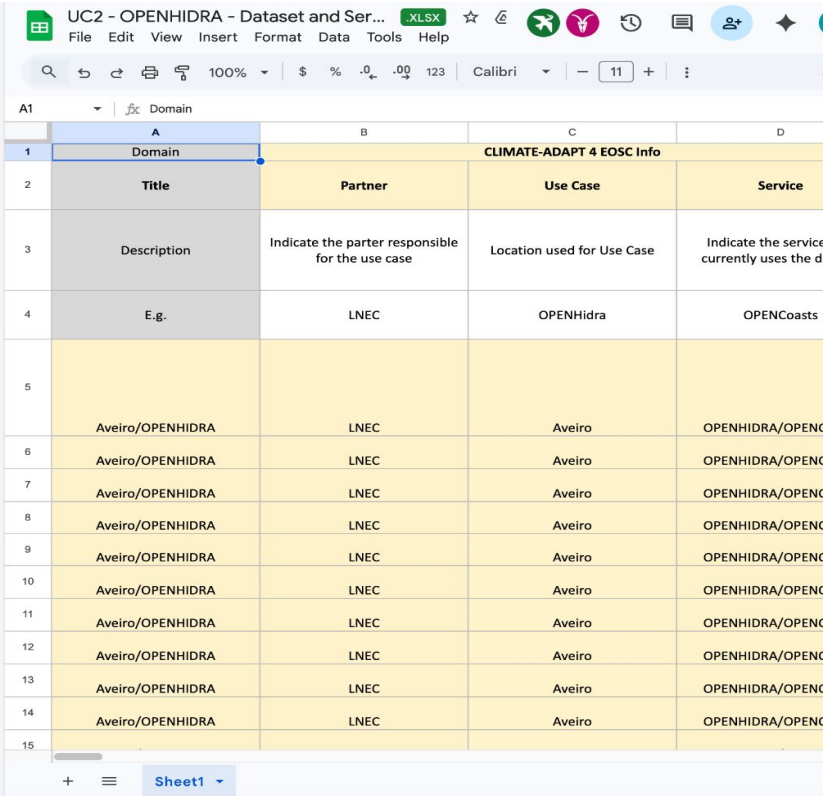
Croissant exporter

Code


```
@context:
  @language:                    "en"
  @vocab:                       "https://schema.org/"
  citeAs:                       "cr:citeAs"
  column:                       "cr:column"
  conformsTo:                   "dct:conformsTo"
  cr:                           "http://mlcommons.org/croissant/"
  data:
    @id:                        "cr:data"
    @type:                      "@json"
  dataBiases:                   "cr:dataBiases"
  dataCollection:               "cr:dataCollection"
  dataType:
    @id:                        "cr:dataType"
    @type:                      "@vocab"
  dct:                          "http://purl.org/dc/terms/"
  extract:                      "cr:extract"
  field:                        "cr:field"
  fileProperty:                 "cr:fileProperty"
  fileObject:                   "cr:fileObject"
  fileSet:                      "cr:fileSet"
  format:                       "cr:format"
  includes:                     "cr:includes"
  isEnumeration:                "cr:isEnumeration"
  jsonPath:                     "cr:jsonPath"
  key:                          "cr:key"
  md5:                          "cr:md5"
  parentField:                  "cr:parentField"
  path:                         "cr:path"
  personalSensitiveInformation: "cr:personalSensitiveInformation"
  recordSet:                    "cr:recordSet"
  references:                   "cr:references"
  regex:                        "cr:regex"
  repeated:                     "cr:repeated"
  replace:                      "cr:replace"
  sc:                           "https://schema.org/"
  separator:                    "cr:separator"
  source:                       "cr:source"
  subField:                     "cr:subField"
  transform:                    "cr:transform"
  wd:                           "https://www.wikidata.org/wiki/"
@type:                          "sc:Dataset"
conformsTo:                     "http://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:                        "7b1201ce6b469796837a035377338c5a"
    contentUrl:                 "stata13-auto.dta"
recordSet:
  0:
    @type:                      "cr:RecordSet"
    field:
      0:
        @type:                  "cr:Field"
        name:                   "make"
        description:            "Make and Model"
        dataType:               "sc:Text"
        source:
          @id:                  "6"
          fileObject:
            @id:                "stata13-auto.dta"
```

Pale Fire AI agent jobs: data depositor





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

Pale Fire jobs: data depositor

Pale Fire Footnotes

Pale Fire Footnotes << 144 145 146 147 148 149 150 151 152 Analyze page Connected

Read https://docs.google.com/spreadsheets/d/1l15wMlgKUR_DkyXiJYSU0xeHTRxcZwKp/edit?gid=250921036#gid=250921036 and deposit first 3 rows as datasets in Dataverse.

