eosc | FAIR-EASE

Building Interoperable Earth Science & Environmental Services

A FAIR-EASE WEBINAR

FAIR-EASE Technical webinar on RO-Crate

29.08 2024

11:00-12:00 CEST

REGISTER NOW

eosc | FAIR-EASE
Building Interoperable Earth Science & Environmental Services

Welcome and Housekeeping

Rob Carrillo
Trust-IT Services

FAIR-EASE has received funding from the European Union's Horizon Europe Framework Programme (HORIZON) - under grant agreement No. 101058785.

Welcome & Housekeeping

- Write on the chat to say hello and where you are connecting from.
- Please mute your microphone and switch off your camera and do not speak unless provided permission by the host. You may ask this via chat.
- This is a recorded webinar. Slides and the recording will be published online on the event page within 1 week of the broadcast.
- Please keep your questions until the end in case your question is addressed in later slides.
- Do share your impressions and screenshots of the webinar. Tag us on our channels:
  - X : @FAIR_EASE
  - LinkedIn: /company/fair-ease

EOSC-Life

BY-COVID

eosc Euro Science Gateway

eosc FAIR-IMPACT
Expanding FAIR solutions across EOSC

eosc | FAIR-EASE
Building Interoperable Earth Science & Environmental Services

RO-Crate

elixir

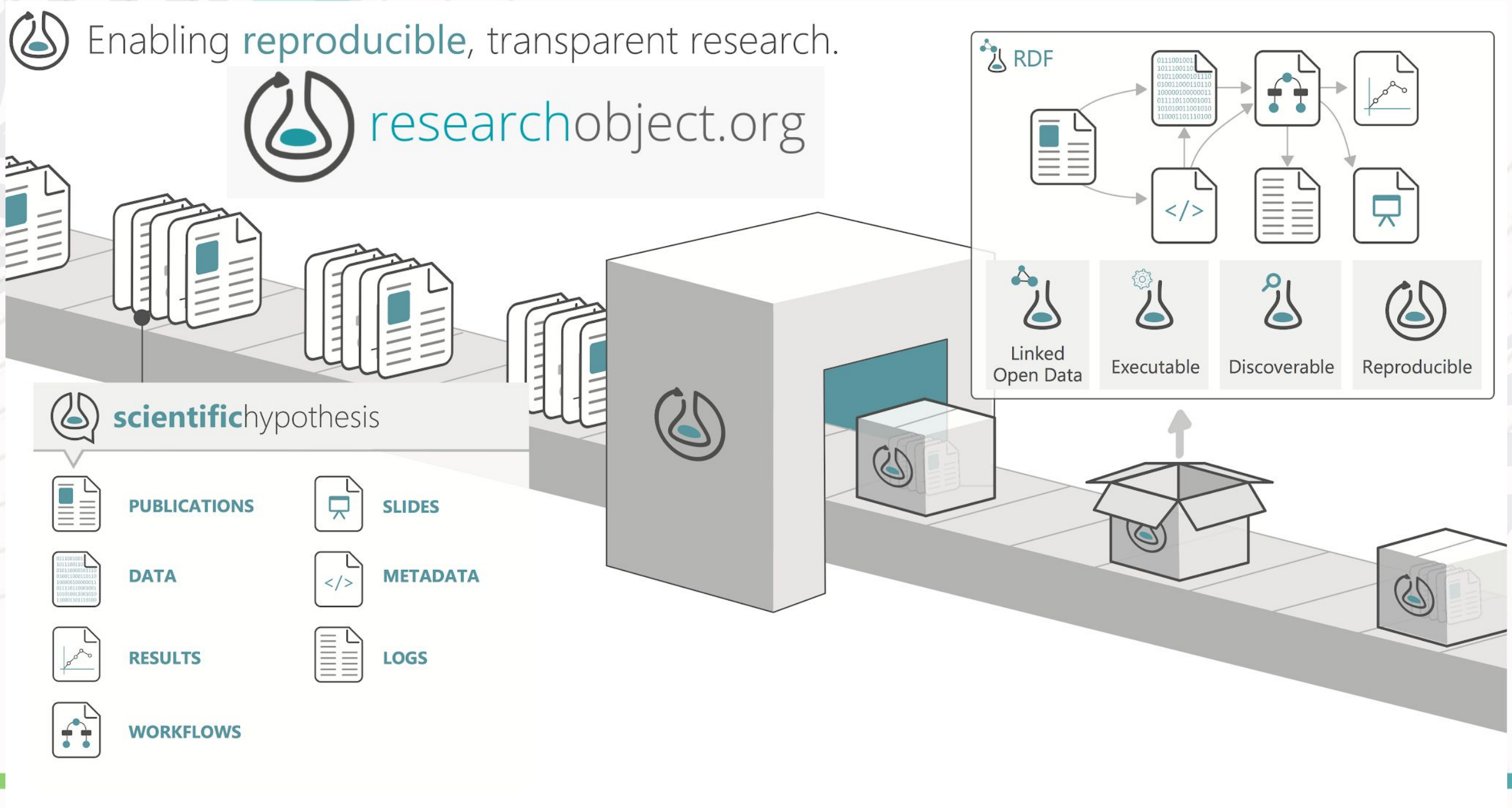
eosc | ENTRUST

eosc | cancer

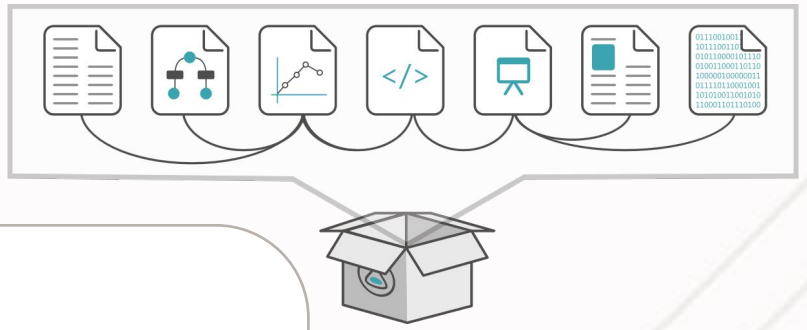
Getting Started with RO-Crate

Stian Soiland-Reyes,
The University of Manchester, UK

FAIR-EASE has received funding from the European Union's Horizon Europe Framework Programme (HORIZON) - under grant agreement No. 101058785.


Enabling reproducible, transparent research.
researchobject.org
RDF
Linked Open Data
Executable
Discoverable
Reproducible
scientifichypothesis
PUBLICATIONS
SLIDES
DATA
METADATA
RESULTS
LOGS
WORKFLOWS

Aims of FAIR Research Objects



Describe and **package** data collections, datasets, software etc.

→ data moves with its **metadata**

Platform-independent object exchange between repositories and services

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

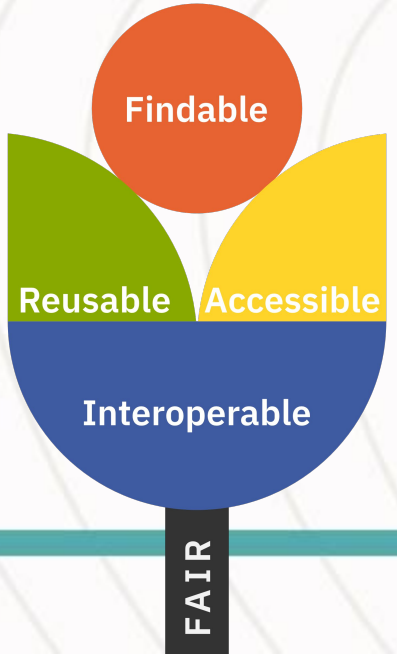
Transfer of **sensitive/large** distributed datasets with persistent identifiers

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

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

Publish and archive **mixed objects** and references

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

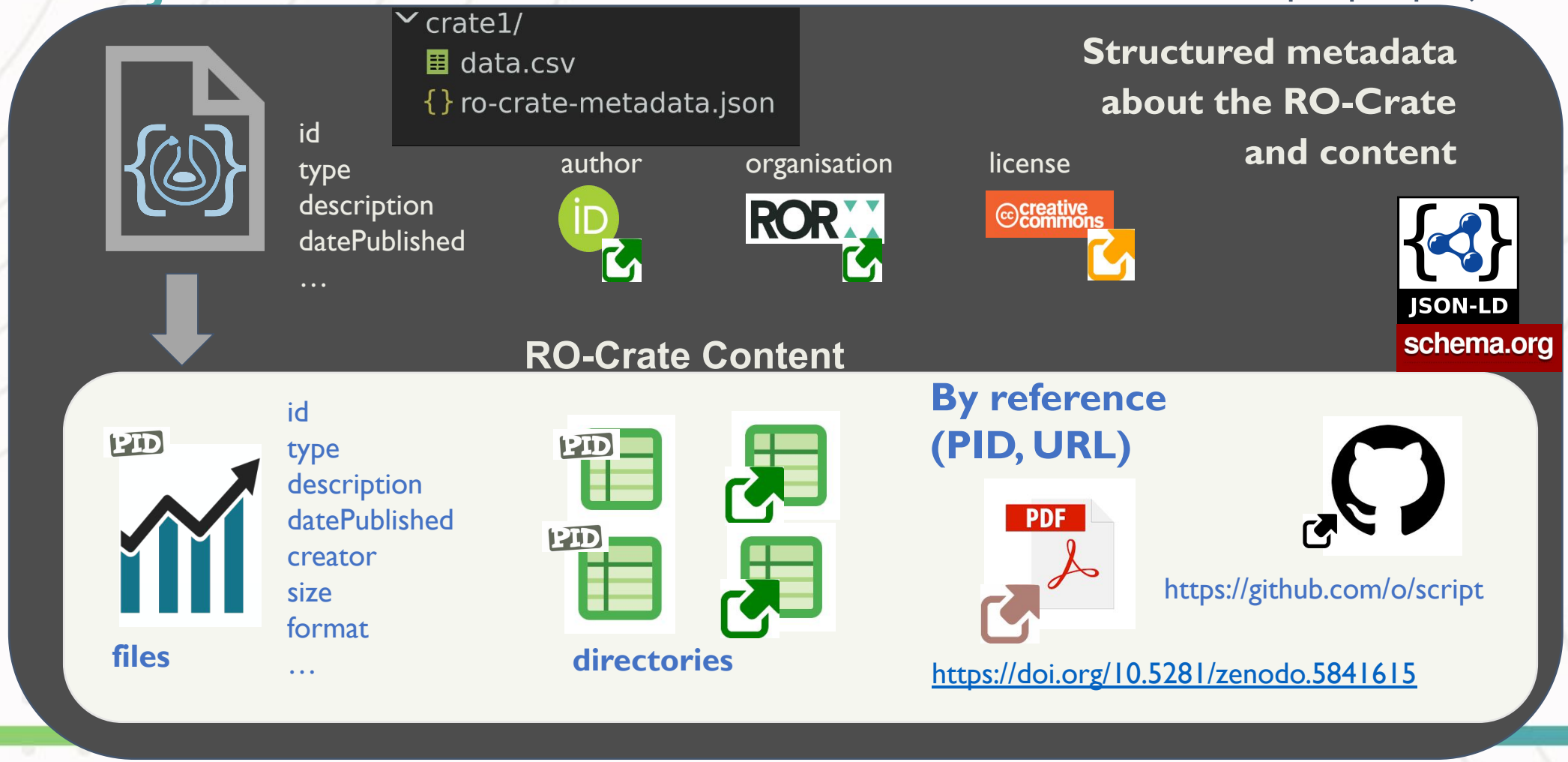


research object.org

eosc | FAIR-EASE

Realizing FAIR Digital Objects with RO-Crate

Reference existing repositories
Re-use Web **standards** (JSON-LD, schema.org)
Persistent identifiers w/FAIR Signposting
Add **context:** people, projects, etc.



https://www.researchobject.org/ro-crate/specification.html

BY-COVID WP5 T5.2 Baseline Use Case

⬇️🏷️ Download all the metadata for BY-COVID WP5 T5.2 Baseline Use Case in JSON-LD format

Check this crate

⬇️📦 Download this dataset: BY-COVID WP5 T5.2 Baseline Use Case

BY-COVID WP5 T5.2 Baseline Use Case

@id	./
name [?]	BY-COVID WP5 T5.2 Baseline Use Case
@type	Dataset
description [?]	This publication corresponds to the Research Objects (RO) of the Baseline Use Case proposed in COVID project on "COVID-19 Vaccine(s) effectiveness in preventing SARS-CoV-2 infection".
funder [?]	European Commission
datePublished [?]	2023-04-19
author [?]	• Francisco Estupiñán-Romero • Nina Van Goethem • Marjan Meurisse • Javier González-Galindo • Enrique Bernal-Delgado
conformsTo [?]	• Process Run Crate • Workflow Run Crate • Workflow RO-Crate
codeRepository [?]	https://github.com/by-covid/BY-COVID_WP5_T5.2_baseline-use-case
hasPart [?]	• BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Anal • BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Caus • COVID-19 vaccine(s) effectiveness assessment (synthetic dataset) • BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Data • BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Stud • Common data model specification • Diagram of analytical pipeline • Conceptual phases • README.md
distribution [?]	https://github.com/by-covid/BY-COVID_WP5_T5.2_baseline-use-case/archive/refs/heads/main.zip
identifier [?]	https://doi.org/10.5281/zenodo.6913045

⬇️ Download: BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Causal Model

@id	vaccine_effectiveness_causal_model/
name [?]	BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Causal Model
@type	File
description [?]	Causal model responding to the research question, using a Directed Acyclic Graph
datePublished [?]	2023-01-26
version [?]	1.1.0
identifier [?]	https://doi.org/10.5281/zenodo.6913045
url [?]	https://zenodo.org/record/7572373
creator [?]	• Francisco Estupiñán-Romero • Nina Van Goethem • Marjan Meurisse • Javier González-Galindo • Enrique Bernal-Delgado
contributor [?]	• Simon Saldner • Lorenz Dolanski-Aghamanoukjan • Alexander Degelsegger-Marquez • Martínez-Lizaga, Natalia
hasPart [?]	• COVID-19 vaccine effectiveness causal model v.1.1.0 (HTML) • COVID-19 vaccine effectiveness causal model v.1.1.0 (QMD)

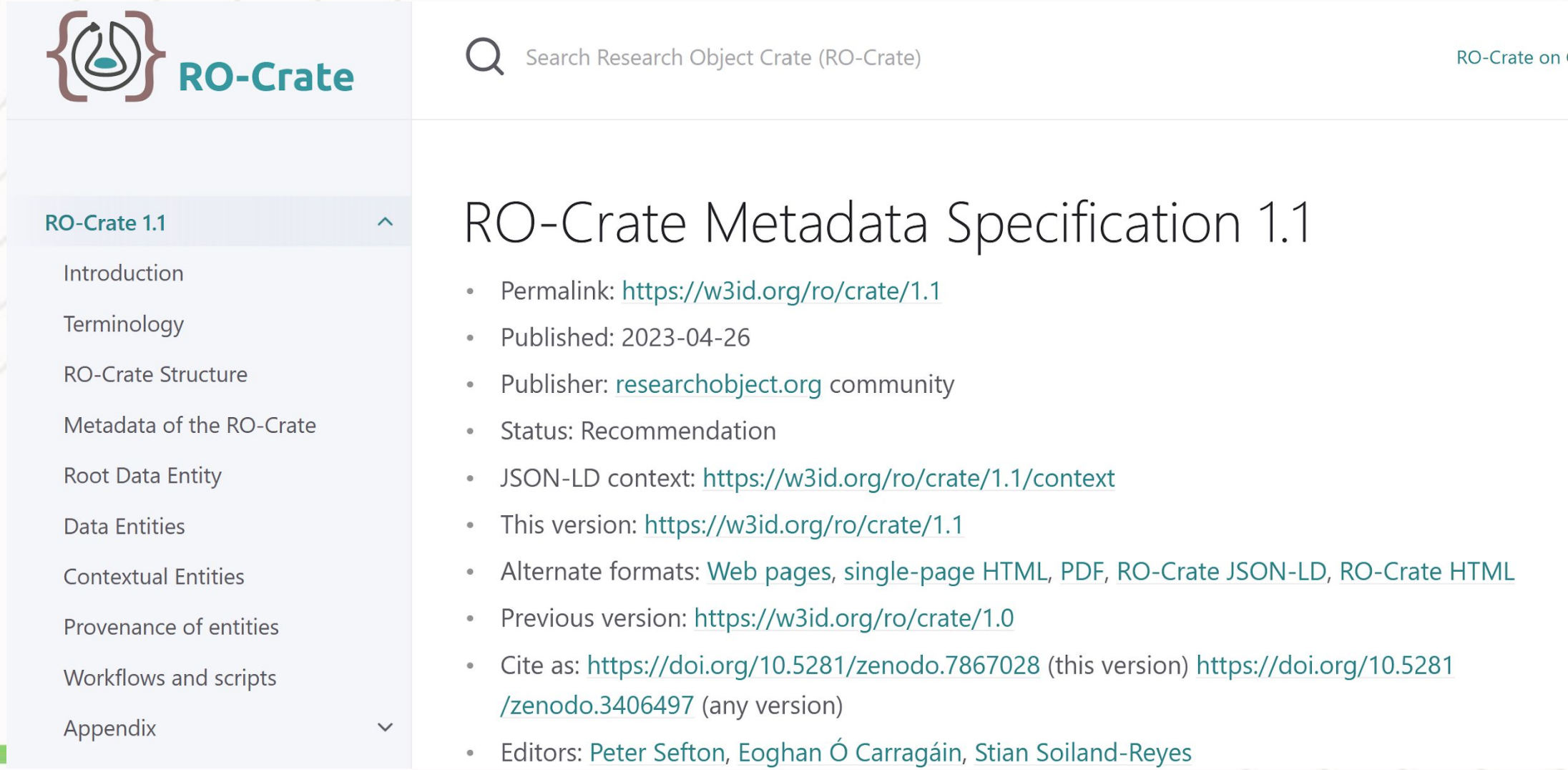
Items that reference this one

hasPart [?]	• BY-COVID WP5 T5.2 Baseline Use Case • BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Data Management Plan • BY-COVID - WP5 - Baseline Use Case: SARS-CoV-2 vaccine effectiveness assessment - Study protocol

https://www.npmjs.com/package/ro-crate-html

https://w3id.org/ro/doi/10.5281/zenodo.6913045

Guidance by examples



https://www.researchobject.org/ro-crate/specification

eosc | FAIR-EASE

RO-Crate in practice

RO-Crate is used by multiple international projects

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

Arkisto

Arkisto was a project website that aimed to outline a principled approach to research data management.

Autosubmit

Autosubmit is an open source Python experiment and workflow manager used to manage complex workflows on Cloud and HPC platforms. Autosubmit uses RO-Crate to package the configuration, traces (logs, metrics, databases, etc.), and data of experiments and workflows.

BioConnect

JAX BioConnect is an index of research data that supports data sharing, high-quality curation, and consistent data description.

COMPSs

The COMPSs programming model is able to record Workflow Provenance in RO-Crate format, for governance and reproducibility of computational experiments

Data Plant

DataPLANT's mission is to provide a sustainable and well annotated data management platform for plant sciences.

Data Stewardship Wizard

The Data Stewardship Wizard (DSW) is an interactive platform for making data management plans

Dataverse & AROMA

AROMA (ARP RO-Crate Manager) is part of Hungarian initiative ELKH ARP, extending Harvard Dataverse to allow dynamic metadata editing of data deposit metadata.

DeSci Nodes

The DeSci Nodes system has been developed by the DeSci foundation, where dPID (distibuted Persistent Identifier) act as an overlay of the Interplanetary File System (IPFS)

EGI AppDB

EGI's Application Database (AppDB) has support for RO-Crate download, either as a JSON-LD metadata file or a ZIP archive.

