

Overview of FAIR data publishing and RO- Crate

MANCHESTER
1824

The University of Manchester

*Stian Soiland-Reyes
The University of Manchester; University of Amsterdam
RO-Crate community co-chair*



*Leyla Jael Castro,
ZB MED Information Centre for Life Sciences
Bioschemas*

Bioschemas: metadata for (life) sciences websites and data feeds

Community initiative built on top of schema.org

Aim

Improve data discoverability and interoperability in Life Sciences

Approach

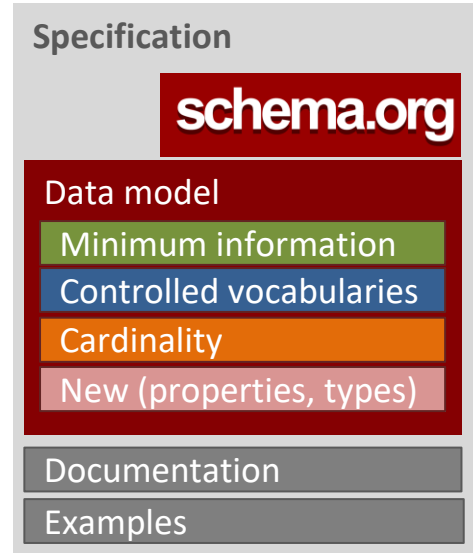
Add Life Science types to schema.org

Provide usage guidelines and examples

Minimum, recommended, optional

Link to domain ontologies

Support software



How does it work?

```
{
  "content_type": "application/json",
  "@context": "https://schema.org",
  "http://purl.org/dc/terms/conformsTo": "https://bioschemas.org/profiles/Dataset/0.3-RELEASE-2019_06_14",
  "@type": "Dataset",
  "id": "https://d-nb.info/gnd/121881389X",
  "identifier": "https://d-nb.info/gnd/121881389X",
  "name": "DaMaLOS 2020",
  "description": "First workshop on Research data management for linked open science - DaMaLOS * and other research objects",
  "keywords": "Research objects, Open Science, Data Management, Linked Data",
  "url": "https://d-nb.info/gnd/121881389X", "https://repository.publiso.de/resourcequery[]?terms=%22https://d-nb.info/gnd/121881389X%22",
  "subject": {
    "@type": "Event",
    "id": "https://zbmed.github.io/damalogs/docs/2020",
    "url": "https://zbmed.github.io/damalogs/docs/2020",
    "location": {
      "@type": "VirtualLocation",
      "url": "https://zbmed.github.io/damalogs/docs/2020"
    },
    "name": "DaMaLOS 2020",
    "description": "First workshop on Research data management for linked open science - DaMaLOS * and other research objects",
    "startDate": "2020-11-02",
    "endDate": "2020-11-02",
    "image": "https://zbmed.github.io/damalogs/img/damalogs.jpg",
    "eventAttendanceMode": "https://schema.org/OnlineEventAttendanceMode"
  }
}
```



- DaMaLOS 2023
- Call for papers
- Important dates
- Submission process
- Camera-ready submission
- Workshop schedule
- ESWC
- Program committee
- Organizing committee
- JBMS
- Contact
- DaMaLOS 2020 proceedings
- DaMaLOS 2021 proceedings

DaMaLOS 2020 - Workshop proceedings

DaMaLOS 2020 took place on the 2nd of November 2020 13:30 to 17:30 CET @ ISWC

Proceedings available at PUBLISSO Fachrepositorium DaMaLOS 2020 collection

- [DOI:10.4126/FRL01-006423280](https://doi.org/10.4126/FRL01-006423280) **Editorial note**
- **Keynote** by Prof. Dr. Carole Goble (The University of Manchester)
- **Data management plans - Chair Sören Auer (TIB)**
 - [DOI:10.4126/FRL01-006423282](https://doi.org/10.4126/FRL01-006423282) ASIO: a Research Management System based on Semantic technologies, presented by Jose Emilio Labra Gayo
 - [DOI:10.4126/FRL01-006423283](https://doi.org/10.4126/FRL01-006423283) Data Management Plans and Linked Open Data: exploiting machine actionable data management plans through Open Science Graphs, presented by Elli Papadopoulou
 - [DOI:10.4126/FRL01-006423289](https://doi.org/10.4126/FRL01-006423289) Towards semantic representation of machine-actionable Data Management Plans, presented by João Cardoso
 - [DOI:10.4126/FRL01-006423291](https://doi.org/10.4126/FRL01-006423291) Research Object Crates and Machine-actionable Data Management Plans, presented by Maroua Jaoua and Tomasz Miksa
- **Data, Life Sciences and beyond - Chair Dietrich Rebholz-Schuhman (ZB MED)**
 - [What is ORKG? - Invited talk](https://www.yaserjaredeh.com/), presented by Yaser Jaradeh (TIB)
 - [Management to Access Patient Data](https://www.patientsdata.com/), presented by
 - [Knowledge acquisition for research data](https://www.knowledgeacquisition.com/), presented by Max Schröder
 - [Semantic Search on Scientific Repositories](https://www.semanticsearch.com/), presented by Thiago Gottardi

The screenshot shows a search engine interface with the search term 'damalos'. Below the search bar, there are tabs for 'Images', 'Videos', 'Maps', 'News', '2023', 'Wikipedia', 'MP3 download', and 'DJ'. The search results section shows 'About 189,000 results (0.32 seconds)'. The first result is from GitHub Pages, titled 'https://zbmed.github.io/damalogs/'. Below it is a link to 'DaMaLOS 2023 @ ESWC' with a description: '3rd Workshop on Metadata and Research objects (e.g. Data Management for Linked Open Science - DaMaLOS 2023 (*) research objects (i.e. data, software, workflows, ...'. The bottom of the page shows a snippet of HTML code: 'html body div#top.main div#main-content-wrap-main-content-wrap div#main-content-main-content script (text)'. There are also logos for GitHub, Google, and YouTube on the left side.