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

Deposited Datasets
Aveiro/OPENHIDRA - Sea level (m)
PID: doi:10.70122/FK2/RE2TCX
Description: Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological.
Aveiro/OPENHIDRA - Salinity (PSU - Practical Salinity Unit)
PID: doi:10.70122/FK2/DV7Q4B
Description: Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological.
Aveiro/OPENHIDRA - Temperature (degrees C)
PID: doi:10.70122/FK2/FUU9N9
Description: Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological.
All datasets were created in the `slava` collection on `demo.dataverse.org` with a **CC0 1.0** license.

Ask Pale Fire Footnotes...

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

Dataverse

Slava Tykhonov Dataverse

Demo Dataverse >

Contact Share Edit

Search this dataverse... Advanced Search Add Data

Dataverses (0)
Datasets (149)
Files (139)

Publication Year
2026 (116)

Publication Status
Published (110)
Draft (39)
Unpublished (33)

Author Name
CDIF Expert (49)
Pale Fire Agent (37)
Copernicus (EU) (30)
IPMA (7)
NOAA (7)
More...

Author Affiliation
CODATA (2)

Subject
Other (144)
Social Sciences (11)
Earth and Environmental Sciences (5)

Keyword Term

1 to 10 of 149 Results Sort

Aveiro/OPENHIDRA - Temperature (degrees C)
Mar 17, 2026
Copernicus (EU), 2026, "Aveiro/OPENHIDRA - Temperature (degrees C)", https://doi.org/10.70122/FK2/FUU9N9, Demo Dataverse, V1
Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological. Original Identifier: yeshttps://doi.org/10.48670/moi-00016.

Aveiro/OPENHIDRA - Salinity (PSU - Practical Salinity Unit)
Mar 17, 2026
Copernicus (EU), 2026, "Aveiro/OPENHIDRA - Salinity (PSU - Practical Salinity Unit)", https://doi.org/10.70122/FK2/DV7Q4B, Demo Dataverse, V1
Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological. Original Identifier: yeshttps://doi.org/10.48670/moi-00016.

Aveiro/OPENHIDRA - Sea level (m)
Mar 17, 2026
Copernicus (EU), 2026, "Aveiro/OPENHIDRA - Sea level (m)", https://doi.org/10.70122/FK2/RE2TCX, Demo Dataverse, V1
Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological. Original Identifier: yeshttps://doi.org/10.48670/moi-00016.

CDIF Dataset: Aveiro/OPENHIDRA - air pressure (Pa)
Mar 13, 2026
Copernicus (EU), 2026, "CDIF Dataset: DRA - air

Feedback

How Pale Fire agent use AI to save data in Dataverse repository

Dataverse

Add Data | Search | About | User Guide | Support | Slava Tykhonov 250

Demo Dataverse > Slava Tykhonov Dataverse >

Aveiro/OPENHIDRA - Temperature (degrees C)

Version 1.0

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

Cite Dataset | Learn about Data Citation Standards.

Edit Dataset | Curation Status | Link Dataset | Contact Owner | Share

Dataset Metrics

0 Downloads

Description	Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological. Original Identifier: yeshttps://doi.org/10.48670/moi-00016.
Subject	Earth and Environmental Sciences
License/Data Use Agreement	PUBLIC DOMAIN CC0 1.0