FAIRSCAPE

FAIRSCAPE is a framework for reusable cloud-based computations using ARK identifiers with rich provenance in an evidence graph and the Evidence Graph Ontology (EVI)

Five Safes Crate

The Five Safes RO-Crate profile extend the Workflow Run RO-Crate profile for use in Trusted Research Environments (TRE).

KEDO Data Lake

Knowledge Enhanced Digital Objects (KEDO) is an experimental approach of building a data lake using a combination of knowledge graphs, RO-Crate and PID records.

Language Data Commons of Australia

LDaCA uses RO-Crate as an interchange and archive format for language data, and is providing data discovery portals and API access to data using RO-Crate-centric APIs.

Life Monitor

Life Monitor is a testing and monitoring service for computational workflows being developed as part of the EOSC-Life project. It aims to facilitate the execution, monitoring and sharing of workflow tests over time, allowing to detect deviations from expected workflow operation and provide useful feedback to the workflow authors

LivePublication

LivePublication is a proof of concept of an executable paper, which interactive visualization and statistical calculations can be regenerated on the fly taking into consideration data sources updated after the paper's publication date.

Machine-actionable data management plans

RDA maDMP Mapper and Ro-Crate_2ma-DMP can convert between machine-actionable data management plans (maDMP) and RO-Crate.

PARADISEC

Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) holds 16,100 hours of audio recordings and 2,800 hours of video recordings that might otherwise have been lost. These recordings are of performance, narrative, singing, and other oral tradition. This amounts to over 220 terabytes, and represents 1,370 languages, mainly from the Pacific region.

PILARS

PILARS is a set of Protocols for Implementing Long-term Archival Repository Services

ROHub

ROHub is a solution for the storage, lifecycle management and preservation of scientific work and operational processes via research objects. It makes these resources available to others, allows to publish and release them through a DOI, and allows to discover and reuse pre-existing scientific knowledge.

RRkive

RRKive is a website aimed to outline a principled approach to research data management with guidance on data storage and metadata

Research Object Composer

Research Object Composer is a REST API for gradually building and depositing Research Objects according to a pre-defined profile.

Sciebo RDS

Sciebo RDS (Research Data Services) is a self-hosted interface between data repositories and file storage solutions, assisting the research data deposition process with annotations made using Describo Online and stored as an RO-Crate

Survey Ontology

The Survey Ontology is an open vocabulary that allows representing, annotating and sharing a representation of the survey structure and the gathered responses.

Time Layered Cultural Map (TLCMap)

TLCMap is a set of tools that work together for mapping Australian history and culture which includes downloads of geographical data packaged in RO-Crate

UTS Cultural Datasets

The UTS Cultural Datasets project is collaborating with Humanities and Social Science (HASS) researchers and is re-using existing UTS Data infrastructure to build interactive services that allow people to use the data.

UTS Research Data Repository

The UTS Research Data Repository is a searchable portal for discovering and accessing public datasets by UTS researchers.

https://www.researchobject.org/ro-crate/use_cases

Wildlive portal

Senckenberg's Wildlive data portal, is a repository and analysis platform for biodiversity monitoring.

Workflow Execution Service (WfExS)

WfExS-backend is a high-level workflow execution command line program that consumes and creates RO-Crates, focusing on the interconnection of content-sensitive research infrastructures for handling sensitive human data analysis scenarios

WorkflowHub

WorkflowHub imports and exports Workflow RO-Crates, using it as an exchange format. They are a specialization of RO-Crate for packaging an executable workflow with all necessary documentation. It is aligned with, and intends to strictly extend, the more general Bioschemas ComputationalWorkflow profile.

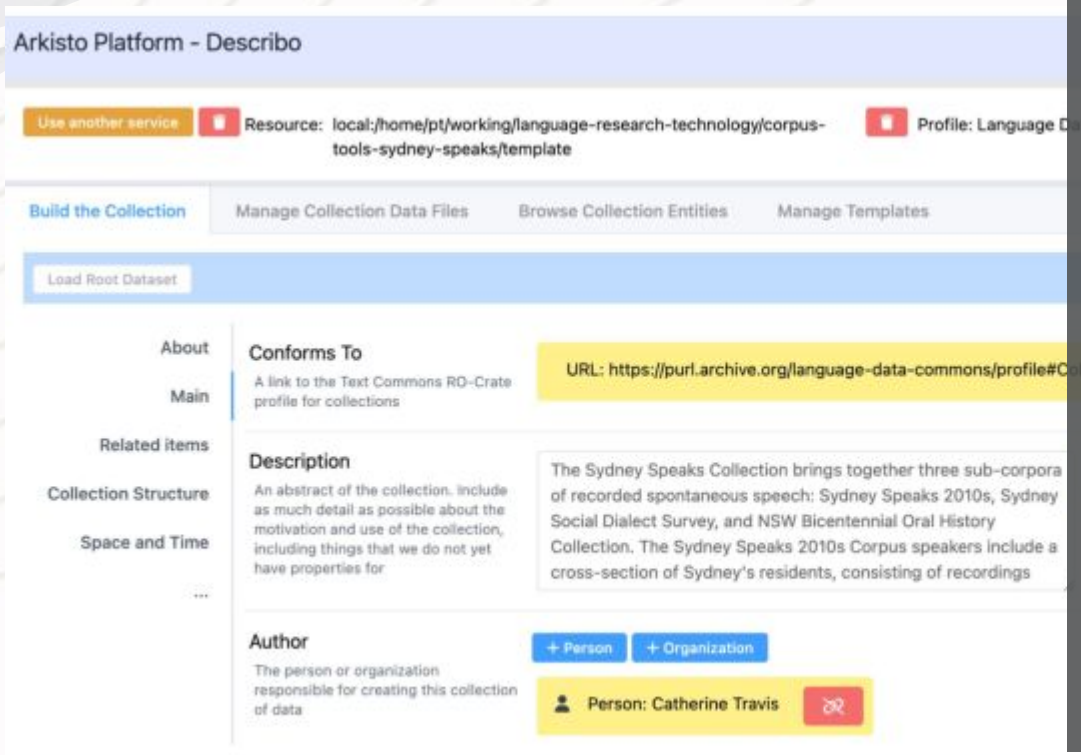
ZBMed SemTec web pages

The Semantic Technologies (SemTec) team in ZB MED uses GitHub pages to share research projects and corresponding research artefacts/outcomes.

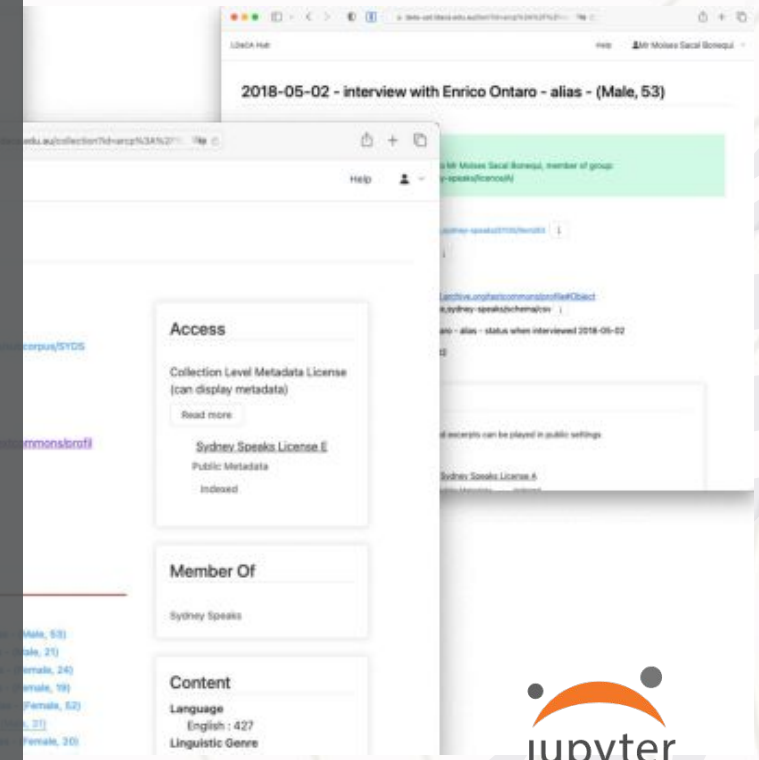
Collecting corpora for a Language Data Commons

eosc | FAIR-EASE

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA




```
{
  "@id": "#place-Kiriwina",
  "@type": "Place",
  "geo": {
    "@id": "#geo-151.0,-8.6-151.2,-8.4"
  },
  "name": "Kiriwina"
},
{
  "@id": "#language-kij",
  "@type": "Language",
  "code": "kij",
  "location": {
    "@id": "#geo-150.294,-8.879-151.2,-8.32"
  },
  "name": "Kilivila"
},
  {
  "@id": "#geo-150.294,-8.879-151.2,-8.32",
  "@type": "Geometry",
  "asWKT": "POLYGON((150.294 -8.32, 151.2
-8.32, 151.2 -8.879, 150.294 -8.879, 150.294
-8.32))"
  },
```



jupyter

https://youtu.be/p-GZbe-Kzwv

atap australian text analytics platform
atap.edu.au

https://www.researchobject.org/ro-crate/ldaca
https://www.researchobject.org/ro-crate/paradisec

eosc | FAIR-EASE

ROHub: Earth observation data

The EOSC project **RELIANCE** used RO-Crate to package data cubes of **earth observation data**, along with documentation, images and workflows
→ FAIR2ADAPT project

Connects to related infrastructures for interactive execution/analysis.

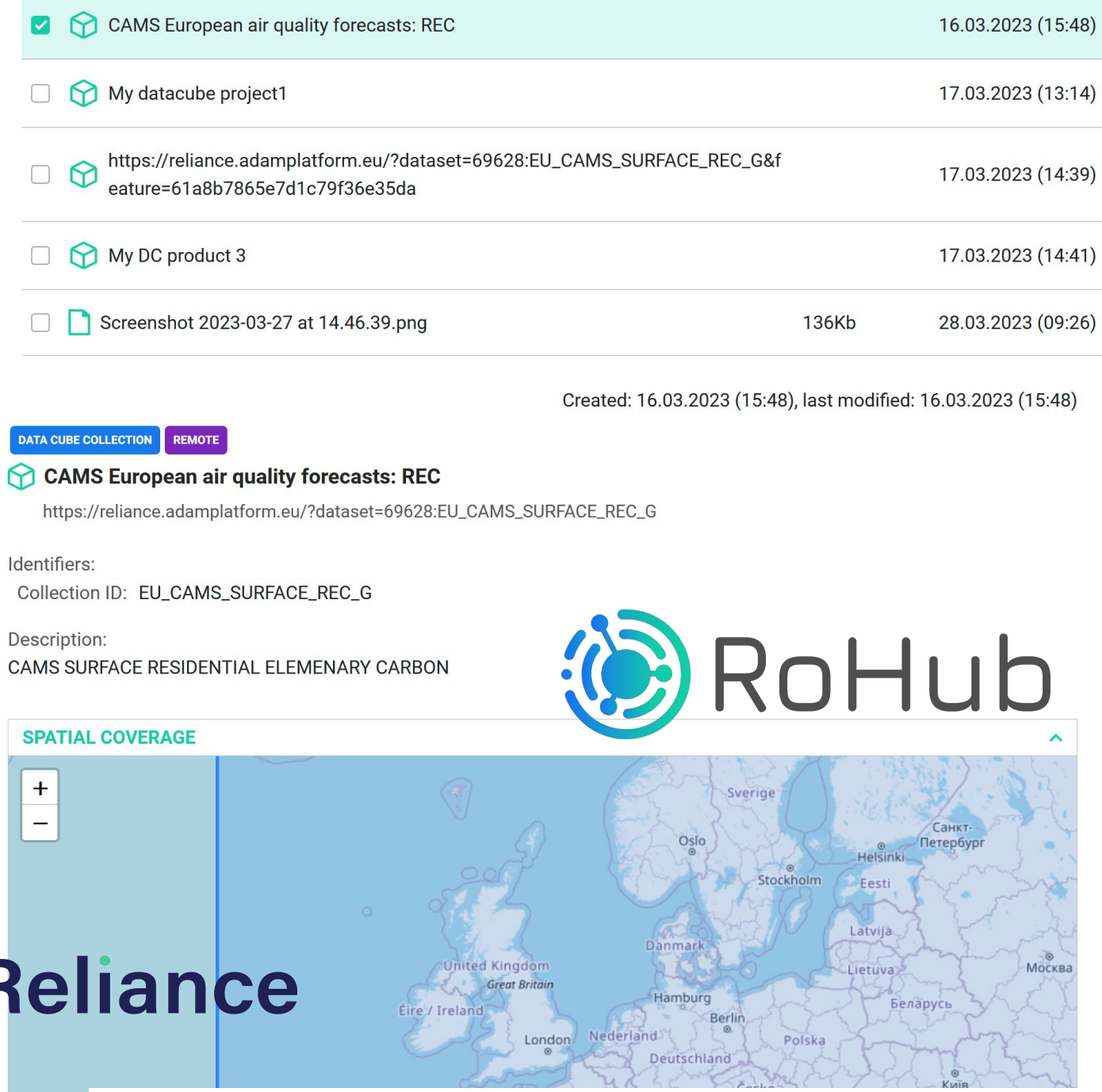
Metadata includes **temporal** coverage, **spatial** coverage and **vertical** coverage.

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

Reliance

https://reliance.rohub.org/

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



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

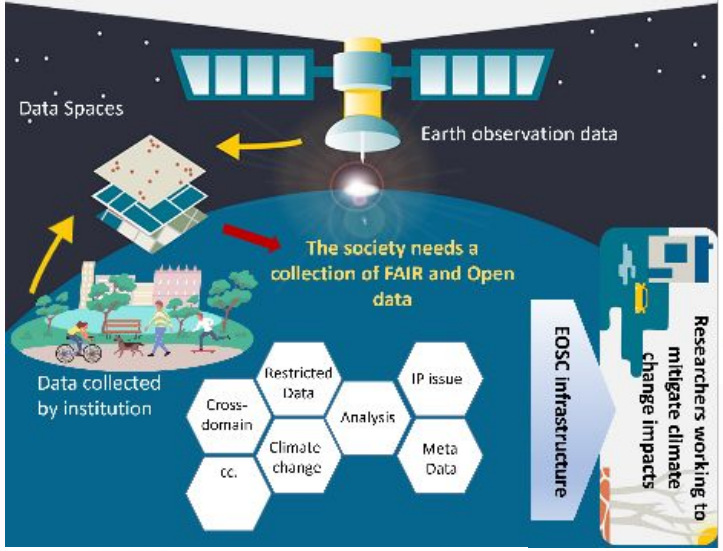
Climate-Adapt4EOSC

New [HORIZON-INFRA-2024-EOSC-01-01](#) projects starting ~ Jan 2025: FAIR2ADAPT, Climate-Adapt4EOSC

ClimateAdapt will use ontologies, knowledge graphs and DDI-CDI to map diverse set of Earth science data spaces, sources and APIs.

RO-Crate captures **provenance** of FAIRification transformations, carries onwards **licenses** and **attributes**.
→ RO-Crate as metadata vessel

Legal & Policy engine reasons on acceptable use of combined data sources.
→ RO-Crate as *evidence store*.



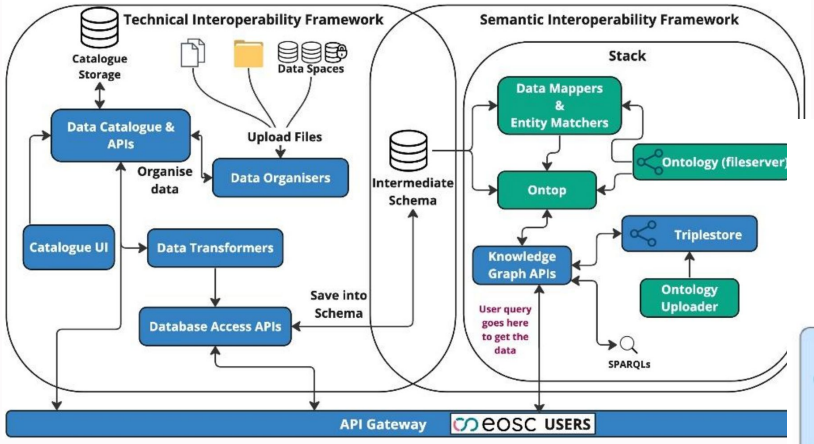
1 ClimateAdapt4 eosc

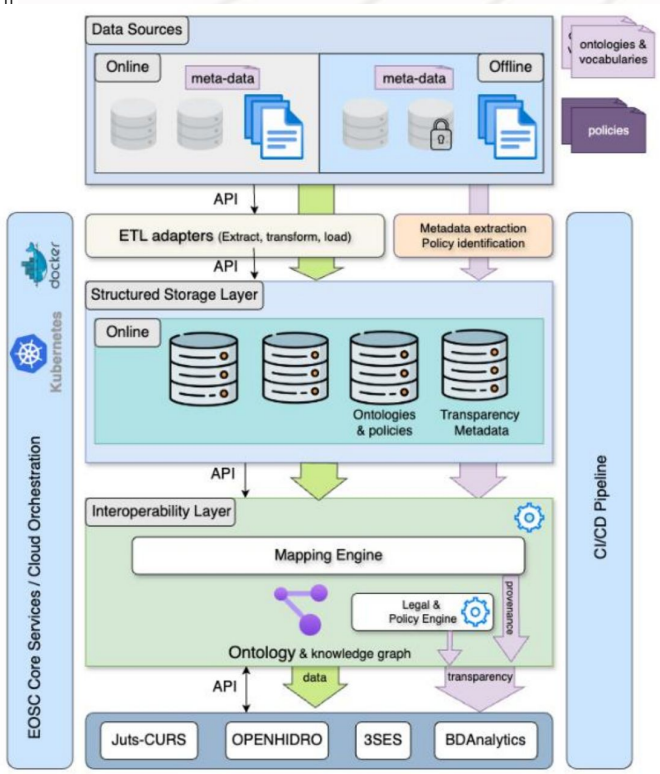
2 ClimateAdapt eosc

3 ClimateAdapt4 eosc

4 ClimateAdapt eosc

Logo to be decided..