The screenshot shows the website for 'DaMaLOS 2023 @ ESWC'. At the top, there is a navigation bar with the text 'DaMaLOS 2023 @ ESWC'. Below the navigation bar, there is a large network diagram consisting of several nodes (circles) connected by lines. The nodes are colored in shades of blue and red. The diagram represents a network of research objects and their relationships. The text below the diagram is partially visible, showing '3rd Workshop on Metadata and Research objects (e.g. Data Management for Linked Open Science - DaMaLOS 2023 (*) research objects (i.e. data, software, workflows, ...'. There is also a small icon of a person and a computer monitor in the bottom right corner.

FAIRness support in Bioschemas


Findable

- F1, F3. Promote - use of unique identifiers
- F2. Provide - rich metadata specifications
- F4. Promote - use of registries


Accessible

- A1. Use - HTTP(S)
- A2. Promote - use of registries


Interoperable

- I1. Use - JSON-LD
- I2. Use - schema.org
- I3. Provide - metadata specifications linking objects to each other


Reusable

- R1. Provide - metadata specifications (minimum, recommended, optional)
- R1.1. Promote - use of licenses
- R1.2 Promote - provenance and attribution
- R.1.3 Provide - community standards



FAIRness limitations in Bioschemas

- Focus on findability
- Focus on structured metadata via markup on webpages (but also options for data feeds/dumps)
- Lightweight rather than strong (e.g., ontology level) semantics
- Separation between metadata and data not always clear
- One profile per type (some use cases could need multiple profiles)
- Some tooling but (maybe) not enough

FAIRness evaluation and Bioschemas validation

June 2023, new release **v1.2.3**

FAIR-Checker

Improve the FAIRness of your web resources



Welcome

FAIR-Checker is a tool aimed at assessing FAIR principles and empowering data provider to enhance the quality of their digital resources.

Data providers and consumers can **check** how FAIR are web resources. Developers can explore and **inspect** metadata exposed in web resources.

Check ✓

Inspect 🔍

<https://fair-checker.france-bioinformatique.fr/>

Step 1: fetch RDF metadata from the resource URL

Examples: [Dataset Databse](#) [Workflow](#) [Publication Daticite](#) [Dataset](#) [Tool](#)

[🔗](#)

Build Knowledge Graph

Step 3: Metadata quality checks

← Controlled vocabularies [✎ Bioschemas](#)

Bioschemas is a community effort aimed at reusing and extending Schema.org for better life science digital resource findability. Several profiles are defined for each kind of Life Science resources, specifying minimal, recommended or optional information. Are minimal information missing? Should other information be provided for better findability?

Check BioSchemas

`https://workflowhub.eu/workflows/18?version=1` has type `https://schema.org/ComputationalWorkflow`
Using `https://bioschemas.org/profiles/ComputationalWorkflow/1.0-RELEASE` for validation, specified from the `dct:conformsTo` property.

Required missing properties

`https://schema.org/input` **must be** provided
`https://schema.org/output` **must be** provided

Improvements

`https://schema.org/citation` **should be** provided
`https://schema.org/contributor` **should be** provided
`https://schema.org/creativeWorkStatus` **should be** provided
`https://schema.org/documentation` **should be** provided
`https://schema.org/funding` **should be** provided
`https://schema.org/hasPart` **should be** provided

How are Bioschemas used?

Bioschemas define domain-specific profiles to add structured metadata to Life Science resources on the Web by using and expanding **schema.org**: *ChemicalSubstance*, *Gene*, *MolecularEntity*, *Protein*, *ProteinStructure*, *Sample*, *Taxon*

General-purpose profiles are being lifted for use *beyond* life sciences (<https://schemas.science/>) including *Dataset*, *Course*, *ComputationalWorkflow*, *ComputationalTool*, *TrainingMaterial*.

Bioschemas are **deployed** widely with **> 180** profile deployments overall

Bioschemas are developed and maintained by working groups in an active **community** (> 150 members)

Name	Group	Use Cases	Cross Walk	Task & Issues	Example
ChemicalSubstance (v0.4-RELEASE) 07 April 2020	Chemicals				
ComputationalTool (v1.0-RELEASE) 11 October 2021	Tools				
ComputationalWorkflow (v1.0-RELEASE) 09 March 2021	Workflow				
Course (v1.0-RELEASE) 13 September 2022	Training				
CourseInstance (v1.0-RELEASE) 13 September 2022	Training				
DataCatalog (v0.3-RELEASE-2019_07_01) 01 July 2019	Data Repositories				
Dataset (v1.0-RELEASE) 12 July 2022	Datasets				
FormalParameter (v1.0-RELEASE) 09 March 2021	Workflow				
Gene (v1.0-RELEASE) 07 April 2021	Genes				
MolecularEntity (v0.5-RELEASE) 07 April 2020	Chemicals				
Protein (v0.11-RELEASE) 07 April 2020	Proteins				
Sample (v0.2-RELEASE-2018_11_10) 10 November 2018	Samples				
Taxon	Biodiversity				

<https://bioschemas.org/profiles/>



Building FAIR data packages with RO-Crate



Is it FAIR to use all these repositories?

