http://tinyurl.com/osf-eosc-datacat

http://tinyurl.com/osf-eosc-datacat-materials

How FAIR Friendly is your data catalogue? Exposing FAIR data in EOSC

Organisers

Rafael Jimenez - ELIXIR-Hub
Donatella Castelli - CNR IT
Natalia Manola - Athena Research & Innovation Centre GR
Carole Goble - The University of Manchester UK

Duration

3.5 hours

Location

Proposed relocated to room 113 (swapped with Book Castle 2nd Floor)

Abstract

Research communities and specially research infrastructures are making a concerted effort to homogenize, collect their (meta)data and publish them in the open through community specific data catalogues. This is a good start towards making data FAIR, but how can we ensure availability of domain specific FAIR data and data-analysis services through a common virtual research environment like the European Open Science Cloud (EOSC)? From vertical domains (e.g., research infrastructures) to horizontal approaches (e.g., OpenAIRE, DataCite) which cover national settings and libraries/repositories, we see different content, data models, interfaces, frameworks, architectures and vocabularies being used. The EOSCpilot data interoperability task aims to establish principles, propose recommendations and demonstrate how FAIR data hosted by domain specific data repositories can be exposed to EOSC to be used and reused by EOSC services and users.

This workshop will build upon the work planned by the EOSCpilot data interoperability task and the BlueBridge workshop held on April 3 at the RDA meeting. We will investigate common mechanisms for interoperation of data catalogues that preserve established community

standards, norms and resources, while simplifying the process of being/becoming FAIR. Can we have a simple interoperability architecture based on a common set of metadata types? What are the minimum metadata requirements to expose FAIR data to EOSC services and EOSC users?

Target audience

Research Infrastructures, e-Infrastructures, libraries, policy makers, funders, researchers.

Agenda

If you are a presenter please add your presentation to the shared google drive folder and link the title of your talk to your presentation.

Session 1: 120 minutes - landscape talks and discussion

Welcome		
2 min	• Agenda and purpose	Carole Goble

EOSC and data interoperability - 30 minutes		
10 min	• EOSC and EOSCpilot	Brian Matthews STFC
10 min	• EOSCpilot data interoperability	Rafael C Jimenez ELIXIR
10 min	• Questions and answers The distinctions between data catalogues and metadata catalogues are blurred Data and metadata quality - example re3Data Validation of metadata	All
Data catalogues - 50 minutes		
6 min	• BlueBridge data catalogue	Donatella Castelli CNR

http://tinyurl.com/osf-eosc-datacat

http://tinyurl.com/osf-eosc-datacat-materials

6 min	• EPOS data catalogue	Keith Jeffery
6 min	• OMICsDI, a data catalogue for life sciences	Rafael C Jimenez on behalf of Yasset Perez Riverol EMBL-EBI
6 min	• OpenAIRE, a horizontal data catalogue	Paolo Manghi CNR
6 min	• FAIRDOMHub, an asset catalogue for systems and synthetic biology projects	Carole Goble The University of Manchester / FAIRDOM Association e.V.
6 min	• FAIRSharing, a metadata catalogue	Rafael C Jimenez on behalf of Susanna A Sansone University of Oxford
6 min	• Bioschemas: schema.org for life sciences	Carole Goble The University of Manchester / ELIXIR-UK
6 min	• SeaDataNet CDI Catalogue Service	Athanasia (Sissy) Iona /HNODC, Greece)
8 min	• Questions and answers Users want to see what is relevant to them not everything Users want to see metadata across disciplines with different metadata Datacite CERIF RDA minimal sets	All

Discussion		
30 min		Carole Goble

http://tinyurl.com/osf-eosc-datacat

http://tinyurl.com/osf-eosc-datacat-materials

Session 2: 90 minutes

Welcome and settling revisited		
5 min	• Agenda and purpose reminder	Carole Goble

Reviews and recommendations 20 minutes		
5 min	• Outcomes and conclusion of the Data catalogues workshop organised by BlueBRIDGE at the RDA meeting	Donatella Castelli CNR
8 min	• Survey results "Towards a common EOSC catalogue"	Massimiliano Assante CNR
2 min	• Survey results "Towards a common EOSC catalogue": a matrix perspective	Rafael C Jimenez ELIXIR
5 min	• Questions and answers	All

Group discussion 70 minutes	
We have preliminarily identified 13 principles grouped into 3 areas. These will drive the work of the EOSCpilot data interoperability working group and the recommendations we will proposed for EOSC. We will divide into breakout groups and discuss for each group • Which are the priority principles? • What are the top 3 obstacles? • What are the top 3 items to put on the roadmap to address them?	
20 min	**Reuse in e-Infras: Leverage the rich legacy of Research Infrastructures** • Making data FAIR is the responsibility of the Research Infrastructures and their data repositories • We must rely on research infrastructure data catalogues

http://tinyurl.com/osf-eosc-datacat

http://tinyurl.com/osf-eosc-datacat-materials

	• We must support an ecosystem of catalogues • We should provide metadata quality recommendations to feedback to RIs
20 min	**Least: The least possible metadata for the most benefit** • Findability should come first • Common and minimum metadata • Focus on common data types: datasets and data repositories • Flexible metadata models to embrace domain specifics • Service requirements and operational metadata first class citizen
20 min	**Practical: Sustainable and pragmatic delivery** • Engage EOSC demonstrator data repositories • Propose methods to expose metadata • Simple to implement, easy to sustain • Deliver guidelines and demonstrators
10 min	Wrap up