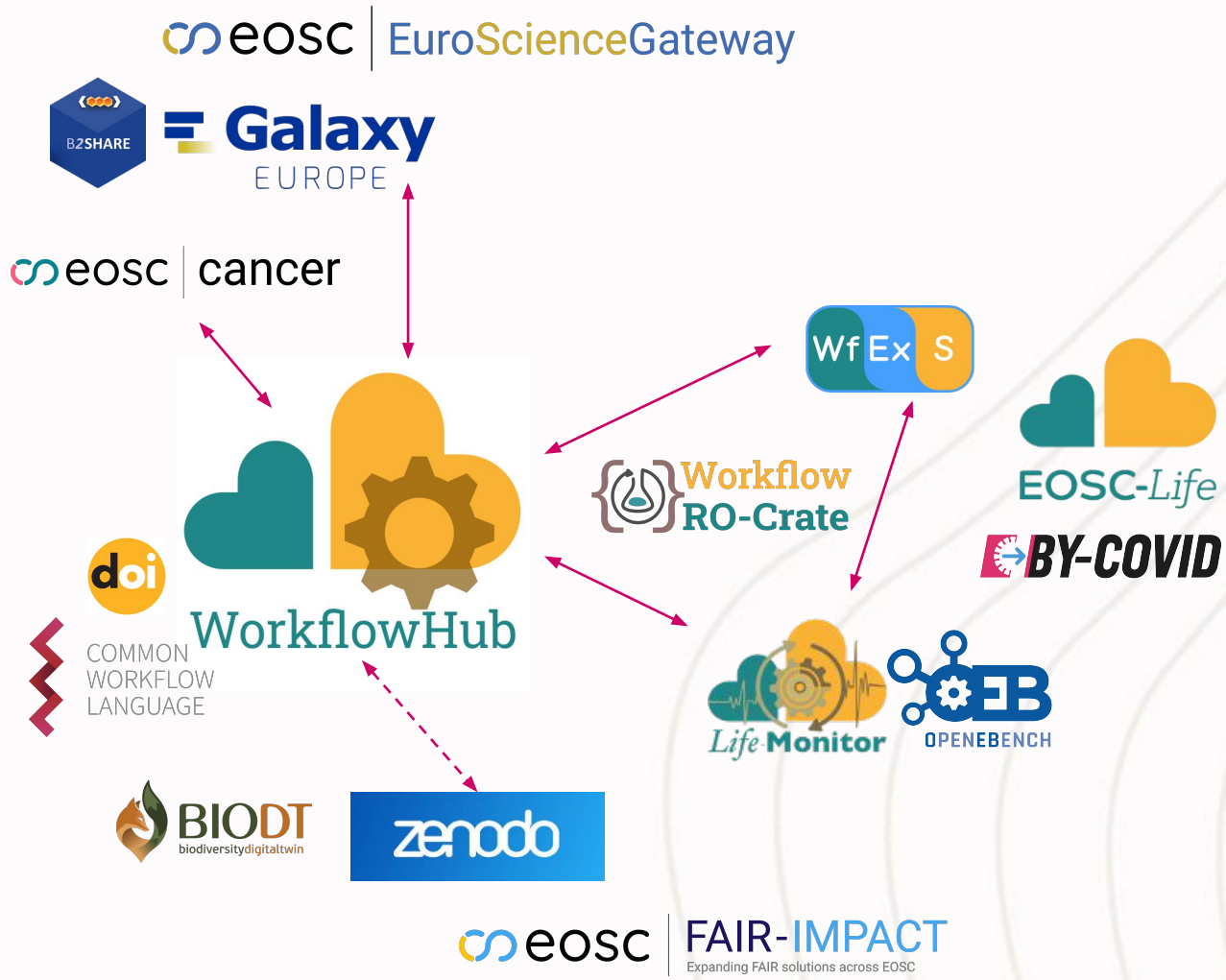




eosc | FAIR-EASE

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/

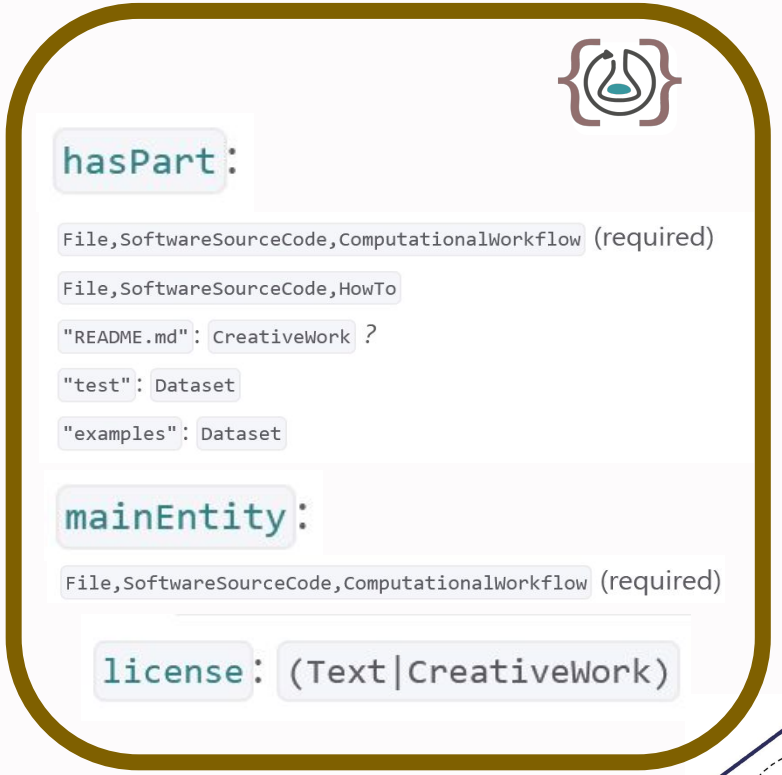
https://www.researchobject.org/ro-crate/profiles.html

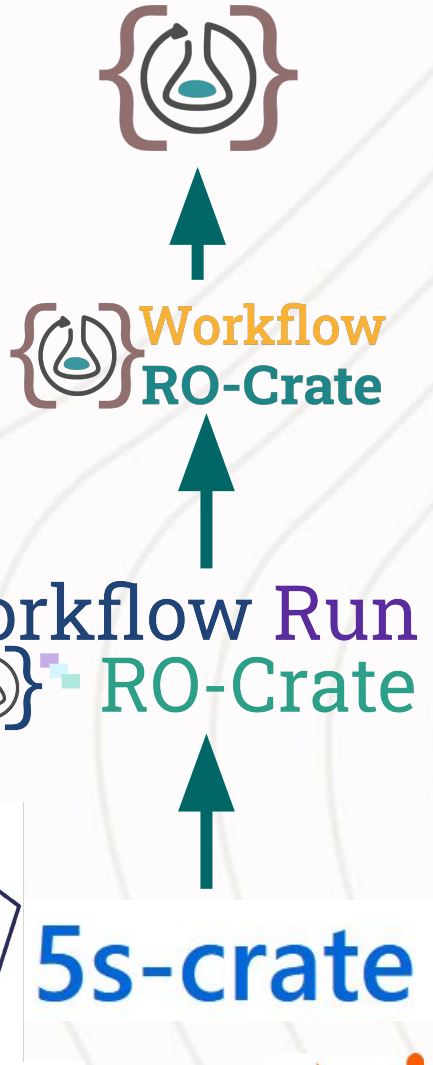
eosc | FAIR-EASE

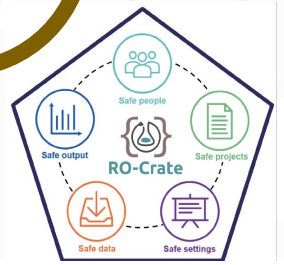
RO-Crate Profiles: Community defined Content Checklists

Profiles give a set of conventions, types and properties that minimally require and expect to be present in that subset of RO-Crates.

- **Duck typing** for creation, consumption, rendering
- **Classification** for finding and comprehension.
- **Profile Crate** for further defining RO-Crates profile resources
- Key to **extensibility and diversity**







Recording provenance of workflow runs with RO-Crate, https://arxiv.org/abs/2312.07852

eosc | FAIR-EASE

RO-Crate tutorials

https://www.researchobject.org/ro-crate/tutorials

Sandbox (experimental): https://ro-crate.ldaca.edu.au/

Python: https://pypi.org/project/rocrate/

Schedule

	Setup	Download files required for the lesson
00:00	1. Introduction	How do I package data in a FAIR way? How can I list the authors of individual files? Can I use multiple licenses in the same data package? How can I visualize JSON-LD metadata?
00:17	2. Turning a folder into an RO-Crate	How can I start a new RO-Crate?
00:24	3. Making a metadata descriptor	Which RO-Crate version is used? How can the crate self-identify as an RO-Crate?
00:28	4. Declaring the root folder	What is the root folder?
00:31	5. Describing the root entity	How can I describe the crate? How do I specify the license of the RO-Crate?
00:39	6. Adding cross-references	How can I describe an entity further? How can I cross-reference different entities?
00:46	7. Data entities	How do I describe the files in my RO-Crate?
00:50	8. Contextual entities	How can I describe things in the world? How can I give details about licenses?
00:54	9. Authorship in crates	How can I list who made the content of the crate? How do I affiliate a person with their place of work?
01:00	10. Validating JSON-LD	How can I validate the JSON-LD?

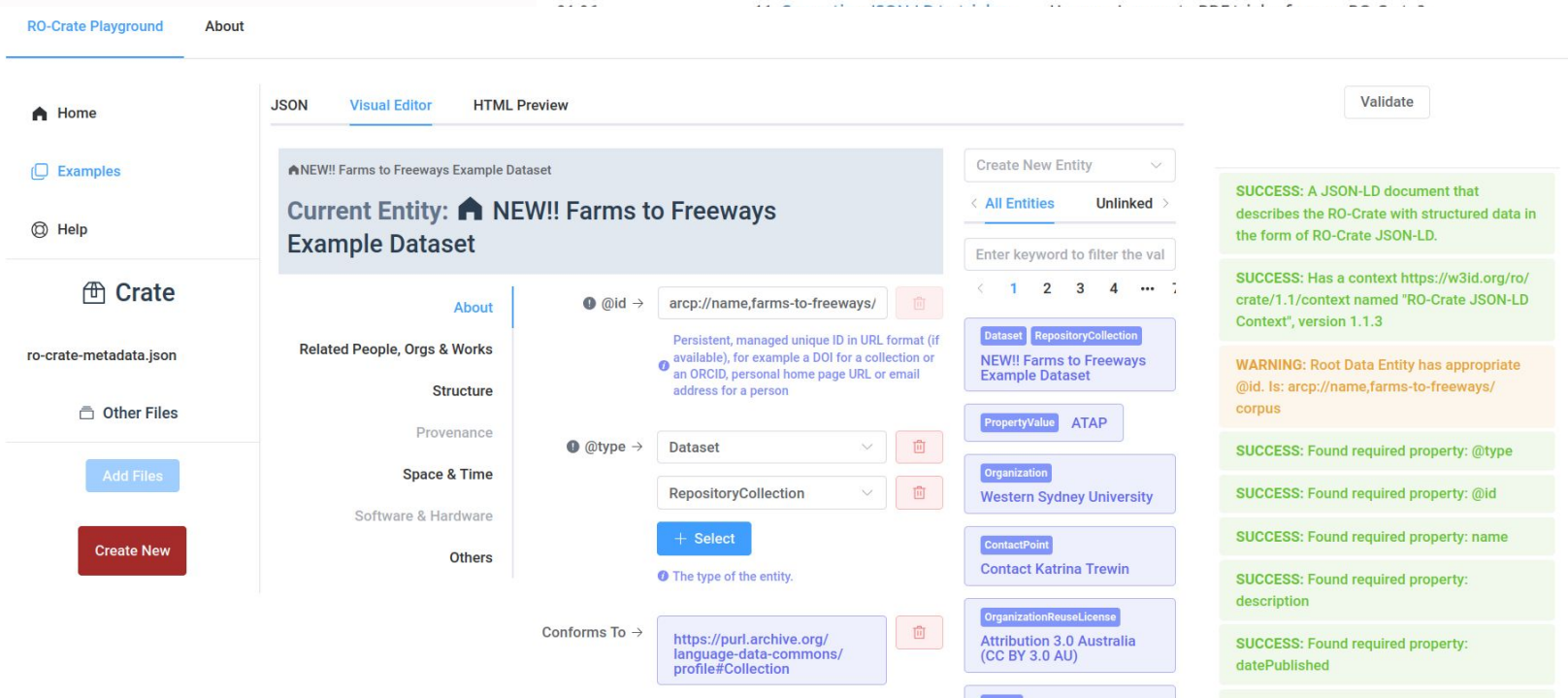
Creating an RO-Crate

In its simplest form, an RO-Crate is a directory tree with an `ro-crate-metadata.json` file at the top level that contains metadata about the other files and directories, represented by data entities. These metadata consist both of properties of the data entities themselves and of other, non-digital entities called contextual entities (representing, e.g., a person or an organization).

Suppose Alice and Bob worked on a research task together, which resulted in a manuscript written by both; additionally, Alice prepared a spreadsheet containing the experimental data, which Bob used to generate a diagram. Let's make an RO-Crate to package all this:

```
from rocrate.rocrate import ROCrate

crate = ROCrate()
paper = crate.add_file("exp/paper.pdf", properties={
    "name": "manuscript",
    "encodingFormat": "application/pdf"
})
table = crate.add_file("exp/results.csv", properties={
    "name": "experimental data",
    "encodingFormat": "text/csv"
})
diagram = crate.add_file("exp/diagram.svg", dest_path="images/figure.svg", properties={
    "name": "bar chart",
    "encodingFormat": "image/svg+xml"
})
```



eosc | FAIR-EASE
Building Interoperable Earth Science & Environmental Services

General Introduction on Interoperability:
Work done in FAIR-Impact

Esteban González Guardia,
Universidad Politécnica de Madrid

FAIR-EASE has received funding from the European Union's Horizon Europe Framework Programme (HORIZON) - under grant agreement No. 101058785.

What is interoperability?

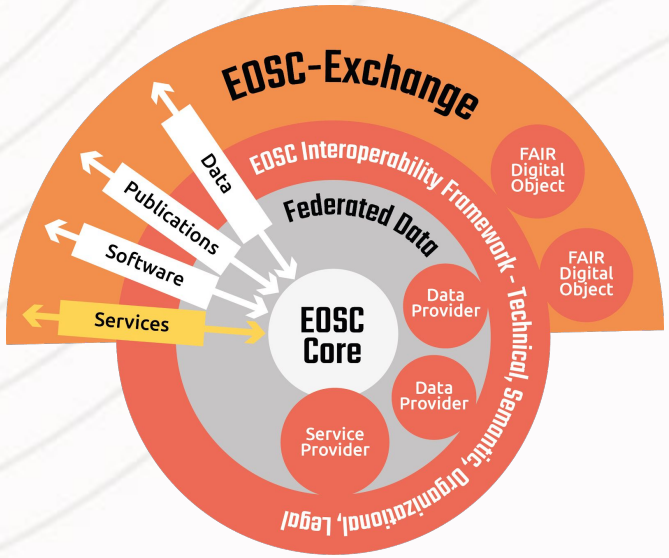
- **Interoperability** - a characteristic of an **IT system**, whose **interfaces are completely understood**, to work with **other IT systems**, at present or in the future, in either implementation or access, without any restrictions or with a controlled access [Source: Wikipedia]

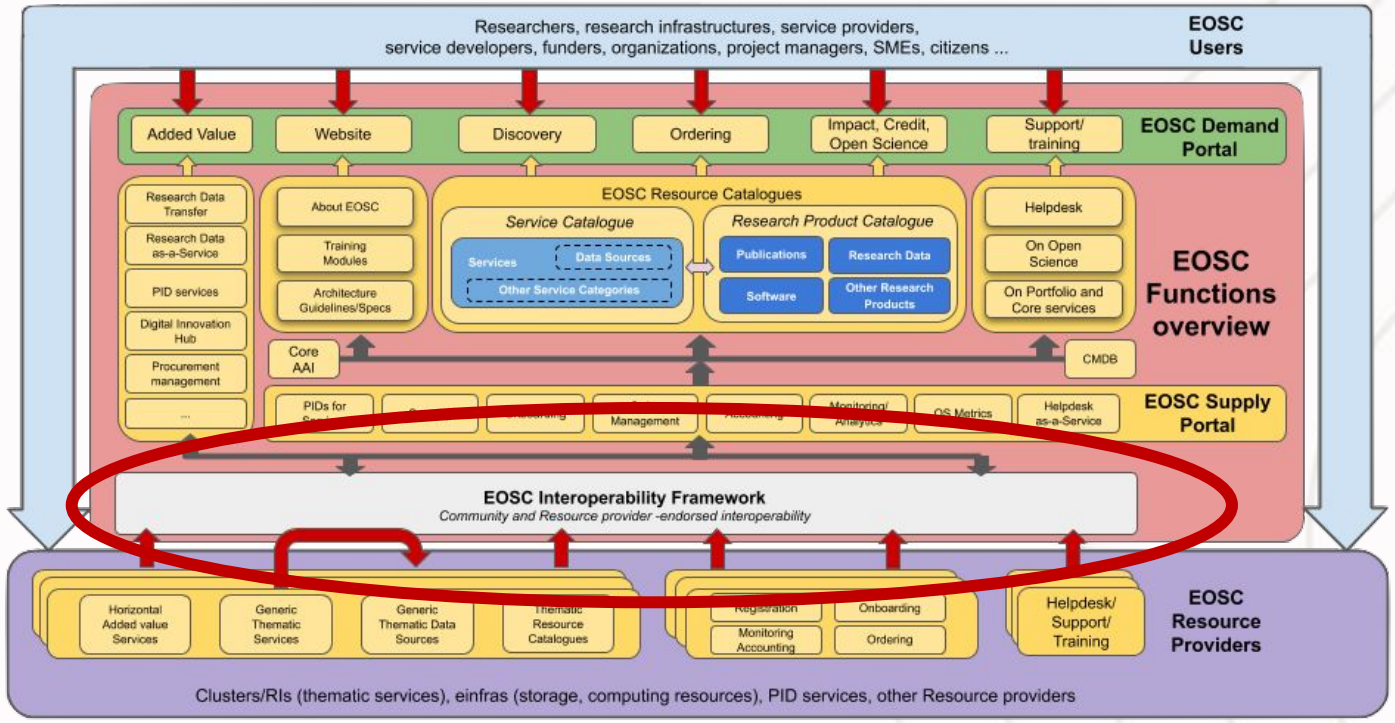
- **Interoperability** is focused on making sure that the **data can be integrated with other data**, and can be **used with applications or workflows** for analysis, storage, and processing. Furthermore, the following principles are identified (for data and its corresponding metadata) [Source: FAIR Data Principles: https://doi.org/10.1038/sdata.2016.18]:
  - I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
  - I2. (Meta)data use vocabularies that follow FAIR principles
  - I3. (Meta)data include qualified references to other (meta)data

Source: Semantic challenges of Open Science, the EOSC perspective. Oscar Corcho (UPM). 2022

EOSC Interoperability Framework

Successful, effective, homogeneous and sustainable "data" sharing inside and across research communities

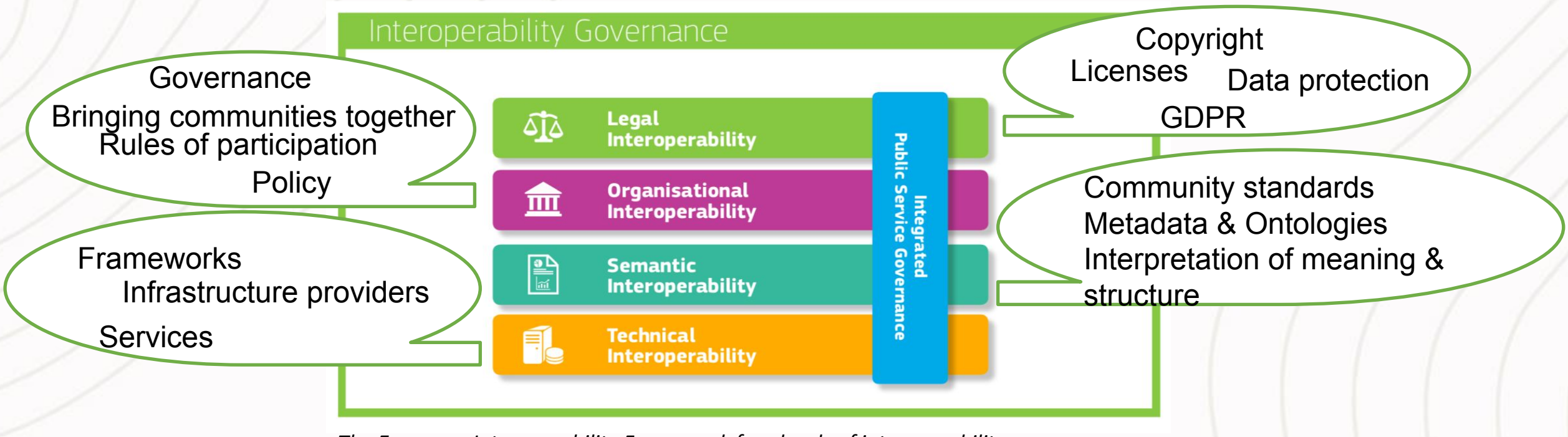




This architecture will change in the future with the adoption of the EOSC Nodes

Source: European Commission: Directorate-General for Research and Innovation, Corcho, O., Eriksson, M., Kurowski, K., Ojsteršek, M. et al., *EOSC interoperability framework – Report from the EOSC Executive Board Working Groups FAIR and Architecture*, Publications Office, 2021, **https://data.europa.eu/doi/10.2777/620649**

Layers on interoperability in EOSC IF



The European Interoperability Framework four levels of interoperability

Source: Semantic challenges of Open Science, the EOSC perspective. Oscar Corcho (UPM). 2022

Interoperability & Research Objects

What is the role of Research Objects & RO-Crate?