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Subjects ⊕

Content Types ☐

- Archived data (504)
- Audiovisual data (335)
- Configuration data (45)
- Databases (473)
- Images (1090)
- Networkbased data (111)
- Plain text (926)
- Raw data (979)
- Scientific and statistical data formats (1429)
- Software applications (368)
- Source code (126)
- Standard office documents (1262)
- Structured graphics (792)
- Structured text (735)
- other (769)

<https://www.re3data.org/>

FAIRCOOKBOOK

- F Findability**
 - EXEMPLAR RECIPES
 - Unique, persistent identifiers
 - Search engine optimization
 - LEARN MORE
- A Accessibility**
 - EXEMPLAR RECIPES
 - Transferring data with BFTP
 - Downloading data with Aspera
 - LEARN MORE
- I Interoperability**
 - EXEMPLAR RECIPES
 - Selecting terminologies and ontologies
 - Creating a metadata profile
 - LEARN MORE
- R Reusability**
 - EXEMPLAR RECIPES
 - Data licenses
 - Declaring data's permitted uses
 - LEARN MORE
- Infrastructure**- LEARN MORE

- Assessments**- LEARN MORE
- Applied Examples**- LEARN MORE
- Maturity model**- LEARN MORE

<https://faircookbook.elixir-europe.org/>

1620 Standards

Terminology Artifact	825
Model/Format	524
Reporting Guideline	229
Identifier Schema	28

[VIEW ALL](#)

1944 Databases

Repositories	1000
Knowledgebases	808
Knowledgebase/Repositories	136

[VIEW ALL](#)

FAIRsharing.org
standards, databases, policies

<https://fairsharing.org/>

Researchers are asked to make their research outputs FAIR – **where** to publish?

Thousands of public, institutional and domain-specific repositories

Help from guidance and **catalogues** (FAIRsharing, re3data, FAIR Cookbook)

..but how to gather and reference outputs **across multiple repositories?**

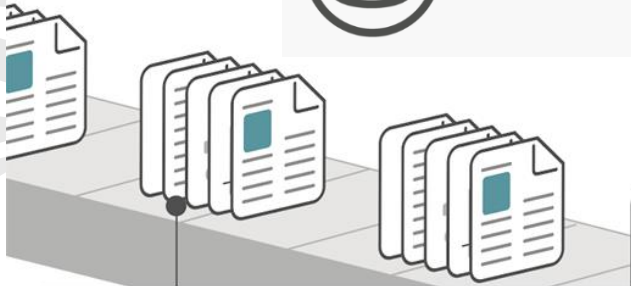
What about **contextual** information?



Gathering diverse research outputs

Building a collection of compounded objects

Enabling **reproducible**, transparent research.



scientifichypothesis



PUBLICATIONS



SLIDES



DATA



METADATA



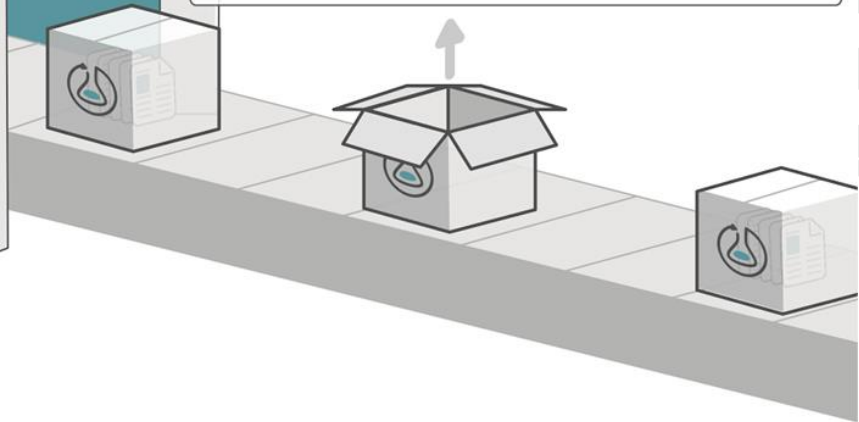
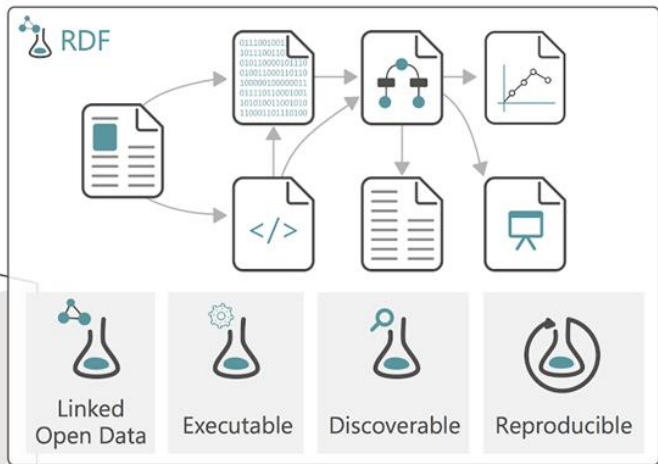
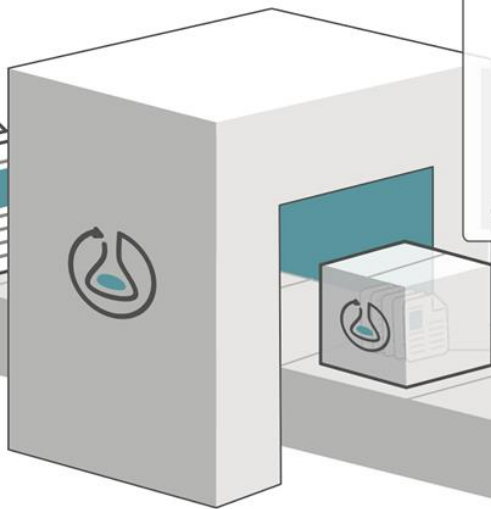
RESULTS