Files | Metadata | Terms | Versions

Add + Edit Metadata | Export Metadata

Citation Metadata

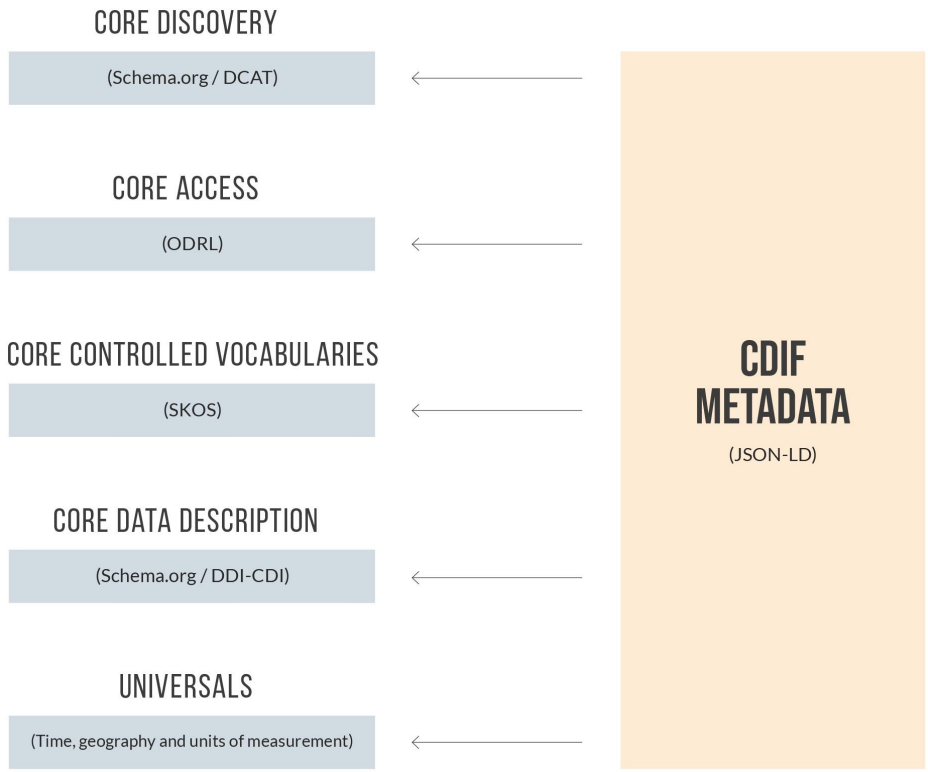
Persistent Identifier	doi:10.70122/FK2/FUU9N9
Publication Date	2026-03-17
Title	Aveiro/OPENHIDRA - Temperature (degrees C)
Author	Copernicus (EU)
Point of Contact	Use email button above to contact. Slava Tykhonov
Description	Dataset Name: Data Marine Copernicus. Factor Subtype: Hydrological. Original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

AI Agent job

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

What is Cross Domain Interoperability Framework (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



Pale Fire Footnotes turning Web pages in data

Pale Fire Footnotes

Pale Fire Footnotes << 142 143 144 145 146 147 148 149 150 Analyze page ● Connected

```
cat << 'EOF' > cache/cdif_variables_paris_air_quality_20260310.csv
Name,Value,Unit,Context
US AQI+,33,index,Air Quality Index (US standard) at 16:00 Mar 10
PM2.5,6,µg/m³,Fine particulate matter concentration
PM10,12.5,µg/m³,Coarse particulate matter concentration
O3,77.6,µg/m³,Ozone concentration
NO2,10.2,µg/m³,Nitrogen Dioxide concentration
Temperature,15,°C,Ambient air temperature
Wind speed,14,km/h,Current wind velocity
Humidity,56,%,Relative humidity
Pollen Index,Low,qualitative,General pollen count level
Tree Pollen,25,%,Tree pollen concentration level
Grass Pollen,0,%,Grass pollen concentration level
Weed Pollen,0,%,Weed pollen concentration level
Stations,16,count,Number of monitoring stations contributing data
Contributors,4,count,Number of data sources/organizations
Forecast AQI (17:00),35,index,Hourly air quality forecast
Forecast Temp (17:00),15,°C,Hourly temperature forecast
Forecast Wind (17:00),11,km/h,Hourly wind speed forecast
Forecast Humidity (17:00),53,%,Hourly humidity forecast
PM2.5 vs WHO Guideline,1.2,ratio,Current PM2.5 concentration relative to WHO an
Historical PM2.5 Average (2019),14.7,µg/m³,Annual average fine particulate matt
EOF
/usr/bin/python3 ../palefire/backend/export_cdif.py cache/cdif_variables_paris_
```

The provided text details the air quality status for **Paris, France**, as of March 10, 2026, at 16:00. The city currently enjoys a "Good" air quality rating with a US AQI of 33, driven primarily by PM2.5 levels at 6 µg/m³. The report includes comprehensive meteorological data (15°C, 14 km/h wind, 56% humidity), pollen counts, and a station-by-station breakdown of pollution levels across the city. It also provides a historical context of Paris's pollution sources (transportation, industry, heating) and its relative ranking within France.

Ask Pale Fire Footnotes...

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

Choose another country or region to see content specific to your location and shop online. France Continue

IQAir

World > ••• > Paris ••• + Add

Air quality in Paris

Air quality index (AQI*) and PM2.5 air pollution in Paris • 13.7M Followers • 16:00, Mar 10

33 US AQI* Good

Main pollutant: PM2.5 6.1 µg/m³

15° 14 km/h 56 %

Data attribution

16 stations from 4 contributors

Hourly forecast

Paris air quality index (AQI*) forecast

Now	17:00	18:00	19:00	20:00	21:00	22:00	23:00	Wed	00:00	01:00	02:00	03:00
33	34	36	39	41	43	46	46		46	47	43	39
15°	15°	15°	14°	13°	12°	12°	11°		11°	11°	11°	10°