Interoperability & Research Objects

- In the context of EOSC IF, they use the term **Digital Object** to refer to the kind of objects that allow binding all critical information about any entity.
- It includes research data, software, scientific workflows, hardware designs, protocols, provenance logs, publications, presentations, etc., as well as all **their metadata** (for the complete object and for its constituents)
- Research Objects, plus RO-Crate, are examples of Digital Objects.
  - It is a good option for sharing information.
  - Adaptable to resource's nature.
  - Metadata defined by semantic artefacts

Source: European Commission: Directorate-General for Research and Innovation, Corcho, O., Eriksson, M., Kurowski, K., Ojsteršek, M. et al., *EOSC interoperability framework – Report from the EOSC Executive Board Working Groups FAIR and Architecture*, Publications Office, 2021, **https://data.europa.eu/doi/10.2777/620649**

Challenges

How to assess the FAIRness of a Research Object

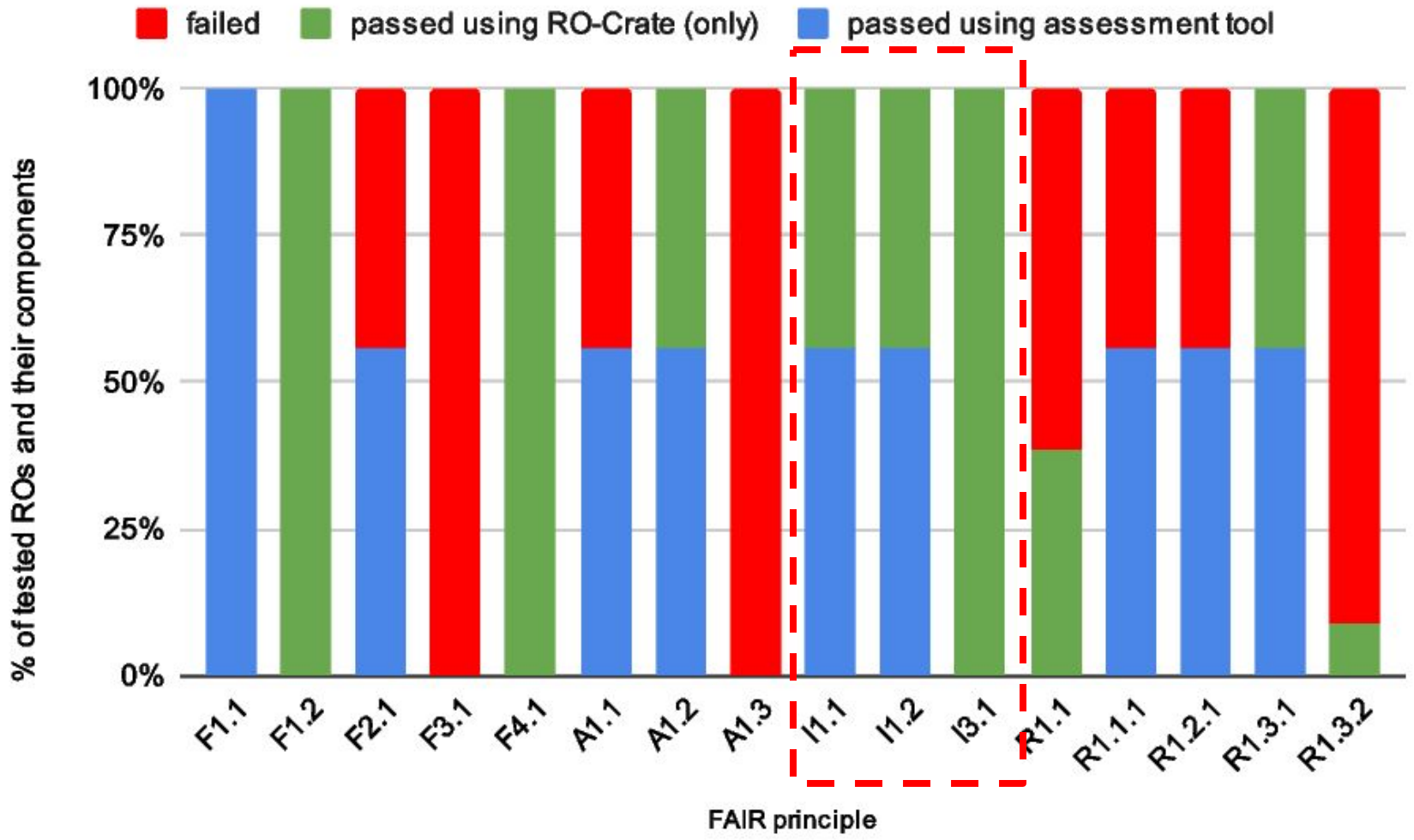
Can we use them for legal interoperability?

FAIR Assessment Approach

- Aggregate the FAIRness of each component and the RO itself
  - Each component of the RO has its own FAIRness score
- We use the tool FAIROs to assess the FAIRness of the Research Objects and other external tools for each resource.
  - Datasets: F-UJI
  - Ontologies -> FOOPS
  - Software -> Custom tool
- How can we aggregate all obtained results?
  - Score system: Using aggregation metrics (we have applied two)
  - Is it a score system enough?
    - Provenance needed!

Source: Wolniewicz, M. FAIR Research Object Assessment: A landscape analysis.

Landscape analysis on Rohub



Source: Wolniewicz, M. FAIR Research Object Assessment: A landscape analysis.

Interoperable

We analyzed ROs from ROHub.

ROHub uses JSON-LD, which represents knowledge using the RDF standard.

RO-Crate relies on Schema.org, a popular vocabulary to describe resources on the Web

Landscape analysis on Rohub: Conclusion

Using an specification like RO-Crate to describe Research Objects gives a more realistic picture of their FAIRness.

Source: FAIR IMPACT

Legal Interoperability

We study the impact of the GDPR on legal interoperability in practice.

We created a landscape analysis to study the foundation for a **legal interoperability framework**, by evaluating the relevance of existing widely adopted metadata schemas and controlled vocabularies used in data repositories for the description of legal constraints.

Source: Rouchon, O., Kraaikamp, E., Gonzalez, E., Fink Kjeldgaard, A. S., Pedersen Tenderup, N., Davidson, J., Hodson, S., Rettberg, N., & Scharnhorst, A. (2024). D6.2 - Core metadata schema for legal interoperability (Versión v1). Zenodo. https://doi.org/10.5281/zenodo.11104269

Legal Interoperability

- The RO-Crate specification does not include a specific element designed explicitly for describing access policies.
- There is only one attribute to refer to the entity of the property rights, *copyrightHolder*
- The RO-Crate specification does not include specific elements to describe data protection and privacy.

Source: Rouchon, O., Kraaikamp, E., Gonzalez, E., Fink Kjeldgaard, A. S., Pedersen Tenderup, N., Davidson, J., Hodson, S., Rettberg, N., & Scharnhorst, A. (2024). D6.2 - Core metadata schema for legal interoperability (Versión v1). Zenodo. https://doi.org/10.5281/zenodo.11104269

General Introduction on Interoperability: Work done in FAIR-Impact

What is interoperability in the EOSC context?

Source: Oscar Corcho, Universidad Politécnica de Madrid

eosc | FAIR-EASE
Building Interoperable Earth Science & Environmental Services

Practical Use-cases of RO-Crate

FAIR-EASE has received funding from the European Union's Horizon Europe Framework Programme (HORIZON) - under grant agreement No. 101058785.

eosc | FAIR-EASE

Use Case 1:
Workflows in Galaxy

Marie Josse, CNRS, FR

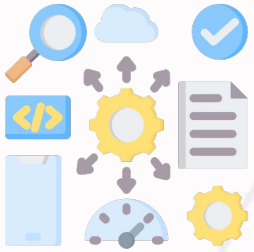
Galaxy an open source platform

https://earth-system.usegalaxy.eu/

Advantages of RO-Crate in Galaxy

- Sustainability of Galaxy's workflows

- Reproducibility and interoperability from one platform to another of Galaxy's workflows

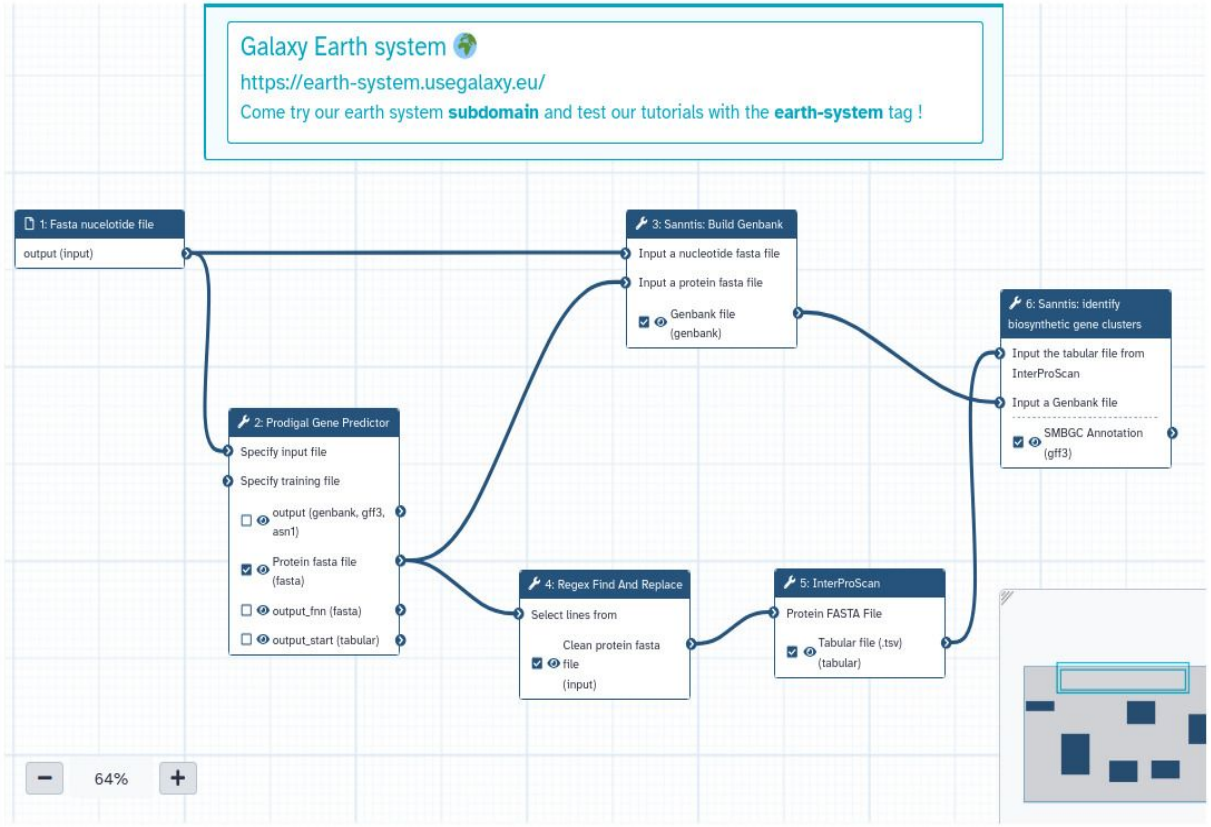


- Implemented within Galaxy's core

Workflows in Galaxy

Workflow Preview

Download | Edit | Run



About This Workflow

Marine Omics identifying biosynthetic gene clusters - Version 5

Author

marie.josse

All published Workflows by marie.josse

Creators

Marie Josse

Description

Secondary metabolite biosynthetic gene cluster (SMBGC) Annotation using Neural Networks Trained on Interpro Signatures

Tags

earth-system | Ocean | Marineomics

License

Creative Commons Attribution 4.0 International

Last Updated

Friday Aug 9th 10:07:53 2024 GMT+2

Sharing

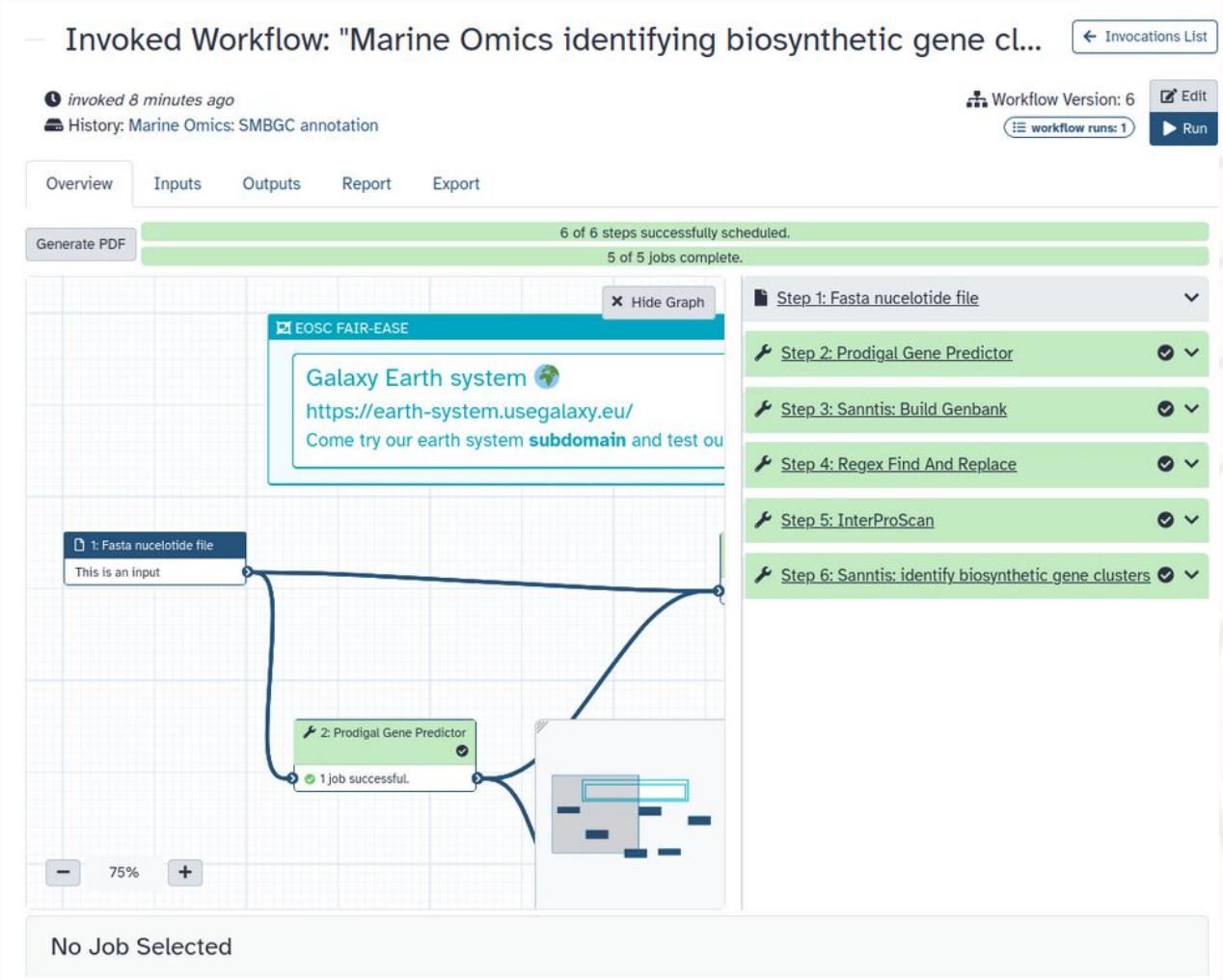
Use the following link to share preview of this workflow: https://usegalaxy.eu/published/workflow?id=b9c938d1af08124b Manage