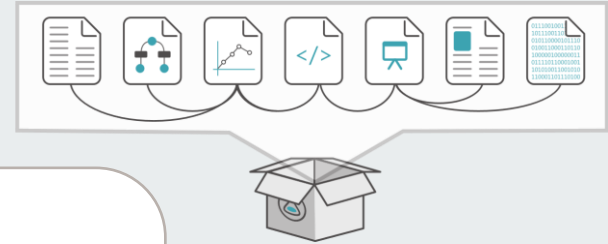
LOGS



WORKFLOWS



Aims of FAIR Research Objects



Describe and **package** data collections, datasets, software etc.
with their **metadata**

Platform-independent object exchange between repositories and services

Support **reproducibility** and **analysis**: link data with codes and workflows

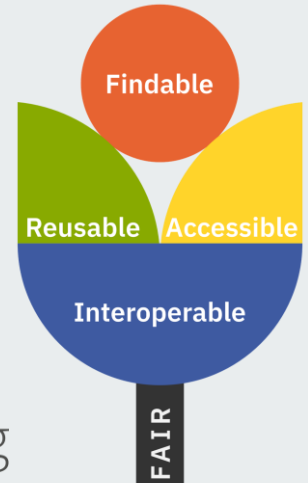
Transfer of **sensitive/large** distributed datasets with persistent identifiers

Aggregate **citations** and **persistent identifiers**

Propagate **provenance** and **existing metadata**

Publish and archive **mixed objects** and references

Reuse existing **standards**, but hide their complexity



RO-Crate: Practical and general purpose



Infrastructure independent – avoiding repository/service silos
Practical, lightweight, robust



Familiar, developer friendly, web native, machine- and human-readable, search-engine accessible
Adoptable Linked Data JSON and PIDs



Embrace diversity, legacy, unknowns, open-ended, multi-interpretation, self-describing, interlingua
Adaptable Metadata Profiles



RO-Crate

<https://www.researchobject.org/ro-crate/>

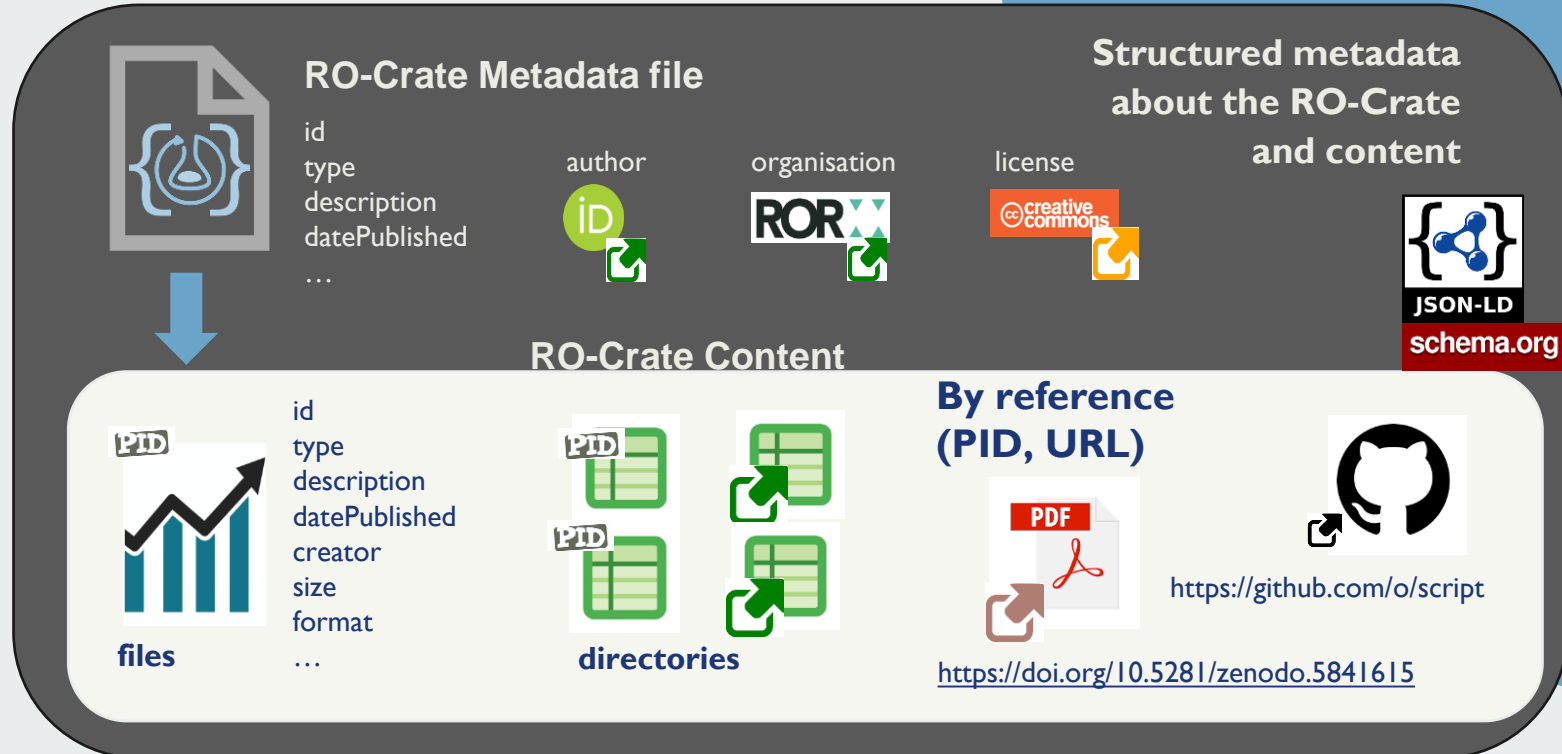
Realizing FAIR Digital Objects with RO-Crate