Pale Fire data scraping job: metadata

Pale Fire Footnotes

Pale Fire Footnotes << 144 145 146 147 148 149 150 151 152 Analyze page Connected

Switching focus to describe the page https://www2.mmm.ucar.edu/wrf/users/download/get_sources_wps_geog.html

The provided text describes the **WPS V4 Geographical Static Data Downloads** page, a repository for static geographical datasets required by the WRF Preprocessing System (WPS) `geogrid` component. The data is organized into three distinct categories: **Mandatory Static Data** (essential for standard WRF runs), **Static Data for Specific Applications** (required for specialized physics schemes like NoahMP or Pleim-Xiu), and **Optional Static Data** (older versions or alternatives). The page offers various resolutions for these fields, ranging from 10 arc-minutes to as fine as 30 arc-seconds or 9 arc-seconds for certain U.S.-specific datasets.

Croissant ML Dataset Summary
The metadata for this geographical data collection has been structured into a Croissant JSON-LD package:

Name: WPS V4 Geographical Static Data
Description: A comprehensive collection of static geographical input data for WPS, including albedo, land use, soil types, and topography at multiple resolutions.
Key Resource: High-resolution mandatory fields package (~29GB uncompressed).
Source URL: https://www2.mmm.ucar.edu/wrf/users/download/get_sources_wps_geog.html
The JSON-LD packages and CSV inventory have been successfully cached in `cache/croissant_wps_geog_20260317.jsonld` and `cache/cdif_variables_wps_geog_20260317.csv`.

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

Ask Pale Fire Footnotes...

WRF

Historical Data from NCAR

Real-time from NCEP

WPS V4 Geographical Static Data Downloads Page

Below you will find static geographical data for WPS (geogrid) input. Three tables are provi types of downloads:

Mandatory Static Data
These are all the mandatory fields required to run WPS and WRF. There are two packaged tar files ava the top of this table:

1) contains the highest resolution of each mandatory field - this is the one that is most common recommended for most applications

2) contains only the lowest resolution of each mandatory field (useful for model testing and edu purposes)

An 'x' beside the field indicates that it is included in the above tar file (within the column).

Static Data for Specific Applications
These data are mandatory for specific applications.
**NOTE: if you wish to use these data, you must use them in addition to one of the mandatory data p above.

Optional Static Data
These data fields are older, and have been replaced/updated by newer fields in the "mandatory" tar fi

- **To download an individual file, click on any data file (link) listed in the left-hand column of ea**
- **To access the old (Version 3) downloads page, click here.**

Pale Fire data scraping job: variables / 1

Pale Fire Footnotes

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

show all CDIF variables

Name	Value	Unit	Other Information / Context
albedo_modis	1	%	MODIS albedo mandatory field
greenfrac_fpar_modis	1	%	MODIS green fraction / FPAR mandatory field
greenfrac_fpar_modis_5m	5	arc-min	MODIS green fraction / FPAR at 5m resolution
lai_modis_10m	10	arc-min	MODIS Leaf Area Index at 10m resolution
lai_modis_30s	30	arc-sec	MODIS Leaf Area Index at 30s resolution
maxsnowalb_modis	1	%	MODIS maximum snow albedo mandatory field
modis_landuse_20class_30s_with_lakes	30	arc-sec	MODIS 20-class land use with lakes at 30s resolution
modis_landuse_20class_5m_with_lakes	5	arc-min	MODIS 20-class land use with lakes at 5m resolution
orogwd_2deg	2	deg	Orography GWD at 2 degree resolution
orogwd_1deg	1	deg	Orography GWD at 1 degree resolution
orogwd_30m	30	arc-min	Orography GWD at 30m resolution
orogwd_20m	20	arc-min	Orography GWD at 20m resolution
orogwd_10m	10	arc-min	Orography GWD at 10m

Ask Pale Fire Footnotes...

WRF Preprocessing System (WPS) Geographical Input Data Mandatory Fields Downloads

Click on file (link) below to download individual data files	Download Highest Resolution of each Mandatory Field *Note: ~29G Uncompressed (2.6G Compressed)	Download Lowest Resolution of Each Mandatory Field
albedo_modis	x	x
greenfrac_fpar_modis	x	x
greenfrac_fpar_modis_5m		x
lai_modis_10m	x	x
lai_modis_30s	x	
maxsnowalb_modis	x	x
modis_landuse_20class_30s_with_lakes	x	
modis_landuse_20class_5m_with_lakes		x
orogwd_2deg	x	
orogwd_1deg	x	x
orogwd_30m	x	
orogwd_20m	x	
orogwd_10m	x	
soiltemp_1deg	x	x
soiltype_bot_5m		x
soiltype_bot_30s	x	
soiltype_top_5m		x
soiltype_top_30s	x	
topo_gmted2010_5m		x
topo_gmted2010_30s	x	