OK

Exporting a workflow invocation in RO-Crate

- Run the workflow

- Galaxy saves all the analytical steps of the workflow in your history

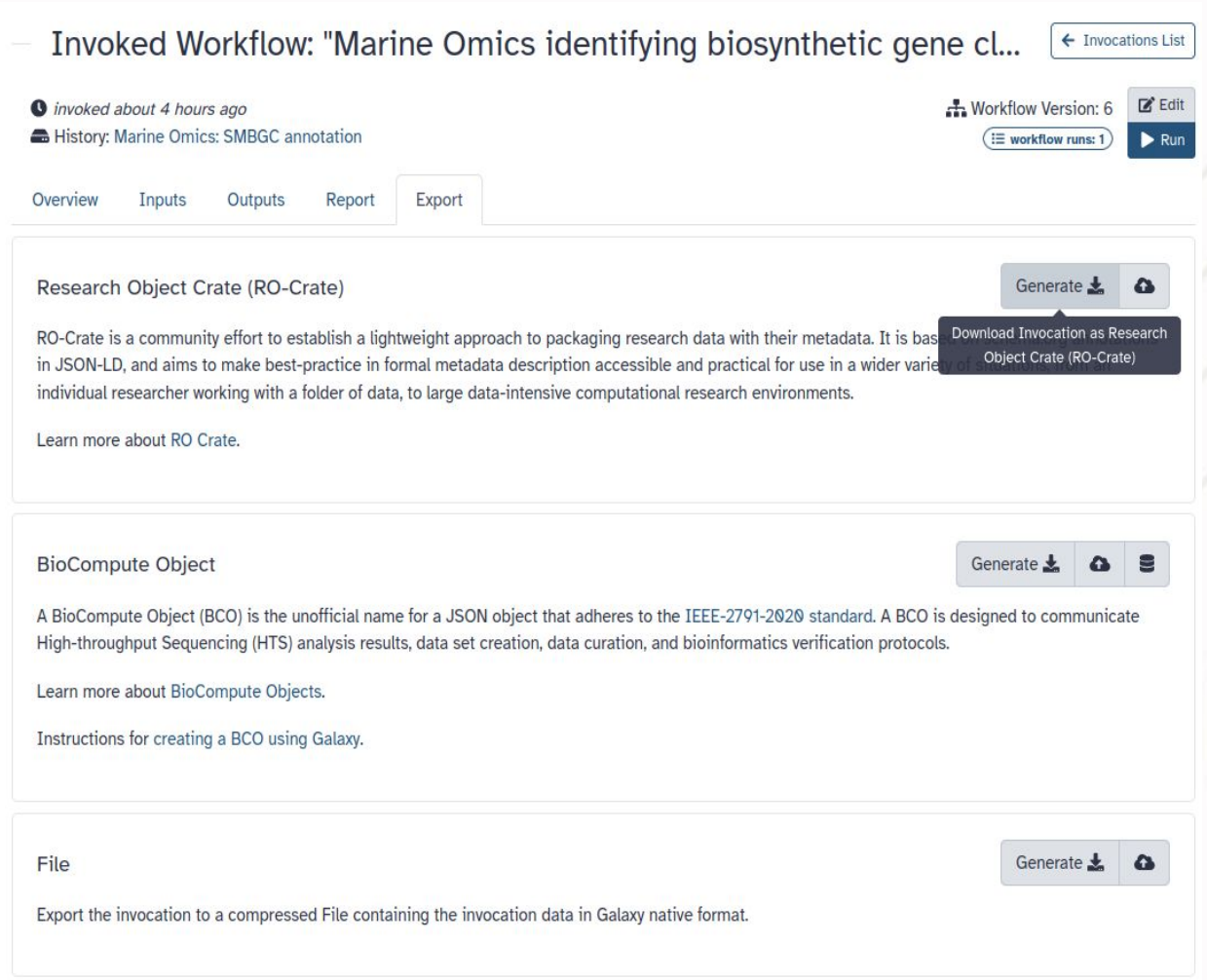


Exporting a workflow invocation in RO-Crate

- Run the workflow

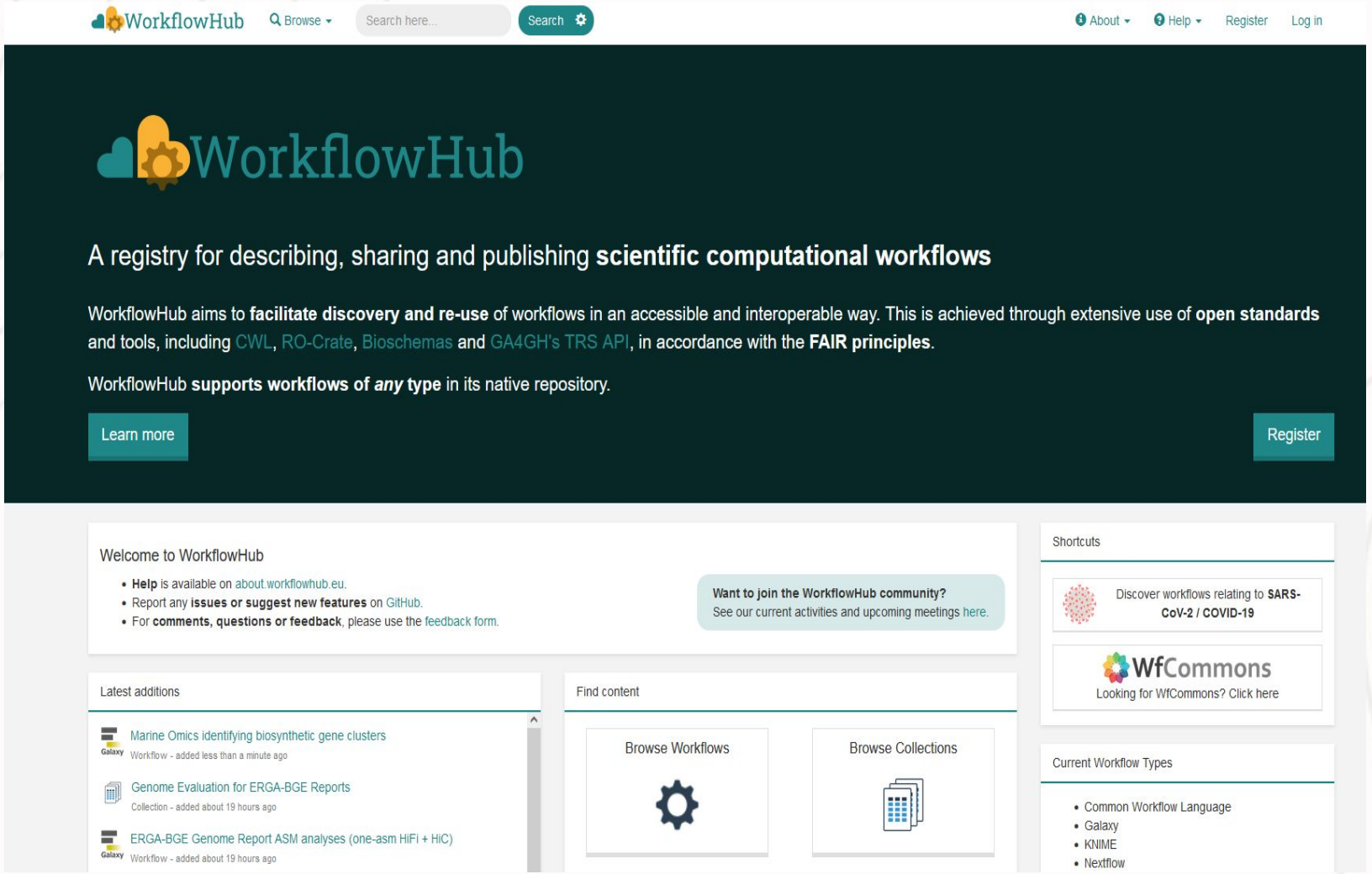
- Galaxy saves all the analytical steps of the workflow in your history

- Generate the RO-Crate from this history

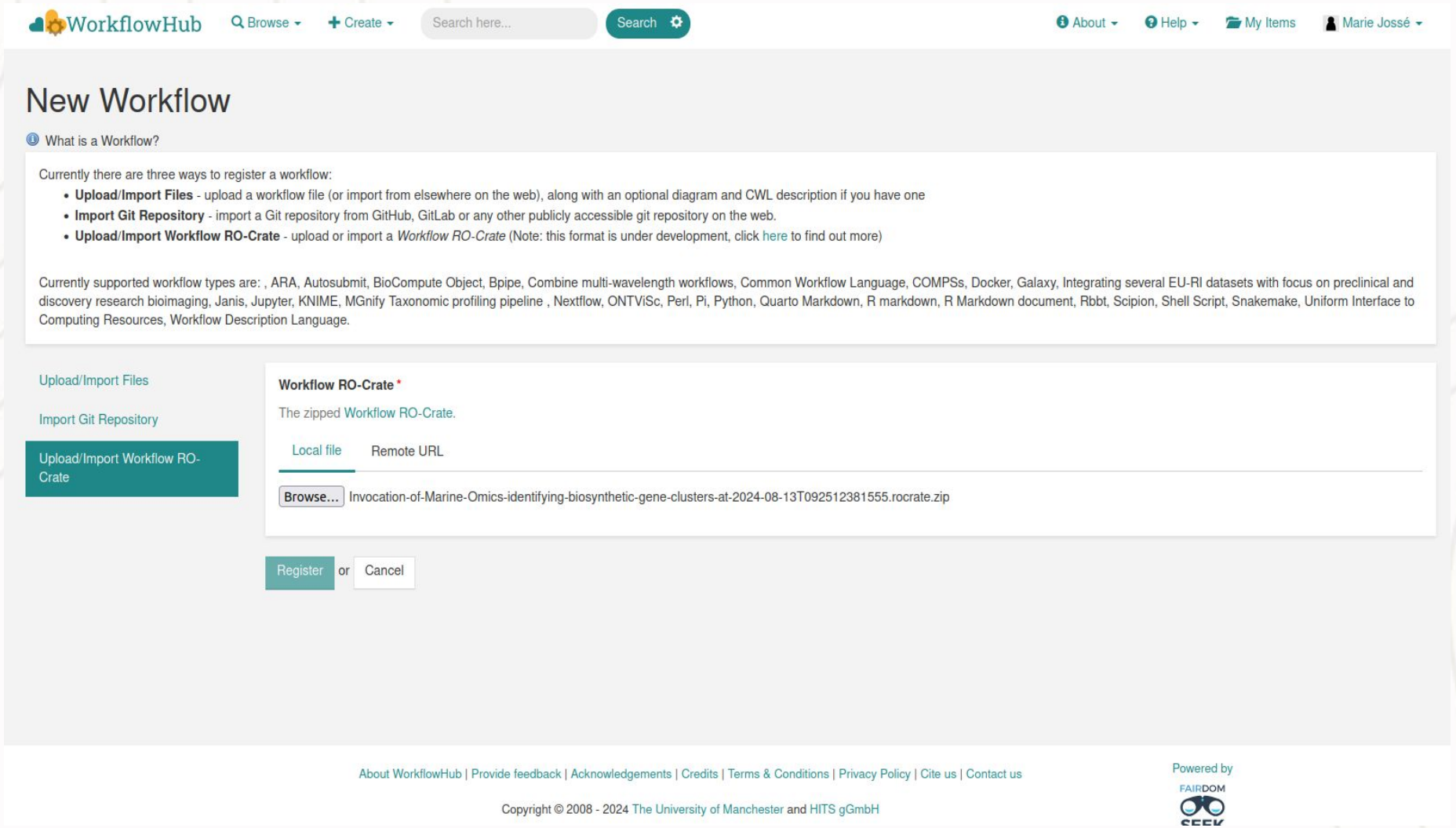


Workflowhub registry

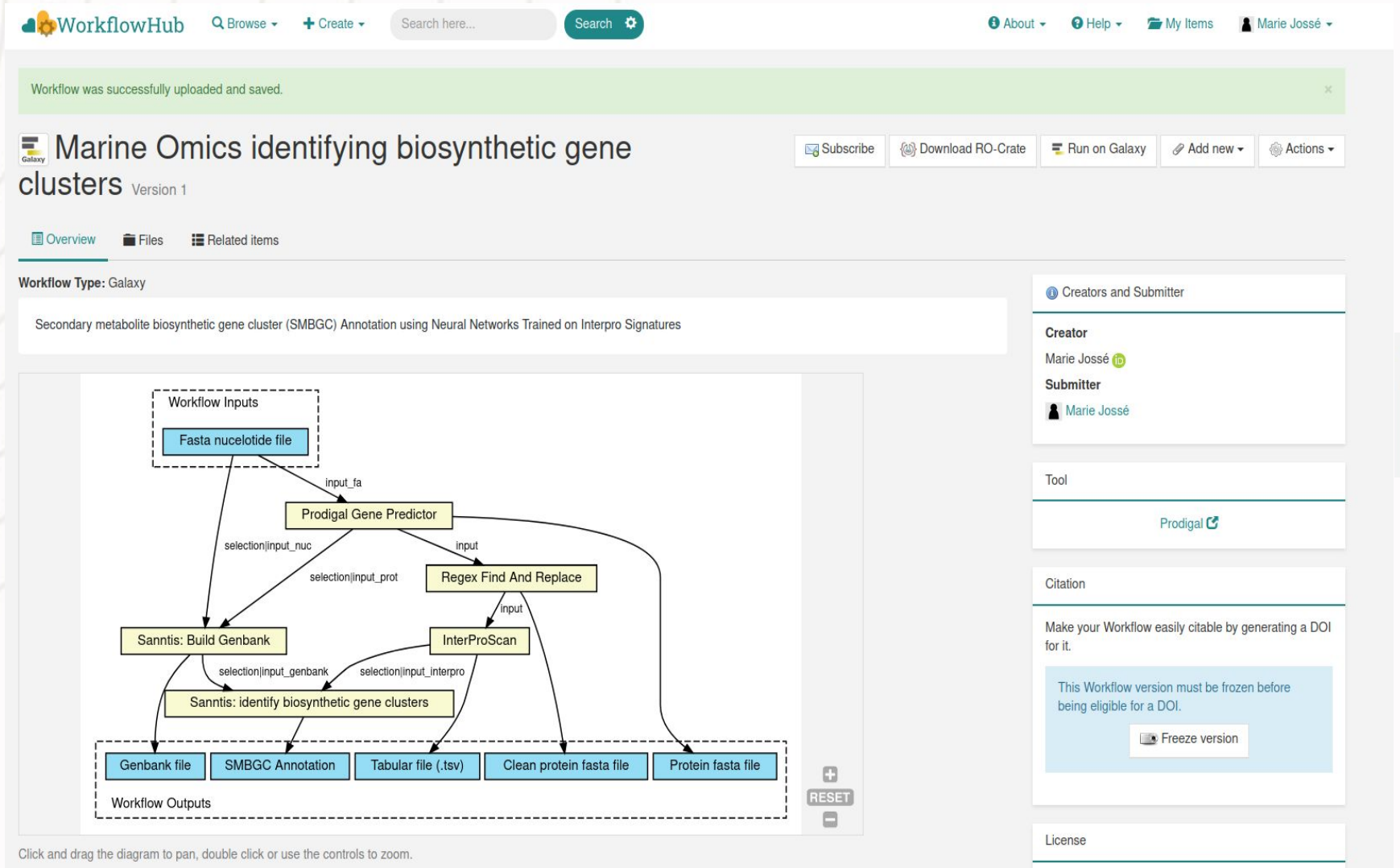
https://workflowhub.eu/



Workflowhub registry



Share and version my workflow



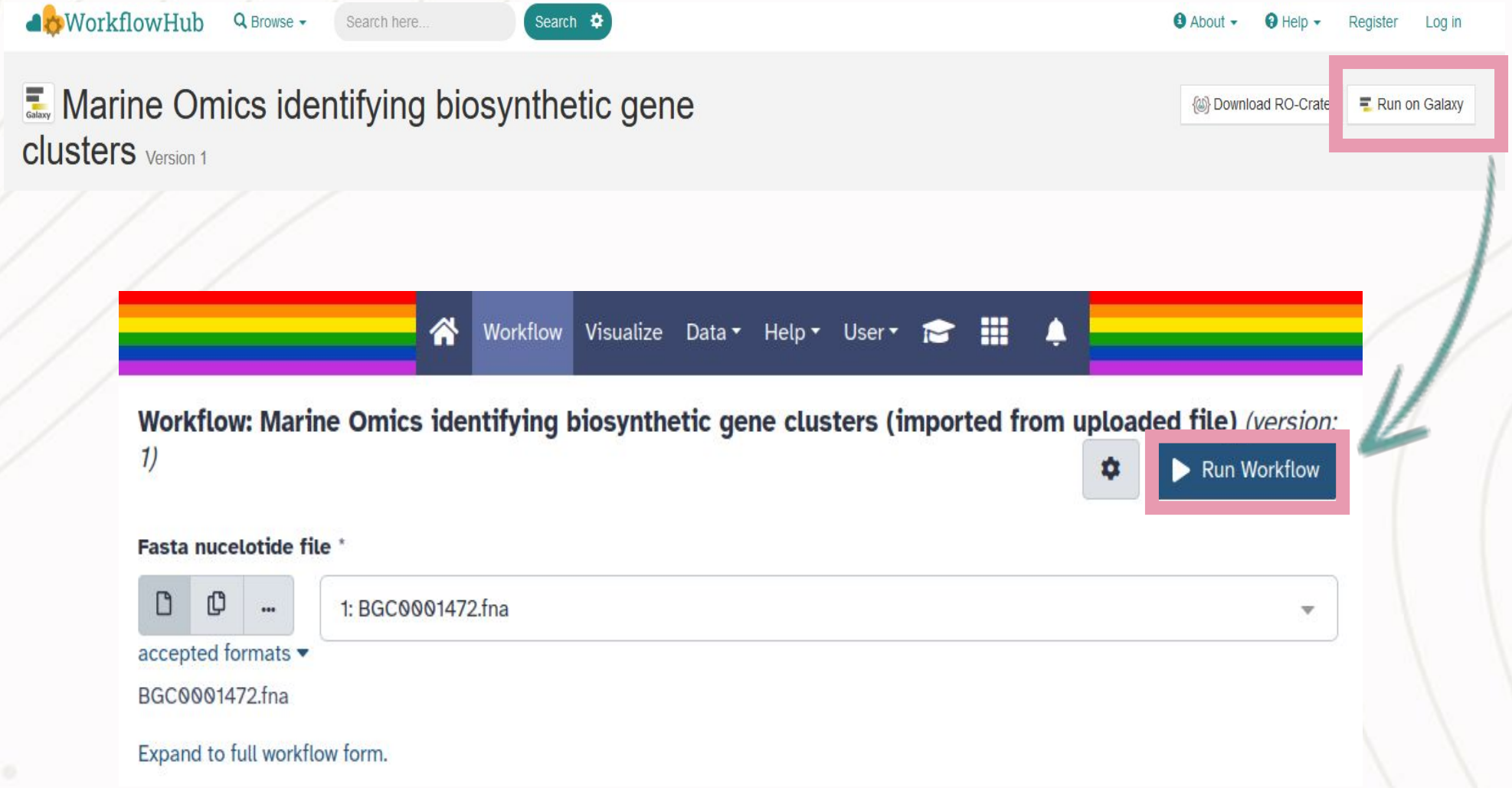
Version History

Version 1 (earliest) Created 21st Aug 2024 at 09:17 by Marie Jossé

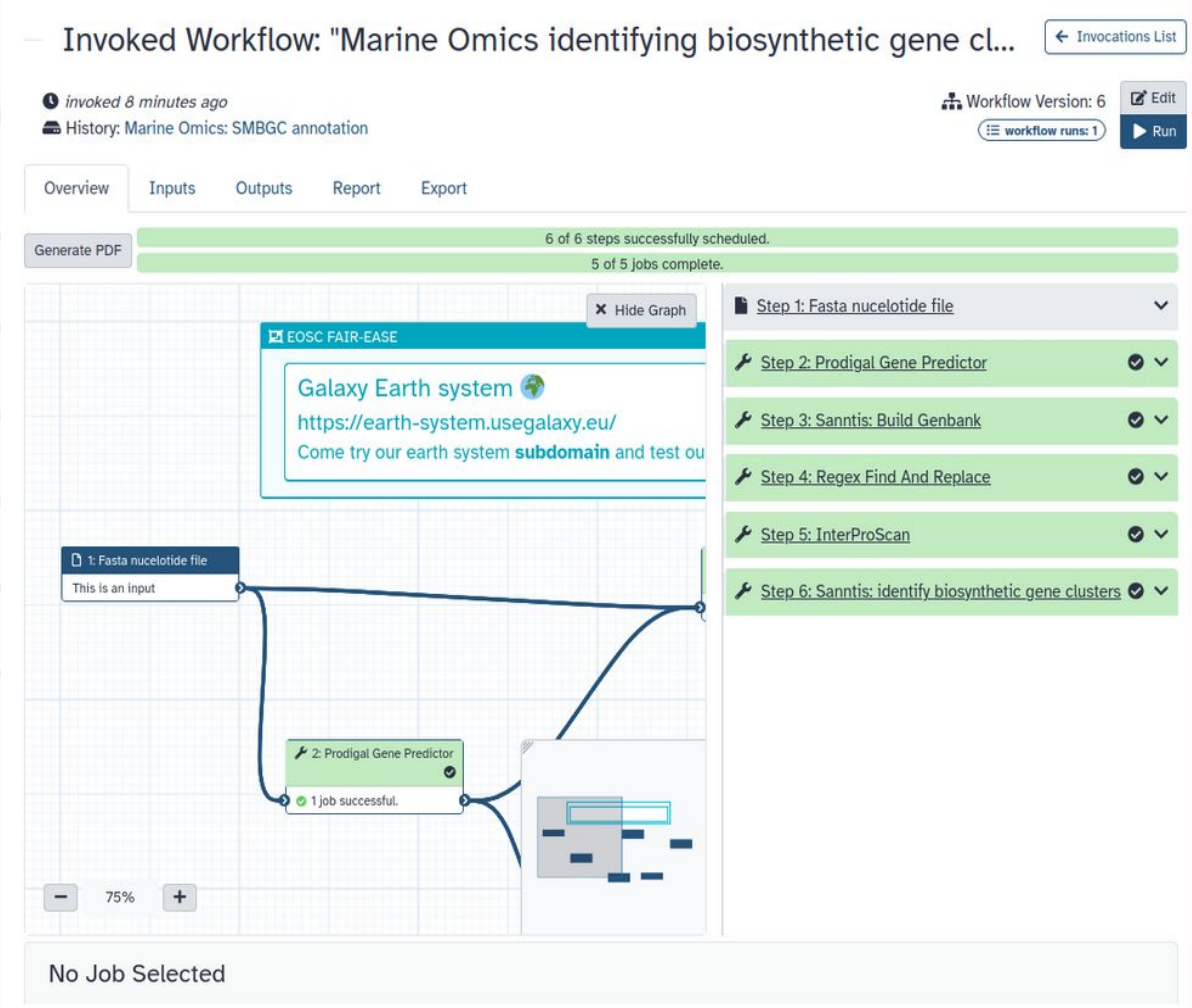
No revision comments

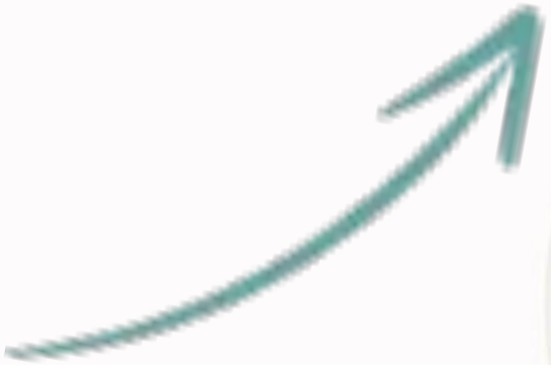
Open master cd239c1

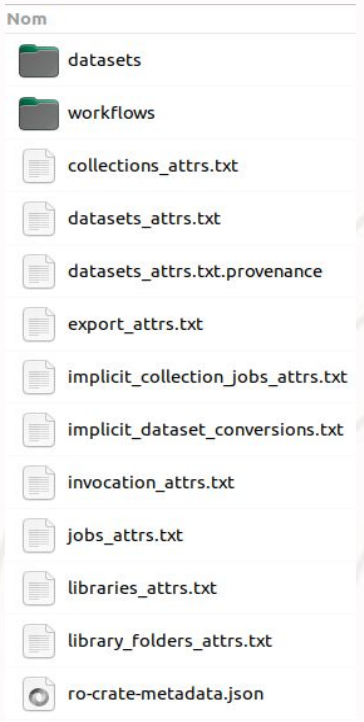
Run my workflow from Workflowhub to Galaxy