Re-use Web
schema.org)

existing repositories
(JSON-LD,

w/FAIR Signposting

Add : people, projects, etc.



Dataset: Survey of Victoria Arch, Wombeyan Caves NSW

[Download this Dataset](#)

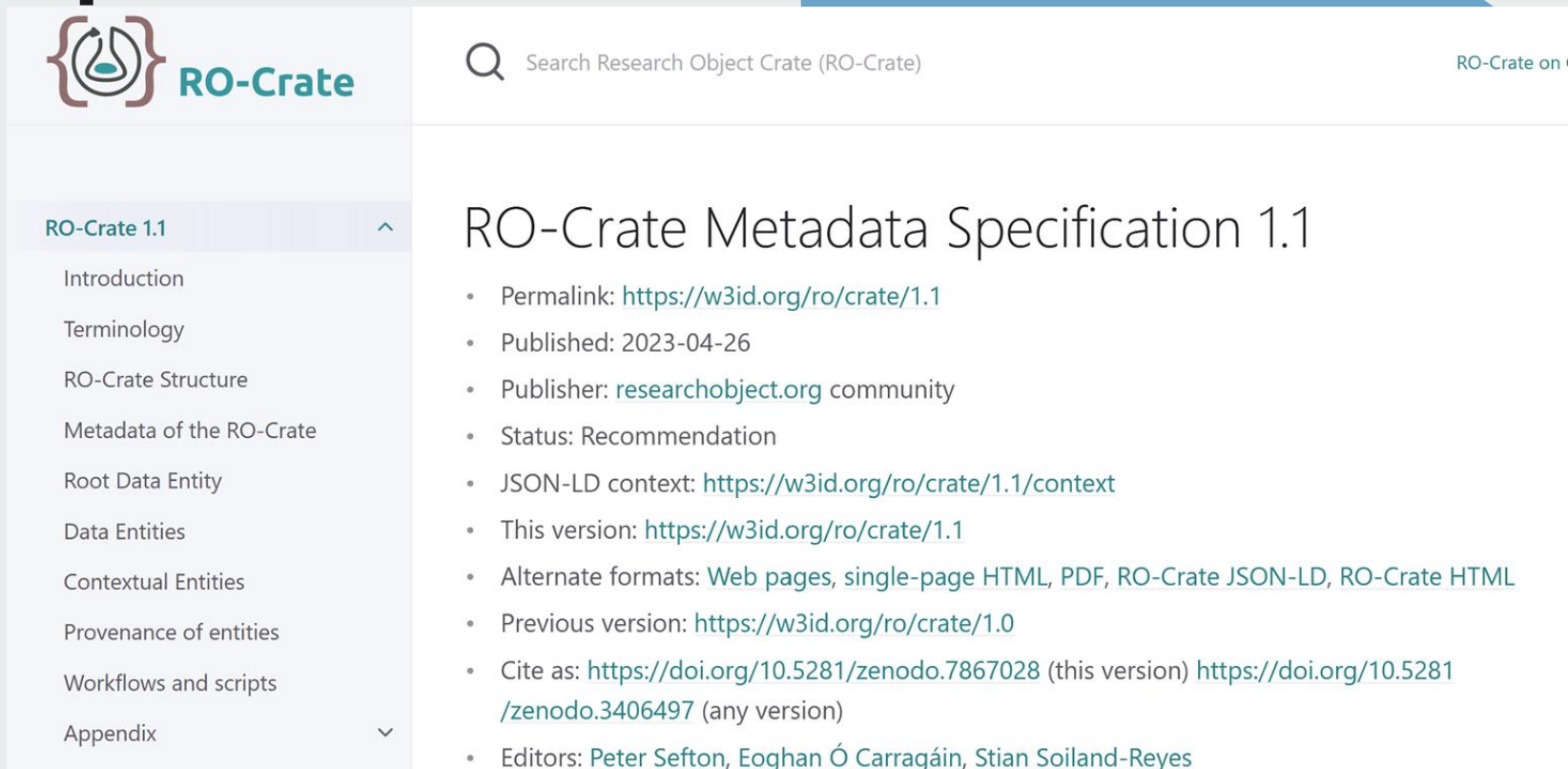
Download all the metadata for Survey of Victoria Arch, Wombeyan Caves NSW in JSON-LD format