Pale Fire data scraping job: variables /2

Pale Fire Footnotes

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

soiltemp_1deg	1	deg	Soil temperature at 1 degree resolution
soiltype_bot_5m	5	arc-min	Bottom soil type at 5m resolution
soiltype_bot_30s	30	arc-sec	Bottom soil type at 30s resolution
soiltype_top_5m	5	arc-min	Top soil type at 5m resolution
soiltype_top_30s	30	arc-sec	Top soil type at 30s resolution
topo_gmted2010_5m	5	arc-min	GMTED2010 topography at 5m resolution
topo_gmted2010_30s	30	arc-sec	GMTED2010 topography at 30s resolution
varsso	1	m^2	Variance of subgrid-scale orography
varsso_10m	10	arc-min	Variance of subgrid-scale orography at 10m
varsso_5m	5	arc-min	Variance of subgrid-scale orography at 5m
varsso_2m	2	arc-min	Variance of subgrid-scale orography at 2m
clayfrac_5m	5	arc-min	Clay fraction for Thompson MP Scheme
erod	1	dimensionless	Erodibility factor
sandfrac_5m	5	arc-min	Sand fraction for Thompson MP Scheme
crop	1	category	Crop/Vegetation data for NoahMP
groundwater	1	m	Groundwater data for NoahMP
soilgrids	1	category	SoilGrids data for NoahMP
drainage_map	1	category	Drainage map for NoahMP
irrigation	1	category	Surface irrigation schemes

Ask Pale Fire Footnotes...

WRF Preprocessing System (WPS) Geographical Input Data Mandatory Fields Downloads

Click on file (link) below to download individual data files	Download Highest Resolution of each Mandatory Field *Note: ~29G Uncompressed (2.6G Compressed)	Download Lowest Resolution of Each Mandatory Field
albedo_modis	x	x
greenfrac_fpar_modis	x	x
greenfrac_fpar_modis_5m		x
lai_modis_10m	x	x
lai_modis_30s	x	
maxsnowalb_modis	x	x
modis_landuse_20class_30s_with_lakes	x	
modis_landuse_20class_5m_with_lakes		x
orogwd_2deg	x	
orogwd_1deg	x	x
orogwd_30m	x	
orogwd_20m	x	
orogwd_10m	x	
soiltemp_1deg	x	x
soiltype_bot_5m		x
soiltype_bot_30s	x	
soiltype_top_5m		x
soiltype_top_30s	x	
topo_gmted2010_5m		x
topo_gmted2010_30s	x	

Pale Fire: AI-Ready dataset produced by AI

Pale Fire Footnotes

Pale Fire Footnotes << 146 147 148 149 150 151 152 153 154 Analyze page Connected

Deposit first 2 variables related to soil type in Dataverse

The first two soil type variables have been successfully deposited and published in Dataverse:

CDIF Variable: soiltype_bot_5m
DOI: `doi:10.70122/FK2/ZPPR7A`
Description: Bottom soil type at 5m resolution.
CDIF Variable: soiltype_bot_30s
DOI: `doi:10.70122/FK2/XC1HVN`
Description: Bottom soil type at 30s resolution.
Both datasets include the metadata and a CSV file containing the variable information, published under a CC0 1.0 license in the `slava` collection.

Ask Pale Fire Footnotes...

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

Dataverse

Dataset Metrics
0 Downloads

Description
Bottom soil type at 30s resolution. Value: 30, Unit: arc-sec. Part of the WPS Geographical Static Data.

Subject
Earth and Environmental Sciences

License/Data Use Agreement
PUBLIC DOMAIN CC0 1.0

Files | Metadata | Terms | Versions

Add + Edit Metadata | Export Metadata

Citation Metadata

Persistent Identifier	doi:10.70122/FK2/XC1HVN
Publication Date	2026-03-17
Title	CDIF Variable: soiltype_bot_30s
Author	CDIF Expert
Point of Contact	Use email button above to contact. Slava Tykhonov
Description	Bottom soil type at 30s resolution. Value: 30, Unit: arc-sec. Part of the WPS Geographical Static Data.
Subject	Earth and Environmental Sciences

Feedback

Pale Fire: Analyzing video

Pale Fire Footnotes

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

Analyzing video https://www.youtube.com/watch?v=YTLnnoZPALI

In this video, Rick Beato draws a provocative parallel between the current AI infrastructure bubble and the historical collapse of the traditional music industry. He argues that the centralized "Big AI" model, characterized by massive data centers and subscription fees, is vulnerable to the same decentralized shift that replaced high-end recording studios with home-based Digital Audio Workstations (DAWs). Beato demonstrates the power of **Local LLMs** (specifically demonstrating **Qwen 2.5** running via **LM Studio**) as a way for users to regain privacy, control, and autonomy, potentially rendering multi-billion dollar data centers as "ghost" facilities in the future.