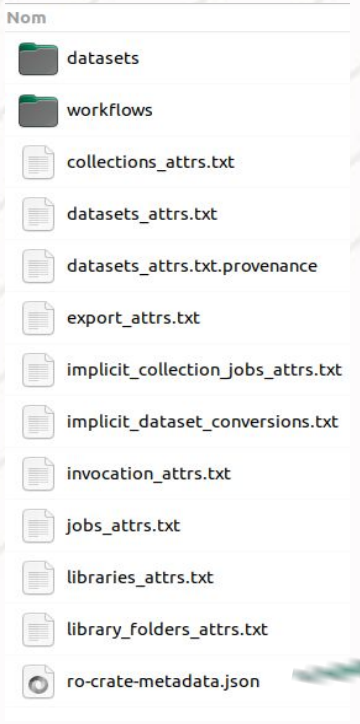
A concrete example of a RO-Crate







A concrete example of a RO-Crate




```
{
    "@context": "https://w3id.org/ro/crate/1.1/context",
    "@graph": [
        {
            "@id": "./",
            "@type": "Dataset",
            "conformsTo": [
                {
                    "@id": "https://w3id.org/ro/wfrun/process/0.1"
                },
                {
                    "@id": "https://w3id.org/ro/wfrun/workflow/0.1"
                },
                {
                    "@id": "https://w3id.org/workflowhub/workflow-ro-crate/1.0"
                }
            ],
            "datePublished": "2024-08-12T07:36:28+00:00",
            "hasPart": [
                {
                    "@id": "workflows/a44ffd6c34f64433.ga"
                },
                {
                    "@id": "workflows/a44ffd6c34f64433.gxwf.yml"
                },
                {
                    "@id": "workflows/a44ffd6c34f64433.abstract.cwl"
                },
                {
                    "@id": "workflows/a44ffd6c34f64433.html"
                },
                {
                    "@id": "datasets/BGC0001472.fna_31e7840b5aedca4365fcd5423bbf227f.fasta"
                },
                {
                    "@id": "datasets/Prodigal_Gene_Predictor_on_data_1___protein_translations_file_31e7840b5aedca43c08e5935886a0a6c.fasta"
                },
                {
                    "@id": "datasets/Regex_Find_And_Replace_on_data_11_31e7840b5aedca43f17b497c109f9250.fasta"
                },
                {
                    "@id": "datasets/Sanntis_output_data_31e7840b5aedca4309892dd15e10502a.gff3"
                },
                {
                    "@id": "datasets/Sanntis_output_data_genbank_31e7840b5aedca43aefa44e8f242ed6d.genbank"
                },
                {
                    "@id": "datasets/InterProScan_on_data_15_(tsv)_31e7840b5aedca43a79360035a3fba29.tabular"
                },
```

A concrete example of a RO-Crate


```
{
"@context": "https://w3id.org/ro/crate/1.1/context",
"@graph": [
    {
        "@id": "./",
        "@type": "Dataset",
        "conformsTo": [
            {
                "@id": "https://w3id.org/ro/wfrun/process/0.1"
            },
            {
                "@id": "https://w3id.org/ro/wfrun/workflow/0.1"
            },
            {
                "@id": "https://w3id.org/workflowhub/workflow-ro-crate/1.0"
            }
        ],
        "datePublished": "2024-08-12T07:36:28+00:00",
        {
            "@id": "workflows/a44ffd6c34f64433.ga"
        },
        {
            "@id": "workflows/a44ffd6c34f64433.gxwf.yml"
        },
        {
            "@id": "workflows/a44ffd6c34f64433.abstract.cwl"
        },
        {
            "@id": "workflows/a44ffd6c34f64433.html"
        },
        {
            "@id": "datasets/BGC0001472.fna_31e7840b5aedca4365fcd5423bbf227f.fasta"
        },
        {
            "@id": "datasets/Prodigal_Gene_Predictor_on_data_1___protein_translations_file_31e7840b5aedca43c08e5935886a0a6c.fasta"
        },
        {
            "@id": "datasets/Regex_Find_And_Replace_on_data_11_31e7840b5aedca43f17b497c109f9250.fasta"
        },
        {
            "@id": "datasets/Sanntis_output_data_31e7840b5aedca4309892dd15e10502a.gff3"
        },
        {
            "@id": "datasets/Sanntis_output_data_genbank_31e7840b5aedca43aefa44e8f242ed6d.genbank"
        },
        {
            "@id": "datasets/InterProScan_on_data_15_(tsv)_31e7840b5aedca43a79360035a3fba29.tabular"
        },
```


By convention, in RO-Crate the @id value of ./ means that this document describes the directory of content in which the RO-Crate metadata is located. This means that where the ro-crate-metadata.json here is the RO-Crate root

Creation time of the RO-Crate 12th August 2024 at 07:36am GMT

Initial dataset uploaded by the user

Data consumed and produced by the workflow

Data produced by the 1st tool and used as input by the following tools

Data produced by the 2nd tool and used as input by the following tools

Data produced by the last tool and used as input by the following tools

Data produced by the 2nd bis tool and used as input by the following tools

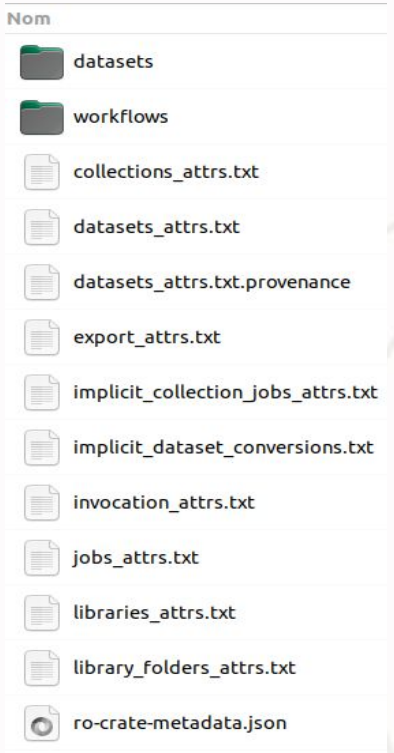
Data produced by the 3rd tool and used as input by the following tools

A concrete example of a RO-Crate


```
            "@id": "datasets/Regex_Find_And_Replace_on_data_11_31e7840b5aedca43f17b497c109f9250.fasta"
        },
        {
            "@id": "datasets/Sanntis_output_data_31e7840b5aedca4309892dd15e10502a.gff3"
        },
        {
            "@id": "datasets/Sanntis_output_data_genbank_31e7840b5aedca43aefa44e8f242ed6d.genbank"
        },
        {
            "@id": "datasets/InterProScan_on_data_15_(tsv)_31e7840b5aedca43a79360035a3fba29.tabular"
        },
        {
            "@id": "datasets_attrs.txt"
        },
        {
            "@id": "jobs_attrs.txt"
        },
        {
            "@id": "implicit_collection_jobs_attrs.txt"
        },
        {
            "@id": "collections_attrs.txt"
        },
        {
            "@id": "export_attrs.txt"
        },
        {
            "@id": "libraries_attrs.txt"
        },
        {
            "@id": "library_folders_attrs.txt"
        },
        {
            "@id": "invocation_attrs.txt"
        }
    ],
    "license": "CC-BY-4.0",
    "mainEntity": {
        "@id": "workflows/a44ffd6c34f64433.gxwf.yml"
    },
    "mentions": [
        {
            "@id": "#5e04475a-635e-4420-bbd0-5a837587018e"
        }
    ]
},
{
    "@id": "ro-crate-metadata.json",
    "@type": "CreativeWork",
```


License of the workflow

Tree structure of the folders and files next to this file the ro-crate-metadata.json



A concrete example of a RO-Crate


```
    "startTime": "2024-08-09T08:07:41.964440"
},
{
    "@id": "datasets/BGC0001472.fna_31e7840b5aedca4365fcd5423bbf227f.fasta",
    "@type": "File",
    "encodingFormat": "text/plain",
    "exampleOfWork": {
        "@id": "#12faaba3-119e-4b9d-a025-c1d7a41229a7"
    },
    "name": "BGC0001472.fna"
},
{
    "@id": "#12faaba3-119e-4b9d-a025-c1d7a41229a7",
    "@type": "FormalParameter",
    "additionalType": "File",
    "description": "",
    "name": "BGC0001472.fna"
},
{
    "@id": "datasets/Prodigal_Gene_Predictor_on_data_1___protein_translations_file_31e7840b5aedca43c08e5935886a0a6c.fasta",
    "@type": "File",
    "encodingFormat": "text/plain",
    "exampleOfWork": {
        "@id": "#3d4ac293-cadb-4ed7-8ceb-35263b29f021"
    },
    "name": "Prodigal Gene Predictor on data 1 : protein translations file"
},
{
    "@id": "#3d4ac293-cadb-4ed7-8ceb-35263b29f021",
    "@type": "FormalParameter",
    "additionalType": "File",
    "description": "",
    "name": "Prodigal Gene Predictor on data 1 : protein translations file"
},
{
    "@id": "datasets/Regex_Find_And_Replace_on_data_11_31e7840b5aedca43f17b497c109f9250.fasta",
    "@type": "File",
    "encodingFormat": "text/plain",
    "exampleOfWork": {
        "@id": "#ab15d3fd-b1db-4584-9c38-71de02b307e6"
    },
    "name": "Regex Find And Replace on data 11"
},
{
    "@id": "#ab15d3fd-b1db-4584-9c38-71de02b307e6",
    "@type": "FormalParameter",
    "additionalType": "File",
    "description": "",
    "name": "Regex Find And Replace on data 11"
```


Date of when the workflow was runned

Start of the workflow with some metadata on the initial dataset used for the analysis

1st tool used: Prodigal

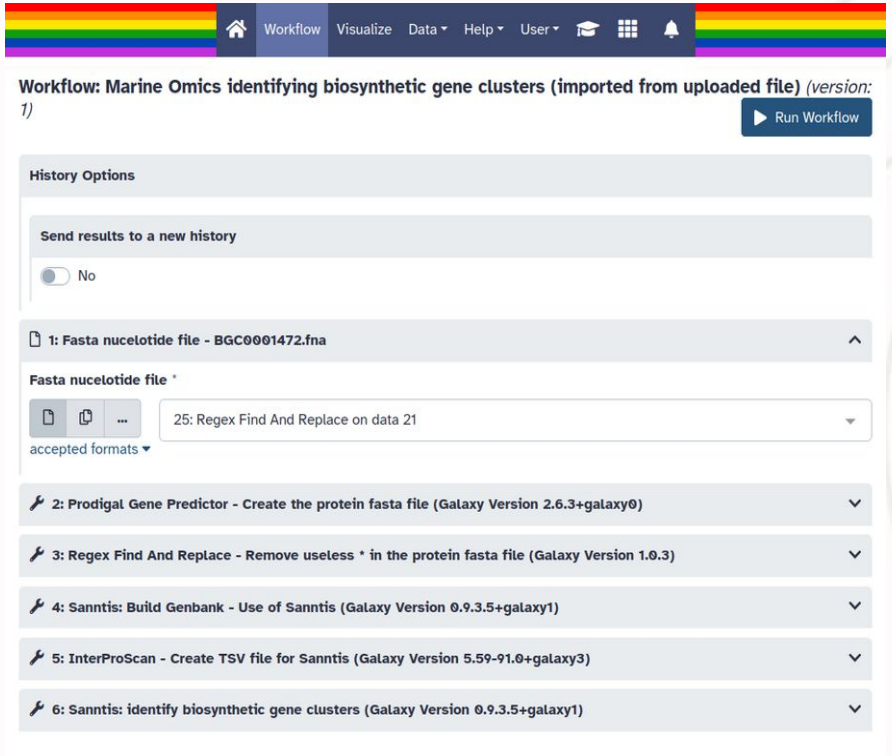
2nd tool used: Regex Find And Replace

Reuse a RO-Crate

All the datasets needed to re-run the exact same workflow

Readable format by galaxy to import the workflow

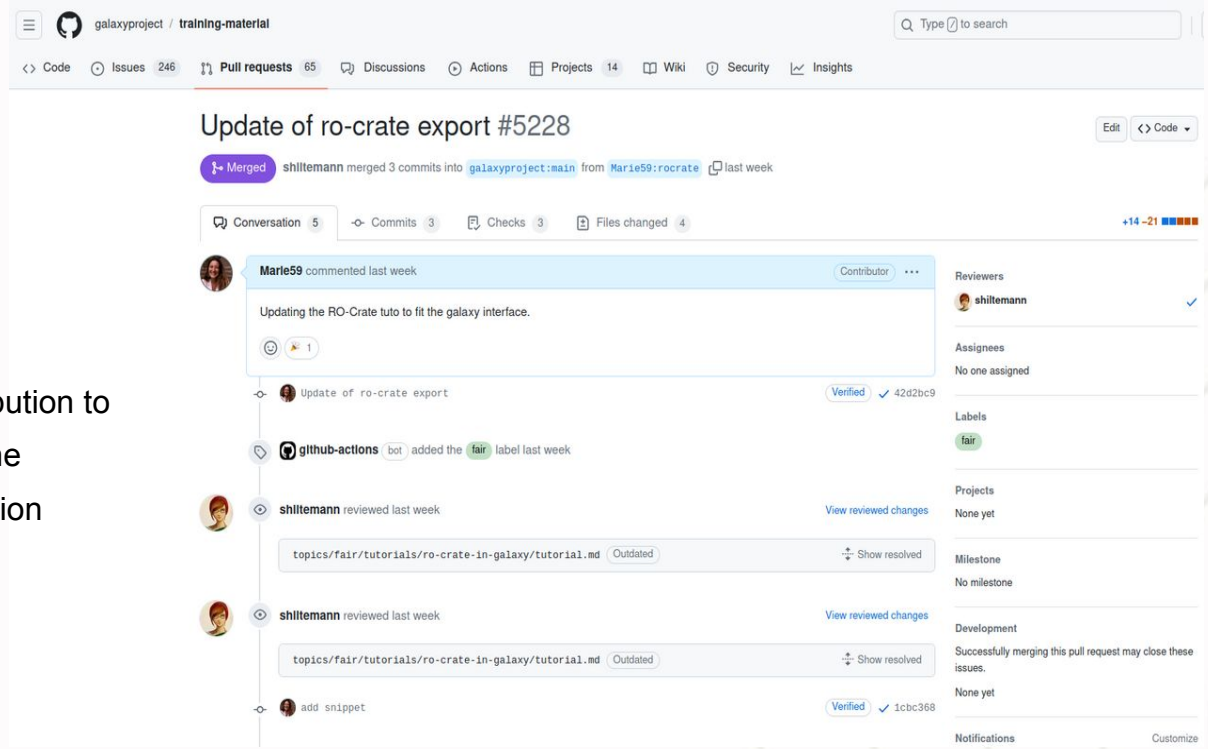
Expand to full workflow form allows you to change all the different parameter of each step before running the workflow



Contribution opportunities

- Ameliorate the RO-Crate profile
- Add details on the resources used (CPU, GPU, run time,...)
- Documentation
- Add the metadata of the tools
- Add possibility to directly import a RO-Crate in Galaxy

A start of contribution to improve the documentation



With FAIR-EASE we wish to build up knowledge on RO-Crate (we're starting from scratch) to contribute to ameliorate the RO-Crate functionalities of Galaxy

THANK YOU !

This a work in collaboration with :

eosc | AquaINFRA

eosc | EuroScienceGateway

eosc | FAIR-EASE

Galaxy EUROPE

Galaxy Training Network

eosc | FAIR-EASE

Use Case 2:
Data management

Marc Portier, VLIZ, BE

Use Case 2: Data management

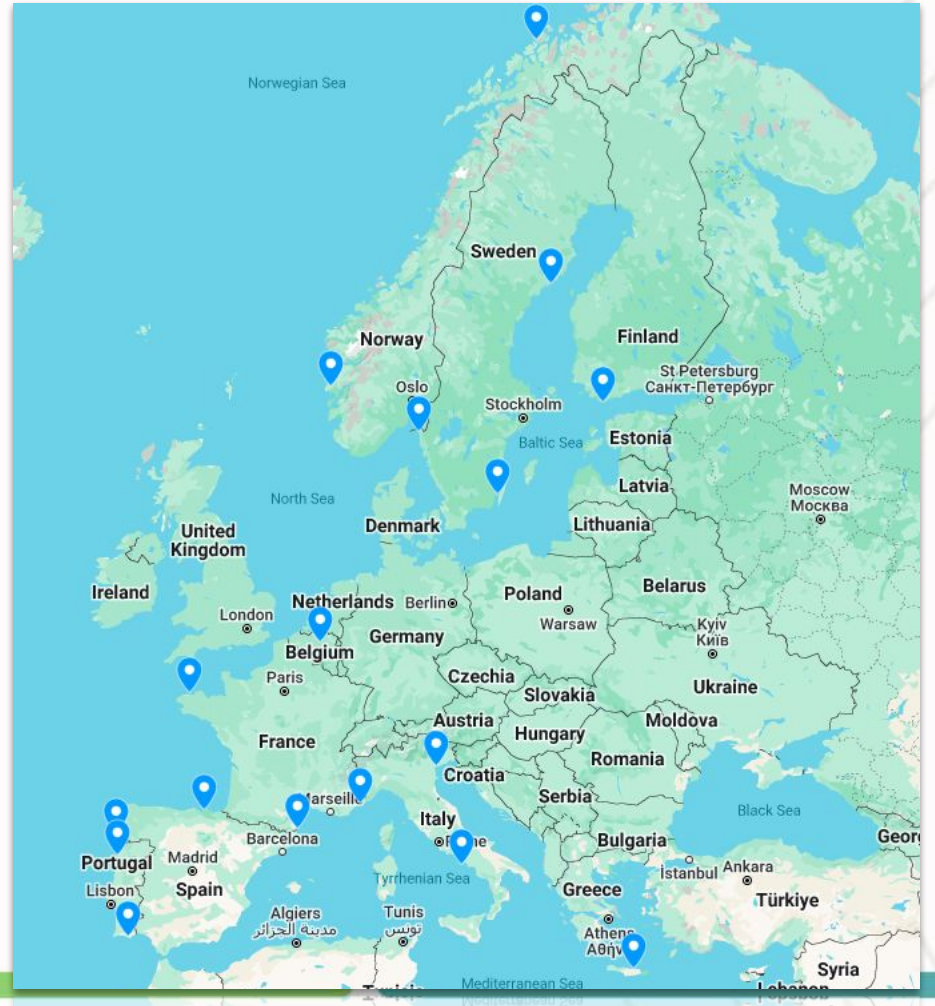
data.emobon.embrc.eu
combining git + rocrates
roprofiles
achieving LOD publication of data through gh actions and
    fair-signposting, linked open data
    large file support tech detail
    support tech harvesting
human consumption
    browsing the web approach
    analysis based on linked open data