[Check this crate](#)

FAIR is not just machine-readable!

@id	https://dx.doi.org/10.4225/59/5a4d9b76d79f4
name	Survey of Victoria Arch, Wombeyan Caves NSW
@type	Dataset
description	This data is part of a project by Michael Lake and supported by the Australian Speleological Federation. Data was acquired at Wombeyan Caves by Robert Zlot in January 2014 using the Zebedee 3D Mapping System developed by CSIRO.
datePublished	2017-10-01
creator	<ul style="list-style-type: none">• Robert Zlot• Mike Lake• Lukas Kaul
path	./
contactPoint	Contact Mike Lake

Guidance by examples



The screenshot shows the RO-Crate website. The top left features the RO-Crate logo, which consists of a stylized blue and green icon resembling a water drop or a flame inside a brown bracket, followed by the text "RO-Crate". To the right of the logo is a search bar with a magnifying glass icon and the text "Search Research Object Crate (RO-Crate)". The main content area is divided into two columns. The left column is a navigation menu with the following items: "RO-Crate 1.1" (highlighted with a blue background and an upward arrow), "Introduction", "Terminology", "RO-Crate Structure", "Metadata of the RO-Crate", "Root Data Entity", "Data Entities", "Contextual Entities", "Provenance of entities", "Workflows and scripts", and "Appendix" (with a downward arrow). The right column displays the "RO-Crate Metadata Specification 1.1" page. The title "RO-Crate Metadata Specification 1.1" is prominently displayed at the top of this section. Below the title is a list of key information points:

- Permalink: <https://w3id.org/ro/crate/1.1>
- Published: 2023-04-26
- Publisher: researchobject.org community
- Status: Recommendation
- JSON-LD context: <https://w3id.org/ro/crate/1.1/context>
- This version: <https://w3id.org/ro/crate/1.1>
- Alternate formats: [Web pages](#), [single-page HTML](#), [PDF](#), [RO-Crate JSON-LD](#), [RO-Crate HTML](#)
- Previous version: <https://w3id.org/ro/crate/1.0>
- Cite as: <https://doi.org/10.5281/zenodo.7867028> (this version) <https://doi.org/10.5281/zenodo.3406497> (any version)
- Editors: Peter Sefton, Eoghan Ó Carragáin, Stian Soiland-Reyes



RO-Crate in practice

RO-Crate is used by multiple
international projects

Applied across research domains –
from **life sciences** to **cultural
heritage**



Collecting corpora for a Language Data Commons



THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

Arkisto Platform - Describo

Use another service Resource: local:/home/pt/working/language-research-technology/corpus-tools-sydney-speaks/template Profile: Language Data Commons top level Collection (corpus) (0.1)

Build the Collection Manage Collection Data Files Browse Collection Entities Manage Templates

Load Root Dataset

About Main Related items Collection Structure Space and Time

Conforms To
A link to the Text Commons RD-Crate profile for collections
URL: <https://purl.archive.org/language-data-commons/profile#Collection>

Description
An abstract of the collection, include as much detail as possible about the motivation and use of the collection, including things that we do not yet have properties for
The Sydney Speaks Collection brings together three sub-corpora of recorded spontaneous speech: Sydney Speaks 2010s, Sydney Social Dialect Survey, and NSW Bicentennial Oral History Collection. The Sydney Speaks 2010s Corpus speakers include a cross-section of Sydney's residents, consisting of recordings

Author
The person or organization responsible for creating this collection of data
+ Person + Organization
Person: Catherine Travis

Sydney Speaks

2018-05-02 - interview with Enrico Ontario - alias - (Male, 53)

LDCA Hub

Sydney Speaks 2010s

Temporal Cover: 2010-2019
age
Conforms To: <https://purl.archive.org/textcommons/profile#Collection>
Identifier: ATAP

Items in Collection: 169

2018-05-02 - interview with Enrico Ontario - alias - (Male, 53)
2018-05-16 - interview with Tony Angello - alias - (Male, 37)
2017-07-23 - interview with Angela Wang - alias - (Female, 24)
2018-11-07 - interview with Cassie Cassan - alias - (Female, 18)
2016-05-28 - interview with Tessa Haglora - alias - (Female, 62)
2018-01-21 - interview with Craig Beer - alias - (Male, 31)
2017-07-04 - interview with Julia Anderson - alias - (Female, 20)

Access
Collection Level Metadata License (can display metadata)
Read more
Sydney Speaks License E
Public Metadata
Indexed

Member Of
Sydney Speaks

Content
Language: English
Linguistic Genre: 427



Peter Sefton et al: <https://ptsefton.com/2022/11/25/ldca-metadata-ecosystem-eresearch-2022/index.html>

<https://youtu.be/p-GZbe-Kzww>



ROHub: Earth observation data

The EOSC project **RELIANCE** use RO-Crate to package data cubes of **earth observation data**, along with documentation, images and workflows

Connects to related **infrastructures** for interactive execution/analysis.