Croissant ML Dataset Summary
The metadata for this analysis has been structured into a Croissant JSON-LD package:

Name: How AI Will Fail Like The Music Industry - Analysis
Description: An analytical dataset based on Rick Beato's critique of centralized AI and the shift toward decentralized local computing.
Key entities: Rick Beato (Speaker), Qwen 2.5 (Model), LM Studio (Software).
Source URL: https://www.youtube.com/watch?v=YTLnnoZPALI
The JSON-LD packages and CSV inventory have been successfully cached in `cache/croissant_ai_bunker.jsonld` and `cache/cdif_variables_ai_bunker.csv`.

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

Pale Fire: chatting with video transcript

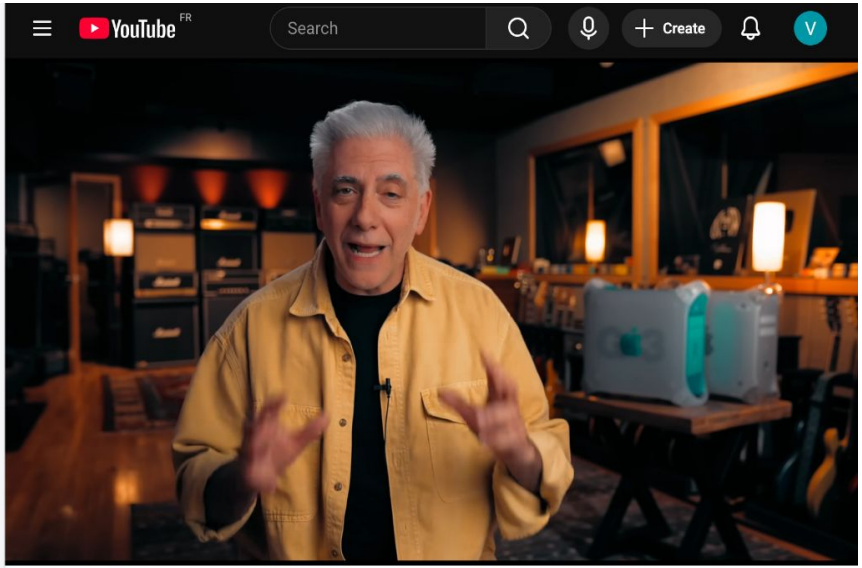
Pale Fire Footnotes

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

Why Rick Beato thinks that AI will fail?

Rick Beato argues that AI—specifically the centralized "Big AI" business model—will fail by following the same path as the traditional music industry. His core thesis is built on four main points:

The "Studio" Parallel (Centralization vs. Decentralization): Beato compares massive AI data centers to the legendary, high-cost recording studios of the 20th century. Just as those studios were rendered obsolete by the rise of home recording software (DAWs), he believes cloud-based AI subscriptions will be replaced by **Local LLMs** (Large Language Models) that people run on their own computers.
Ghost Data Centers: He warns of an infrastructure bubble. He suggests that the billions currently being spent on data centers may result in "ghost" facilities, similar to the over-investment in fiber optics during the dot-com crash. As models become more efficient and hardware becomes more powerful, the need for centralized supercomputers will diminish.
Privacy and Corporate Overreach: Beato emphasizes that users will eventually reject subscription-based models to escape corporate oversight, data collection, and "censorship." Running a model locally (using tools like **LM Studio**) provides total privacy and permanent access without a monthly bill.
Technological Democratization: He demonstrates how rapidly powerful models (like **Qwen 2.5**) are becoming accessible to average users. He argues that once "the cat is out of the bag" and high-quality models can run on a standard Mac or PC, the leverage held by companies like OpenAI or Google will evaporate.
In summary, he believes AI won't "fail" as a technology, but rather the **centralized business model** will fail because the power is rapidly shifting from corporate data centers to the individual's desk.



How AI Will Fail Like The Music Industry

Rick Beato
5.53M subscribers
Subscribe
44K Share Save Clip

943K views 5 days ago
In this episode I compare the future of AI to the failure of the music industry in the early 2000's.

Open Source AI Models: https://huggingface.co/ ...more

How it works: NER Enrichment System

- Hybrid Extraction Engine: Combines spaCy's high-precision (95%+) models with robust regex fallbacks.
- Extensive Entity Support: Recognizes 18+ types including Persons, Locations, Organizations, Dates, Laws, and Events. It’s being integrated with Croissant metadata.
- Structural Annotation: Injects metadata hints into the text to guide LLMs in accurate graph construction.
- Contextual Disambiguation: Intelligently distinguishes between entities like 'Apple' (company) and 'apple' (fruit).