EMBRC.eu/emo-bon

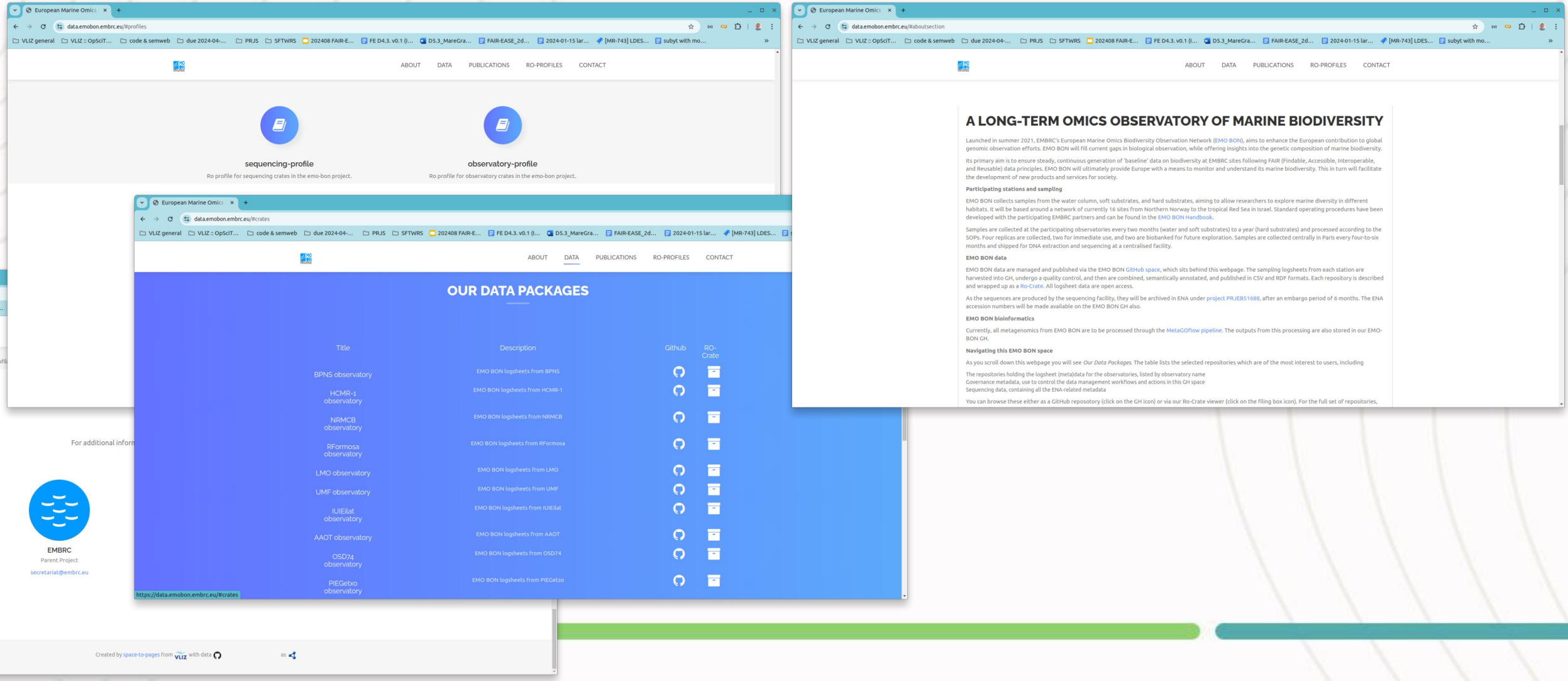
- European Marine Omics Biodiversity Observation Network

  = A long-term omics observatory of marine biodiversity

- Shared procedures & datamanagement
- @FAIR-EASE aka pilot 5 (5.3.1 Marine Omics Observatory)
- The distributed data management for embrc/emo-bon is
  - based on github & RO-Crate
  - targeting a Linked-Open-Data publication (5* semantic web) of the data



data.emobon.embrc.eu/** (LOD)



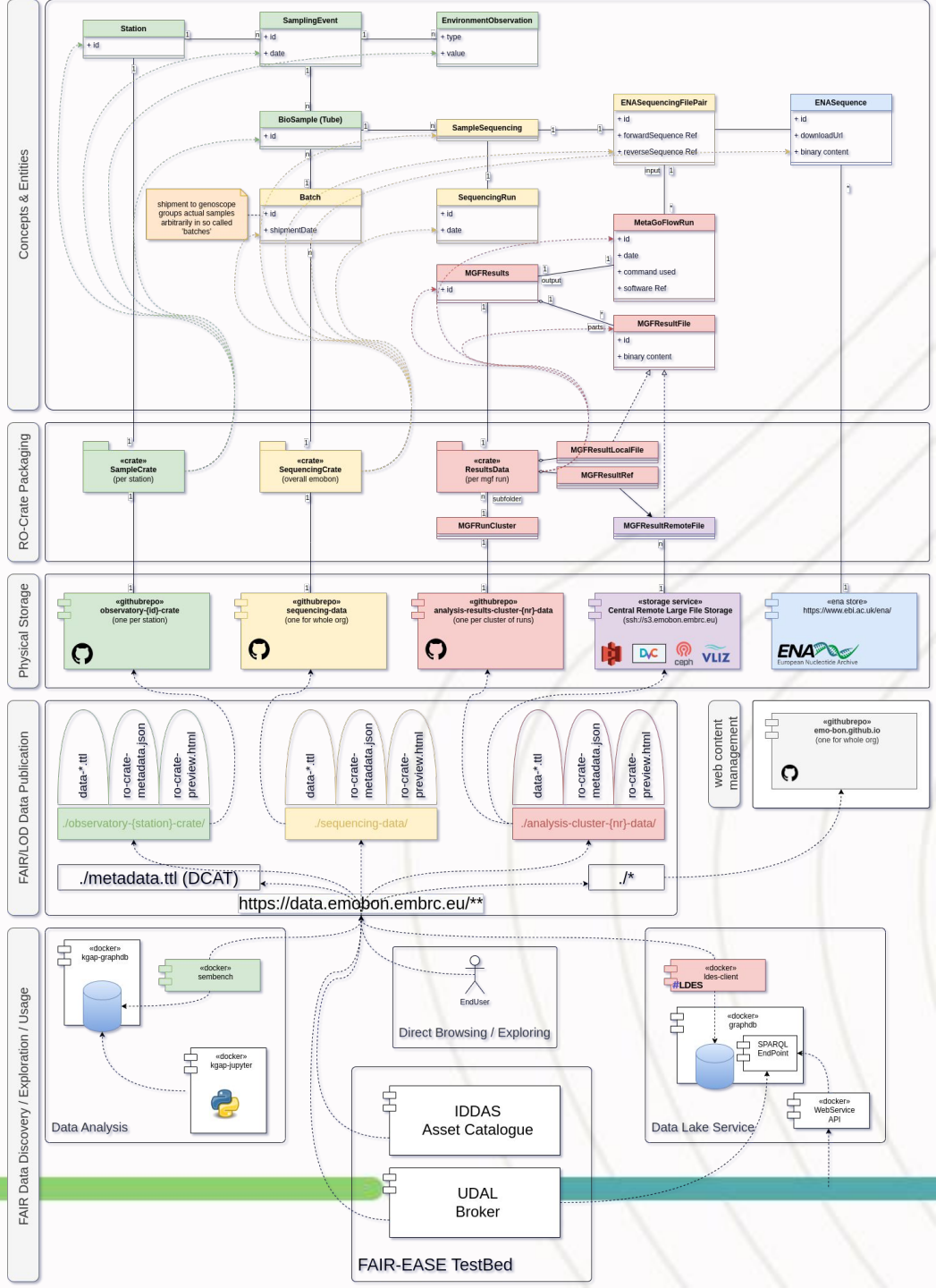
The full picture :

Agenda / Ambition - explain this image

- The model / flow behind this BON
- Where and how data is managed
- How the FAIR / LOD publication of this data is achieved
- How that enables open usage and consumption

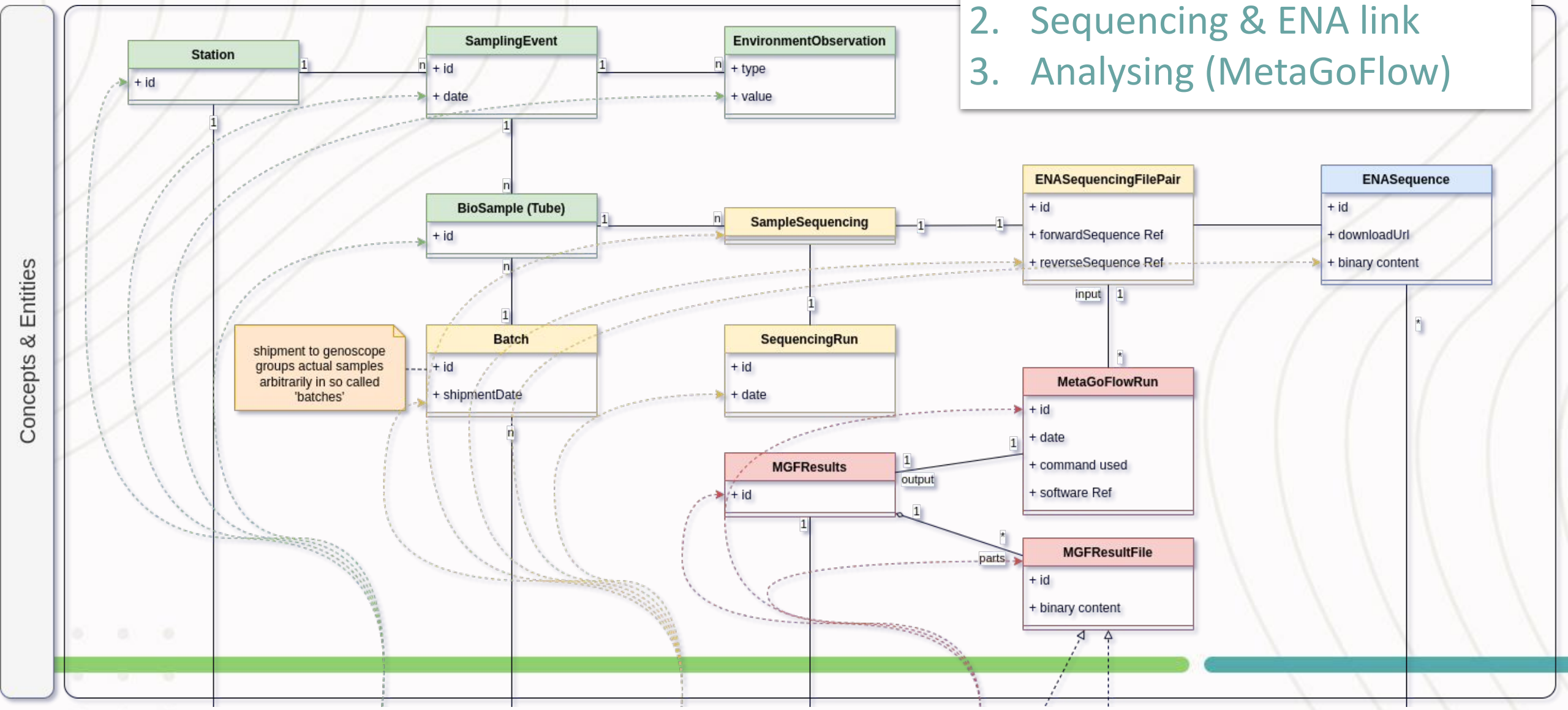
Note:

- Focus on Plan & Vision
- Implementation at ~60%



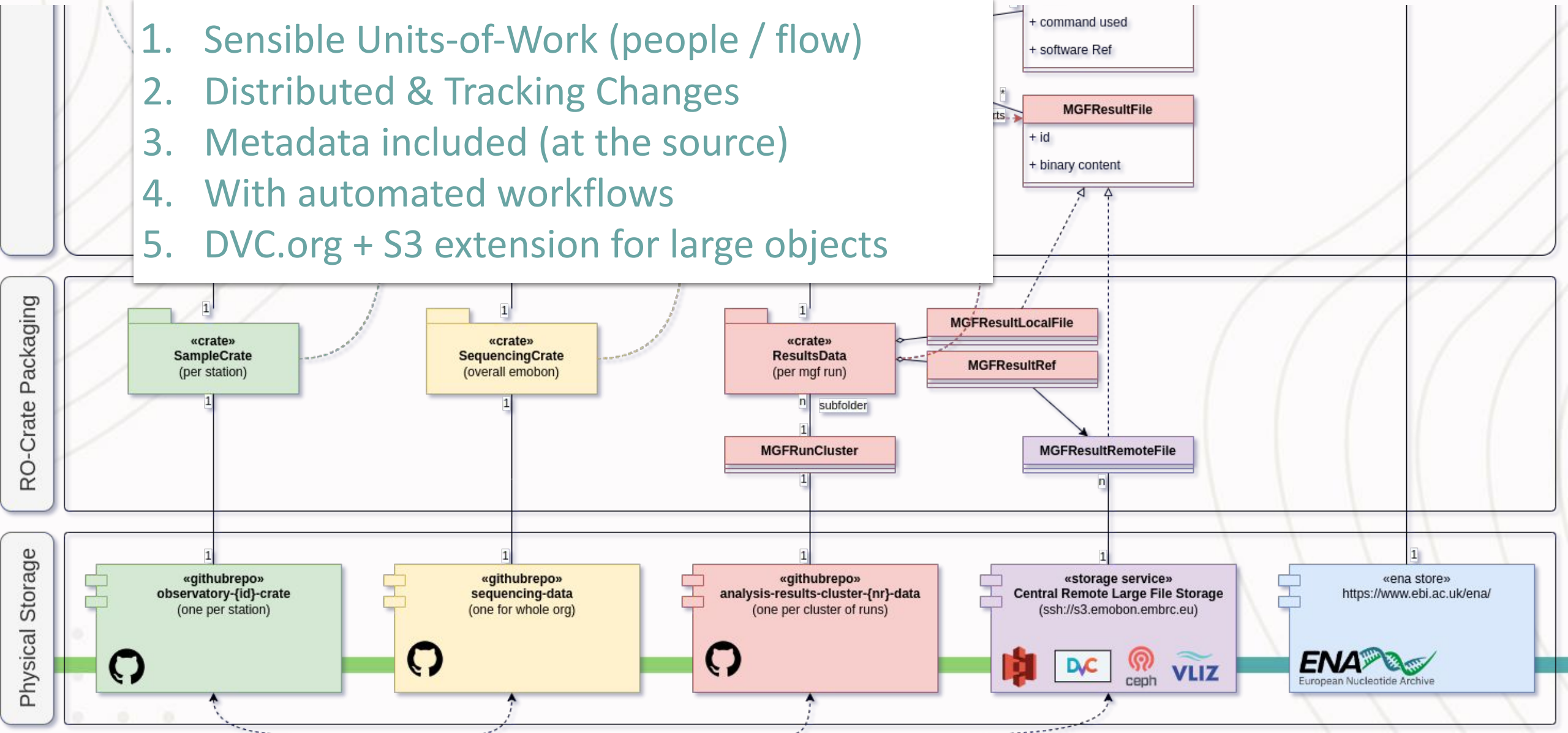
The model & Flow

1. Sampling (field & biobank)
2. Sequencing & ENA link
3. Analysing (MetaGoFlow)



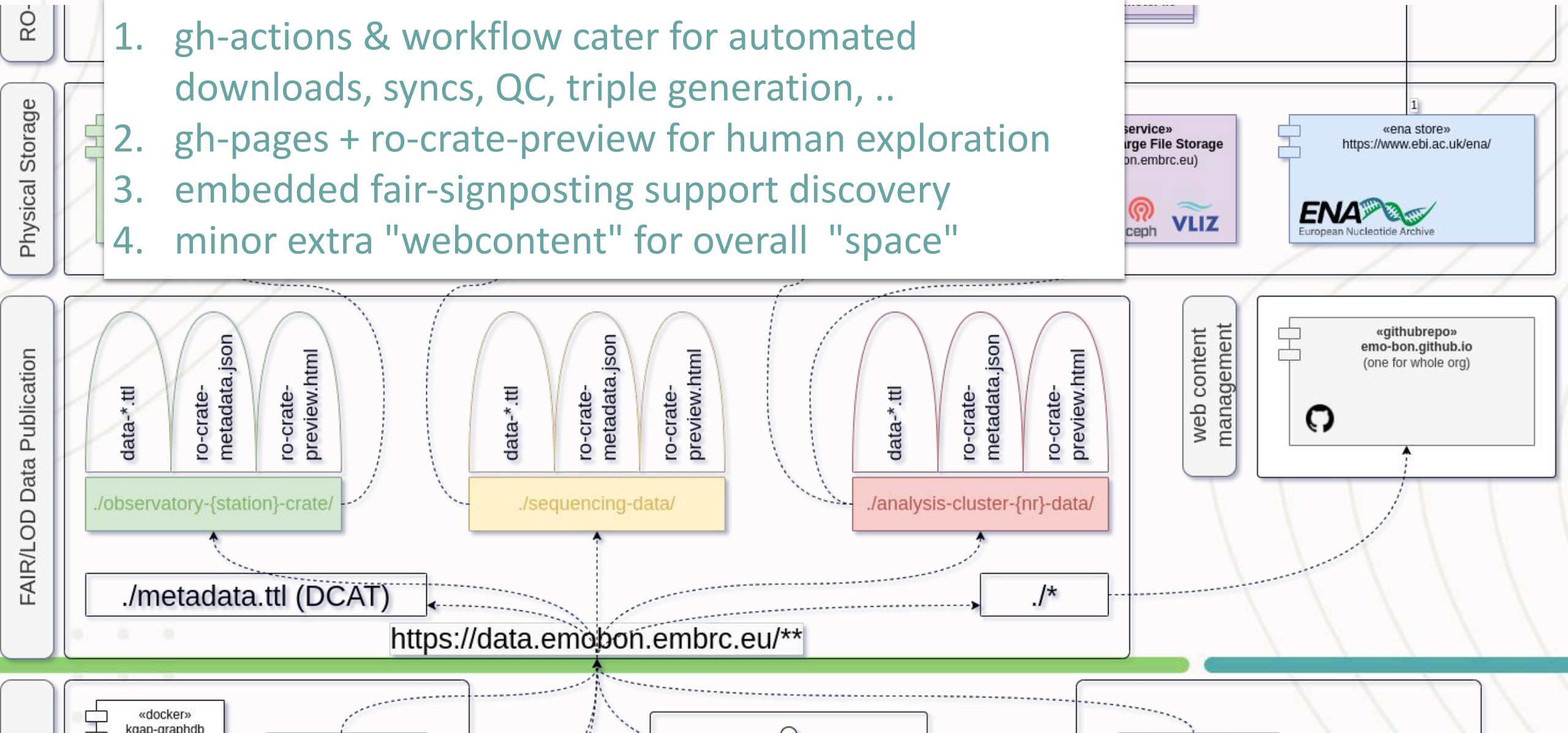
Data Management & Storage

1. Sensible Units-of-Work (people / flow)
2. Distributed & Tracking Changes
3. Metadata included (at the source)
4. With automated workflows
5. DVC.org + S3 extension for large objects