Metadata includes **temporal** coverage, **spatial** coverage and **vertical** coverage.

ROHub publishes the archived RO-Crates to general-purpose repositories (Zenodo, B2Share) for longevity and PIDs.



<input checked="" type="checkbox"/>		CAMS European air quality forecasts: REC	16.03.2023 (15:48)
<input type="checkbox"/>		My datacube project1	17.03.2023 (13:14)
<input type="checkbox"/>		https://reliance.adamplatform.eu/?dataset=69628:EU_CAMS_SURFACE_REC_G&feature=61a8b7865e7d1c79f36e35da	17.03.2023 (14:39)
<input type="checkbox"/>		My DC product 3	17.03.2023 (14:41)
<input type="checkbox"/>		Screenshot 2023-03-27 at 14.46.39.png	136Kb 28.03.2023 (09:26)

Created: 16.03.2023 (15:48), last modified: 16.03.2023 (15:48)

DATA CUBE COLLECTION REMOTE

CAMS European air quality forecasts: REC

https://reliance.adamplatform.eu/?dataset=69628:EU_CAMS_SURFACE_REC_G

Identifiers:

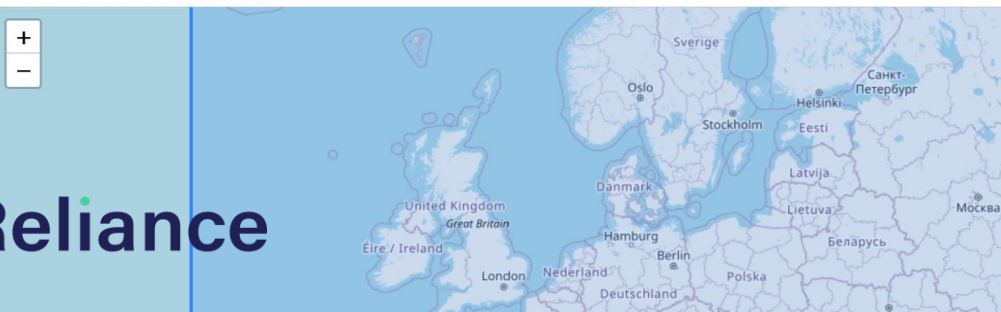
Collection ID: EU_CAMS_SURFACE_REC_G

Description:

CAMS SURFACE RESIDENTIAL ELEMENTARY CARBON



SPATIAL COVERAGE



<https://reliance.rohub.org/>

<https://www.researchobject.org/ro-crate/in-use/rohub.html>

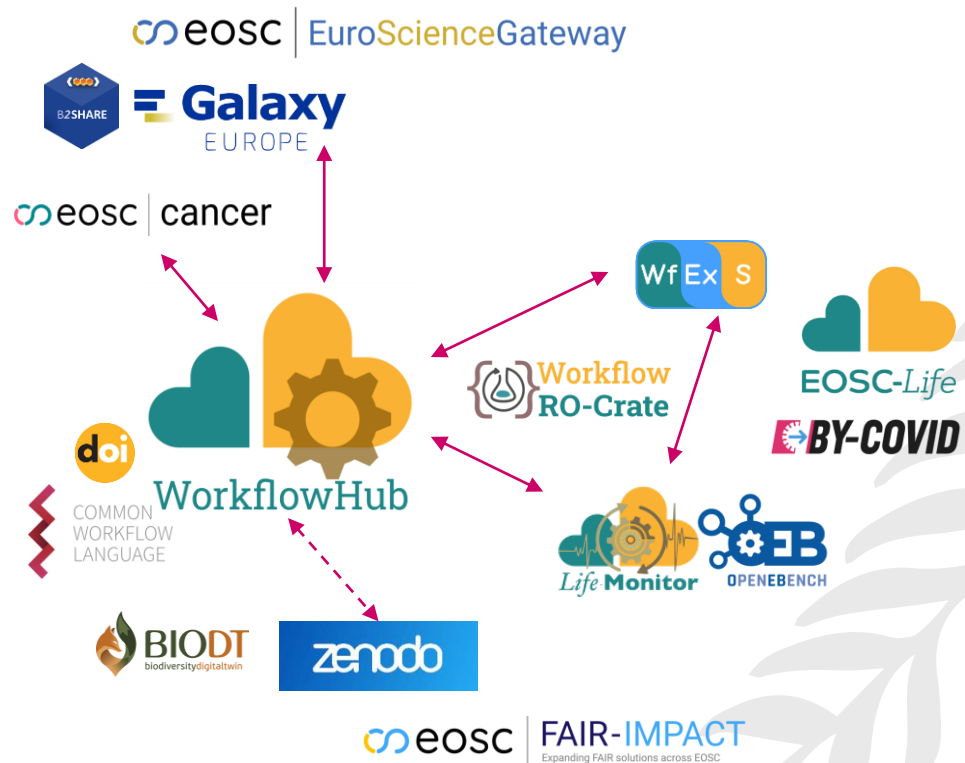
<https://w3id.org/ro-id/6fa27870-c1a4-4386-8d51-855d5ac932e2>