How it works: Intent Intelligence

- Query Type Detection: Automatically identifies question intent (WHO, WHERE, WHEN, WHY, HOW MANY).
- Dynamic Weighting: Boosts relevant entity types based on intent (e.g., doubling Person weights for 'WHO' queries).
- Strategic Search: Tailors the ranking algorithm to prioritize the most semantically relevant data points (Qdrant).
- Improved Accuracy: Reduces noise by aligning search results with the specific nature of the user's inquiry.

Pale Fire AI Agent Daemon

- Zero Cold-Start: Keeps heavy spaCy and Gensim models pre-loaded in memory for instant NLP processing.
- Shared Memory Architecture: Enables concurrent API calls to access the same model instances, reducing RAM overhead.
- High Performance: Optimized to handle over 100 concurrent requests per second on standard hardware.
- Robust Stability: Includes automatic PID tracking, health monitoring, and self-recovery mechanisms.

Ghostwriter in Pale Fire: Intelligent RAG (Qdrant)

- Infinite Context: Implements semantic chunking to ingest and process any URL or local document regardless of length.
- High-Speed Retrieval: Fully integrated with Qdrant vector database for sub-second similarity searches.
- Private Inference: Supports a 'Local-First' stack using Ollama for secure, offline data analysis.
- Verifiable Citations: Maintains direct mapping between generated text and the original source chunks/URLs.

🤔 Ask Questions

Example: How to manage permissions in Dataverse?

Enter your question:

How to manage permissions in Dataverse?

🔍 Ask

Answer

Based on the available information: Groups can be defined as a collection of Dataverse installation user accounts, a collection of IP addresses (e. g. all users of a library's computers), or a collection of all users who log in using a particular institutional login (e. g. everyone who logs in with a particular university's account credentials). Admins of a Dataverse collection can assign roles and permissions to the users of that Dataverse collection. If you are an admin on a Dataverse collection, then you will find the link to the Permissions page under the Edit dropdown on the Dataverse collection page. Clicking on Permissions will bring you to this page: When you access a Dataverse collection's permissions page, you will see three sections: Permissions: Here you can decide the requirements that determine which types of users can add datasets and sub Dataverse collections to your Dataverse collection, and what permissions they'll be granted when they do so. Users/Groups: Here you can assign roles to specific users or groups, determining which actions they are permitted to take on your Dataverse collection. You can also reference a list of all users who have roles assigned to them for your Dataverse collection and remove their roles if you please.

Sources

Title	Relevance	Source URL
Dataverse Collection Management — Dataverse.org	73.1%	http://guides.dataverse.org/en/latest/user/dataverse-management.html

Pale Fire AI Expert Skills

DATAVERSE EXPERT

Handles complex metadata schemas (Citation, Geospatial). Automates dataset creation and API-based file uploads (Deposits).

EXCEL/GSHEET EXPERT

Loads remote Google Sheets via Pandas and generates CDIF-compliant data inventories automatically.

CDIF & CROISSANT EXPERT

Extracts measurements and geospatial variables to generate standard `Croissant` JSON-LD packages for AI training.