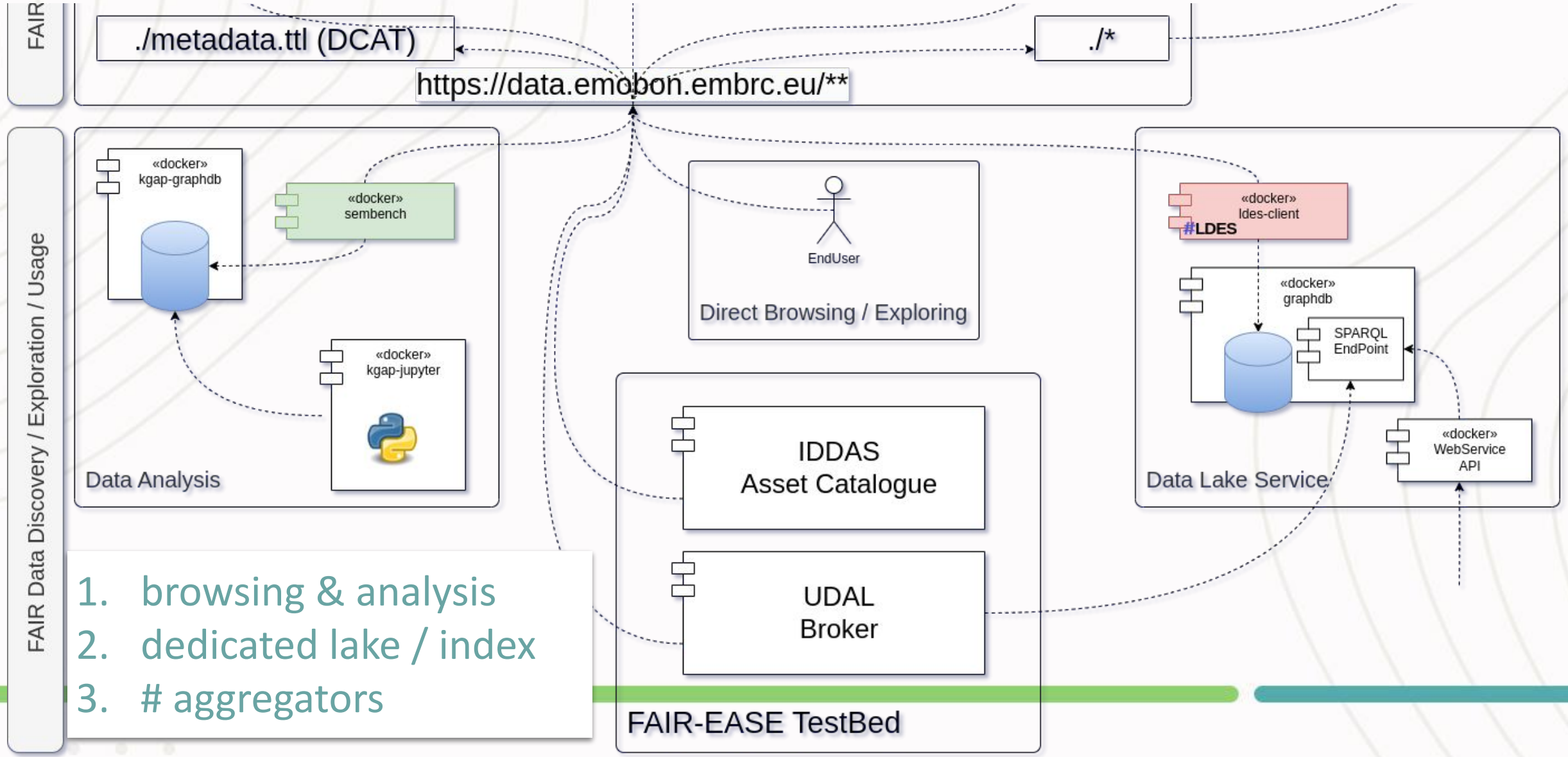


FAIR & Linked Open Data Publication

1. gh-actions & workflow cater for automated downloads, syncs, QC, triple generation, ..
2. gh-pages + ro-crate-preview for human exploration
3. embedded fair-signposting support discovery
4. minor extra "webcontent" for overall "space"



Open & Interoperable Data Reuse

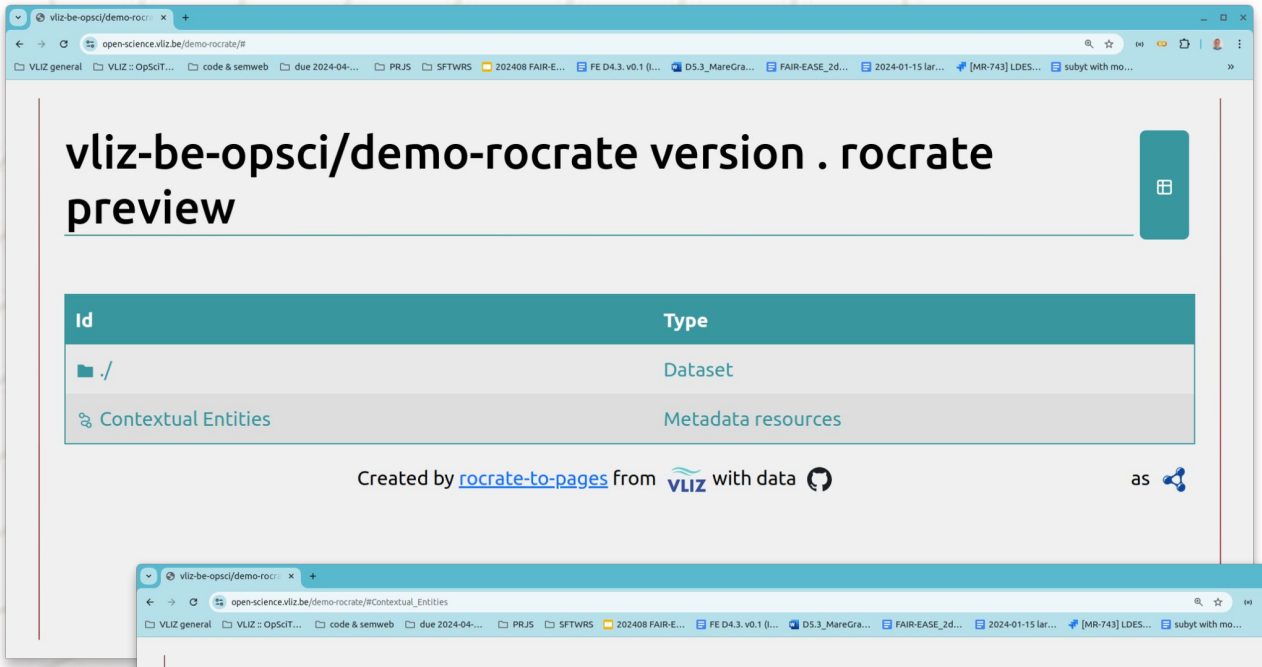


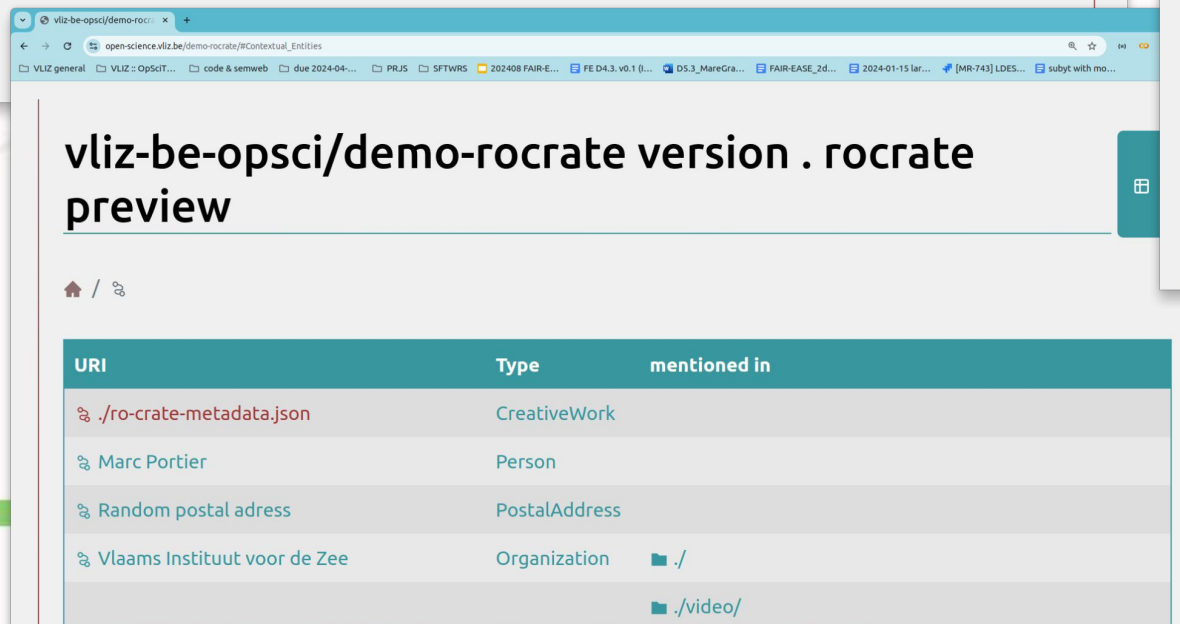
1. browsing & analysis
2. dedicated lake / index
3. # aggregators

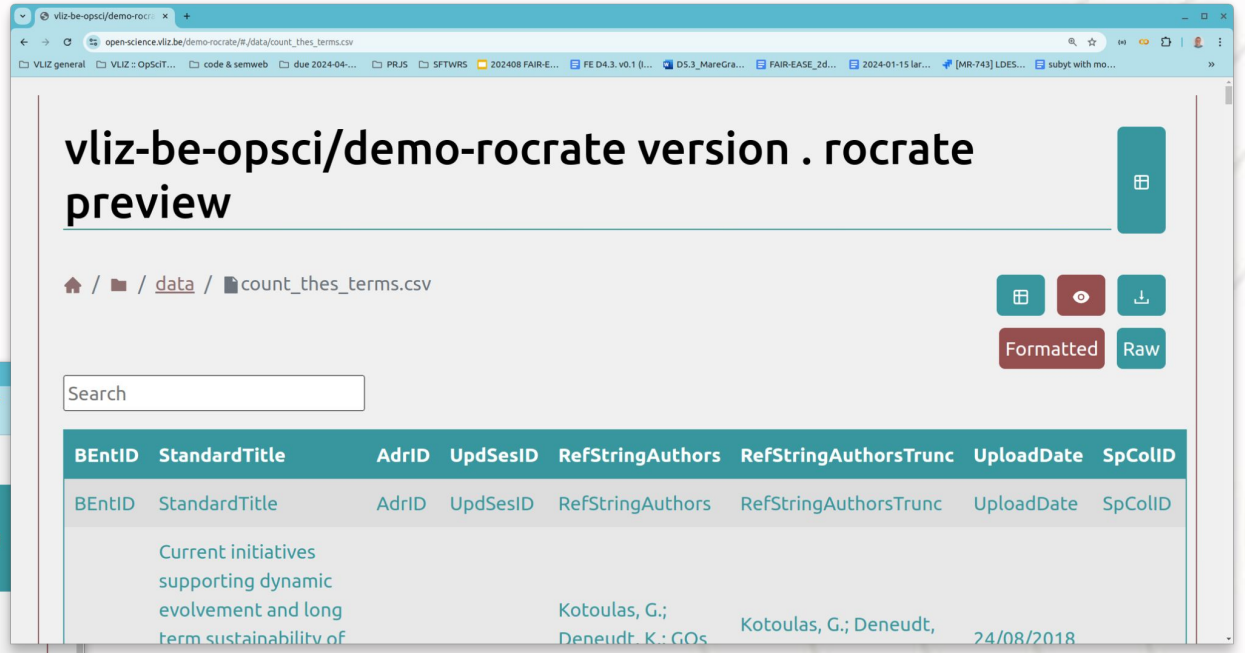
«RO-Profiles»

- RO-Profiles
  - easy plugable external rules to RO-Crates
  - simple identifier ⟶ of additional expectations
- Hook for tools:
  - assisted entry & templating (inside authoring tool)
  - Processing
  - validating, QC reports, …
- Step towards Machine-Actionable DMPs
- Allowing to govern some cross-dataset “conformity”

ro-crate-preview.html (demo)







Github-Actions and Github Pages

- https://github.com/vliz-be-opsci/rocrate-to-pages «github action»
- Converts
  - From «git repo» (an authored basic ro-crate)
  - Into «mini website» (a published ro-crate github.io)

- More GH-Actions at emo-bon adding:
  - Google Docs import
  - ENA Synchronisation and Linking of metadata
  - QC Reports & Issues
  - Semantic Uplifting (triple generation) of some data
  - TO-BE:
    - provenance
    - LD Fragments: change feeds (LDES) and/or indexes

Can you express your knowledge ?

- Living under "Central Repository" Control
  - often first place for "metadata"
  - too late → often only at publication time, not continuously during work
  - too limited → least common denominator
  - lagging on innovation → revisit old records when model is extended

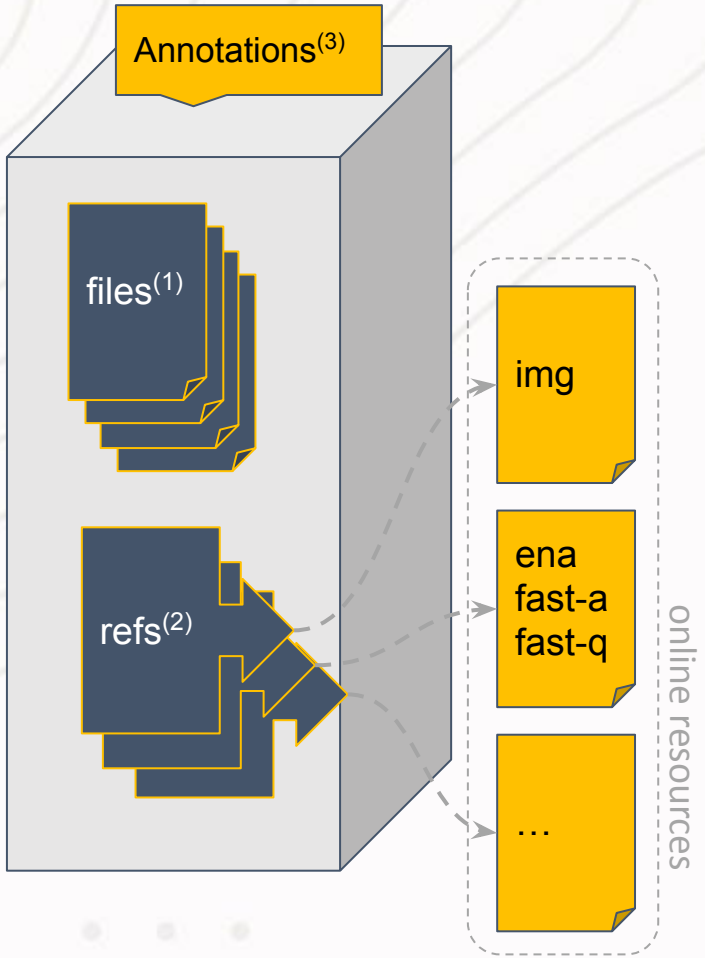
vs.

- «Get it while it is hot» & «Fix it at the Source» (once)
  - "specimen by John" - syndrome
    - lower the cost of capturing valuable side-wise info (e.g. provenance)
  - allow the expert to express and extend
  - repository can still harvest and decide on what to search (index)
  - reverse control → repositories to adapt under growing innovation

eosc | FAIR-EASE

Semantic Public Datasets

RO-Crates as FAIR Digital Objects (FDO)



- (meta)data included with semantic annotation

`./ro-crate-metadata.json(ld)`

- Self-contained «package»
  - (1) all parts (files, any format)
  - (2) all remote parts (online references) «Data Entities»
  - (3) all descriptions (semantics) «Contextual Entities»

- Addressable on the web
  - available as online resources (if needed protected)
  - (self) published like mini-web-sites

eosc | FAIR-EASE

Stepping up Simple → Complex

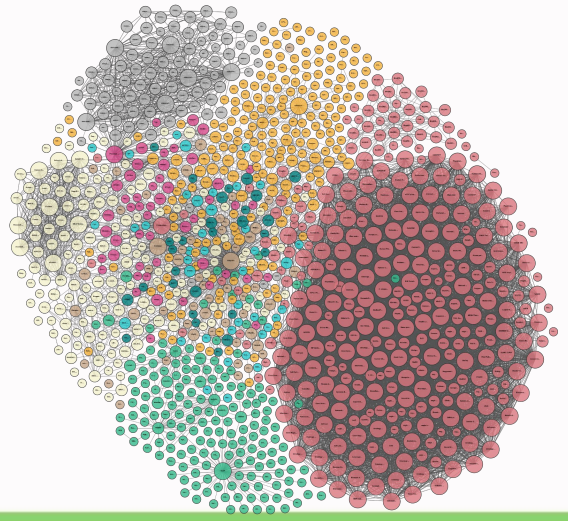
Distribution 🤝 Uniformisation

one simple block

whatever you imagined

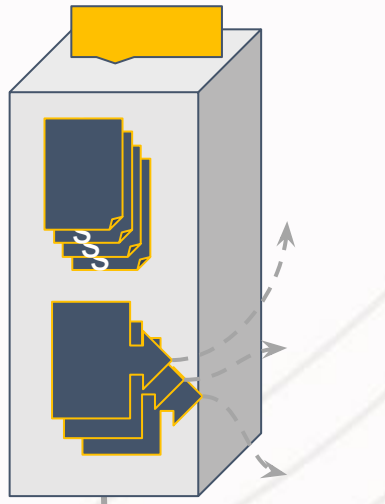
one individual website

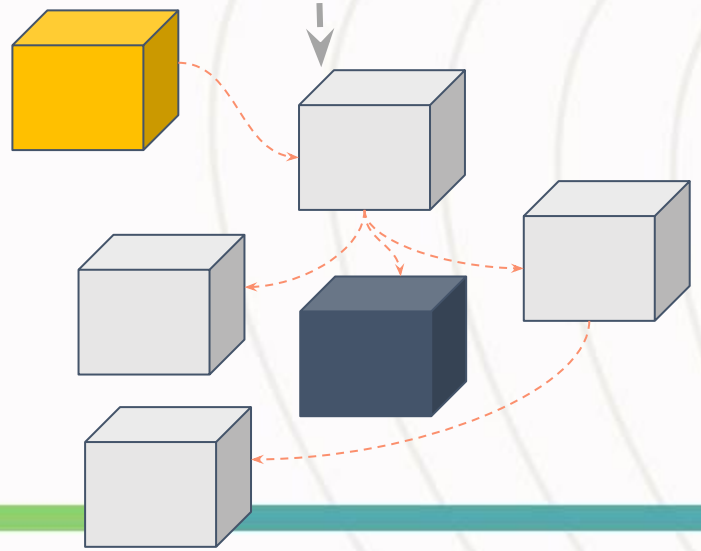
`html | url | http`



the global knowledge graph

one data crate





a cloud of interlinked datasets

On having an «open» Style

- Known Hallmarks of Good (Standards) Design
  - Make Simple Things Easy, Make Hard Things Possible
  - Have a lots of Spaces Within
  - Atomic (do one thing well)
  - Collaborative (play along with others)
- e.g. web-standards → HTML, URL, HTTP, RDF
- RO-Crate design principles…
  - conform to these best practices
  - healthy community that reflects carefully on these aspects

eosc | FAIR-EASE
Building Interoperable Earth Science & Environmental Services

Open Q&A

Moderators:
Rob Carrillo & Gael Lymer

FAIR-EASE has received funding from the European Union's Horizon Europe Framework Programme (HORIZON) - under grant agreement No. 101058785.

eosc | FAIR-EASE

Building Interoperable Earth Science & Environmental Services

Thank you!