Building an EOSC ecosystem of FAIR Workflows

- EOSC projects **BY-COVID**, **EOSC-Life**, **EuroScienceGateway**, **BioDT** exchange rich Workflow RO-Crates within an emerging EOSC ecosystem of workflow services

Workflow Crates transfer

- identifiers, authors, license, workflow system
- executable workflows in their **native format** (e.g. Galaxy)
- interoperable **CWL** description of the workflow
- **software citations** (e.g. tools used)
- required data **sources**
- **test** suites
- workflow **execution** provenance



<https://workflowhub.eu/>
<https://w3id.org/workflowhub/workflow-ro-crate/>
<https://w3id.org/ro/wfrun/>

Importance of profiles

*“ Brian: Look, you’ve got it all wrong! You don’t need to follow me. You don’t need to follow anybody! You’ve got to think for yourselves! You’re **all individuals!**”*

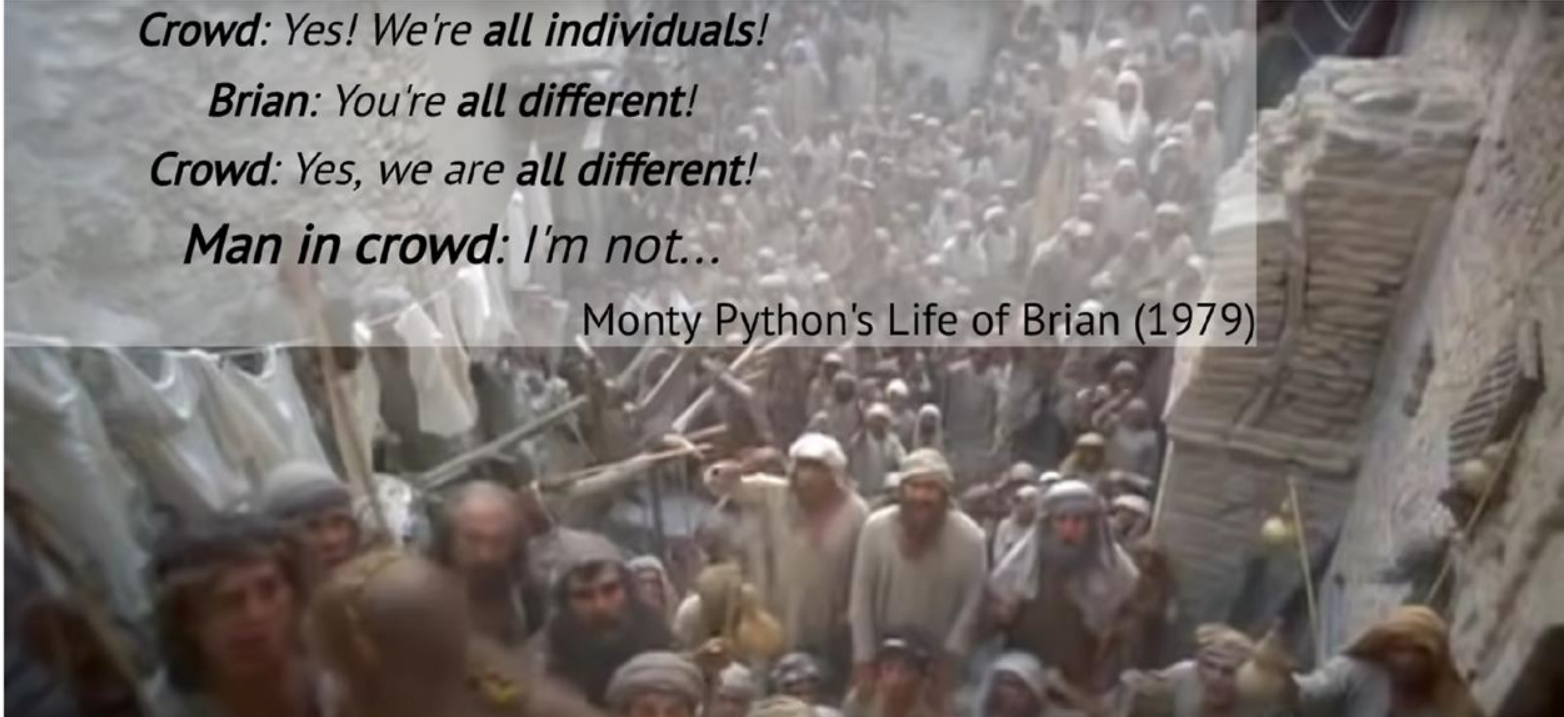
*Crowd: Yes! We’re **all individuals!***

*Brian: You’re **all different!***

*Crowd: Yes, we are **all different!***

Man in crowd: I’m not...

Monty Python's Life of Brian (1979)





OPEN MEET. SHARE. INSPIRE. CARE.
SCIENCE
FESTIVAL

#OSF2023DE

Thank you!

*Stian Soiland-Reyes
The University of Manchester
RO-Crate community co-chair*

*Leyla Jael Castro,
ZB MED Information Centre for Life Sciences
Bioschemas*