LIBRARIAN SKILL

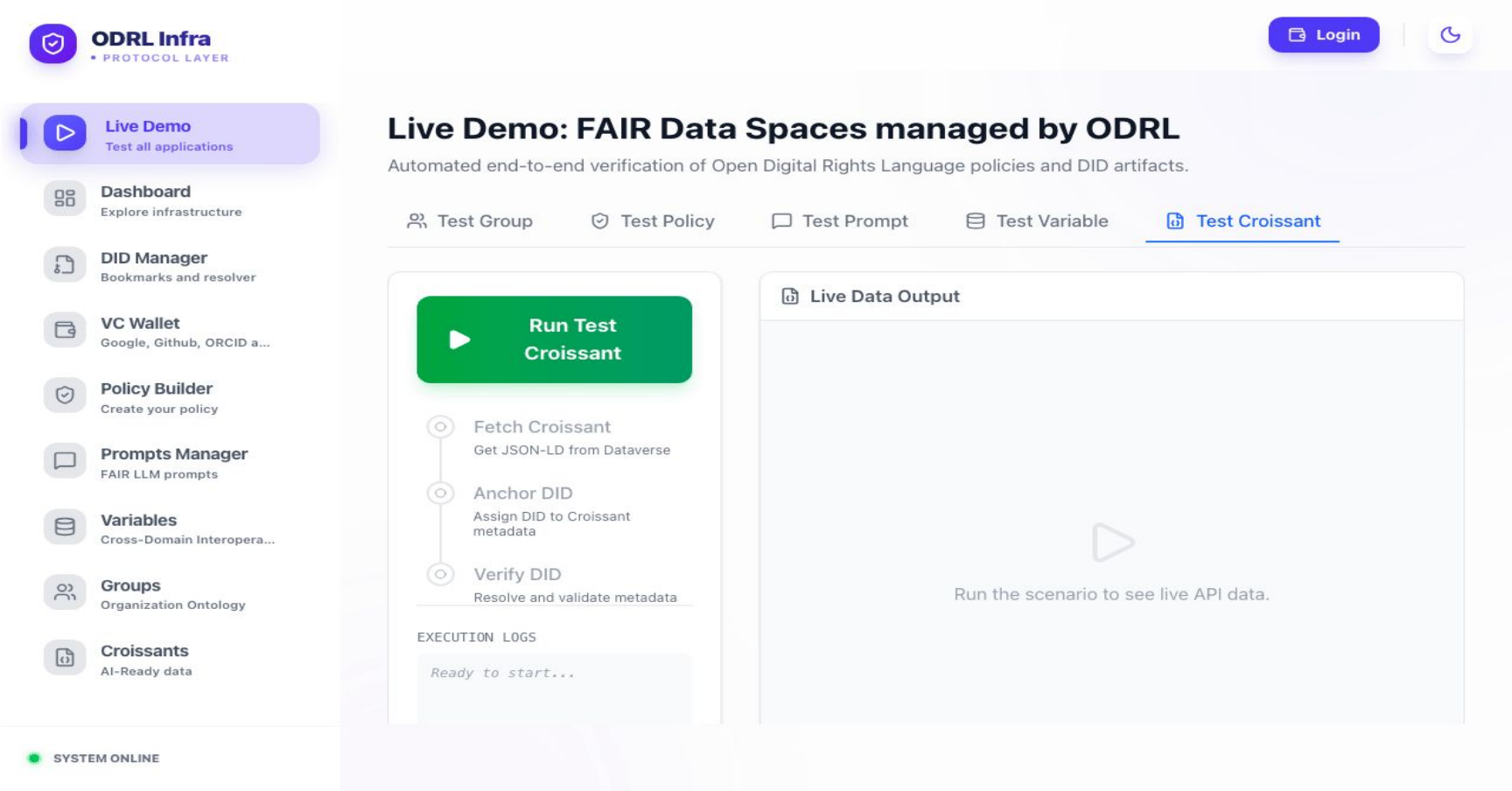
Resolves Decentralized Identifiers (DIDs) and searches ODRL (Open Digital Rights Language) services.

Expert skills are supporting Pale Fire ecosystem

Multimodal & Translation

- Advanced Transcriber: Converts video content into structured text using multimodal models like Gemini Pro.
- Expert Translation: Preserves technical terminology and nuanced meaning across multiple languages.
- Precision Extraction: Near-human accuracy in extracting entities from messy, unstructured web content.
- Rights Governance: Additional work to generate policies to ensure data usage compliance (DID and ODRL).

Open Digital Rights Languages (ODRL)



Model Context Protocol (MCP)

- Native Claude Integration: Exposes Pale Fire's knowledge graph as a tool for the Claude Desktop client.
- Tool-Enabled Search: Provides `ingest_url`, `ask_question`, and `search_content` capabilities to AI assistants.
- Fact-Based Workflows: AI clients use Pale Fire to provide deep factual backing for generated answers.
- Interoperability: Fully compatible with any modern client supporting the Model Context Protocol.

Modular Architecture

- Fully Containerized: Ready-to-use Docker-Compose setup for Neo4j, Qdrant, and the AI Agent Daemon.
- Unified Logic Path: Ensures identical processing whether accessed via CLI or the FastAPI REST interface.
- DevOps Automation: Simplified workflows with `make setup`, `make ingest-demo`, and `make logs`.
- Open Standards: Built entirely on open technologies like Neo4j/Cypher, Python, and FastAPI.

Credits

- Pale Fire: Created by Slava Tykhonov https://www.linkedin.com/in/vyacheslavtikhonov/

  CODATA Head of AI https://www.codata.org
- Mission: Bridging the gap between raw data, intelligence, and narrative synthesis.
- Future-Proof: Open-source and human-readable knowledge search designed for the AI era.
- Source Code: Available at https://github.com/agstack/palefire
- Pale Fire Footnotes: https://github.com/CA4EOSC/footnotes funded by ClimateAdapt4EOSC project (Horizon 2020 Europe)
- Community: hosted by AgStack / Linux Foundation.