Speakers

- Donatella Castelli (CNR, IT)
- Rafael C Jimenez (ELIXIR Hub)
- Paolo Manghi (CNR, IT)
- Carole Goble (The University of Manchester, UK / FAIRDOM Association e.V. / ELIXIR-UK)
- Brian Matthews (STFC, UK)
- Keith Jefferrey (ERCIM)
- Massimiliano Assante (CNR, IT)

Workshop living document

Feel free to contribute with notes in this section

Outcome of Q&A discussion of the morning

Find is really the key driver
EOSC wants domain agnostic metadata
Context, as you soon as you become generic you lose context
Associating training with accessibility of data

http://tinyurl.com/osf-eosc-datacat

http://tinyurl.com/osf-eosc-datacat-materials

Hands-on session

After discussion within the EOSCpilot data interoperability working group and mainly based on community feedback collected from EOSCpilot workshops like the BlueBridge workshop and via surveys we have defined the scope of our tasks based on guiding principles mentioned below. We would like feedback to evaluate these principles which will drive 1.- the work of this working group and 2.- the recommendations we will propose to EOSC.

Guiding principles

Reuse: Leverage the rich legacy of Research Infrastructures

Making data FAIR is the responsibility of the Research Infrastructures and their data repositories

The role of the EOSCpilot data interoperability working group is not to define how to make data FAIR but to define and demonstrate a simple data interoperability architecture to expose FAIR data to EOSC services and EOSC users. We believe the responsibility of defining how to make data FAIR lies on Research Infrastructures and specially their participant data repositories. Moreover there is already a working group funded by the European Commision which started in parallel to define a roadmap to make data FAIR across data repositories.

We must rely on research infrastructure data catalogues

Many data repositories exist per scientific domain. Domain specific research infrastructures maintain data catalogues which collect, integrate, harmonise and enrich metadata from many dispersed and diverse data repositories to facilitate data discovery. We plan to rely on existing metadata catalogues as a main providers of scientific metadata for EOSC. We expect domain specific data catalogues will collect metadata from relevant data repositories.

We must support an ecosystem of catalogues

We believe in an ecosystem of coordinated data catalogues where domain specific data catalogues collect specific metadata from data repositories and generic data catalogues collect a subset of metadata from domain specific data catalogues. Ideally the generic data catalogues should pull information from specific data catalogues and recommend metadata submission to domain specific catalogues.

We should provide quality recommendations to feedback to RIs

With the analysis of data catalogues, metadata models and standards we aim to provide recommendations about how to improve the quality of the metadata provided by data catalogues and data repositories.

We should recommend associating competency levels associated with data use

It's not straightforward to work with data from specialised contexts, so would need to augme~~t~~nt~~n~~ data with appropriate training and understanding of the data. POinters on how to understand the data rather than specialist domain data.

http://tinyurl.com/osf-eosc-datacat

http://tinyurl.com/osf-eosc-datacat-materials

Least: The least possible metadata for the most benefit

Findability should come first

Findability is the first step to make data FAIR and the main condition to access and reuse data. We will focus on how EOSC services and EOSC users can find data taking into account their access, interoperability and reusability requirements.

Common and minimum metadata

We do not aim to create a new data model to describe datasets or data repositories but create a recommendation of minimum metadata properties common across data catalogues. Properties that will help EOSC services and users to find data repositories and datasets and will facilitate data access, interoperability and reusability. We aim to evaluate existing data models and recommend how to expose data reusing one or several data models.

Focus on common data types: datasets and data repositories

We will focus our work on few data types which are common across different scientific disciplines to start with. These data types are datasets and data repositories.

Flexible metadata models to embrace domain specifics

Each scientific domain work with standards to define their specific scientific entities which might or not be described with a standard format. We want to respect the existing formats and let research infrastructures and scientific communities decide on how better describe their data. We are looking for a set of minimum properties among models but we should be flexible enough to allow space for custom or domain specific properties.

Service requirements and operational metadata first class citizen

It is about the scientific metadata but also importantly about the operational metadata required by services to be able to find, access and use the data.

Practical: Sustainable and pragmatic delivery

Engage EOSC demonstrator data repositories

Most of the EOSC demonstrators involved datasets at least from one data repository. We will engage these data repositories to demo how their datasets can be discovered and accessed in EOSC via data catalogues.

Propose methods to expose metadata

We will evaluate existing methods and guidelines to expose metadata and propose one or more technologies to expose in data catalogues minimum and common properties. We will rely on work done by initiatives like RDA and GO-FAIR as well as the expertise of our data catalogues.

Simple to implement, easy to sustain

Any proposed solution should be looking at a high impact low effort strategy specially in the short term. It should be simple to implement and easy to maintain providing just enough functionality to facilitate discovery, access and use of data in the EOSC.

http://tinyurl.com/osf-eosc-datacat

http://tinyurl.com/osf-eosc-datacat-materials

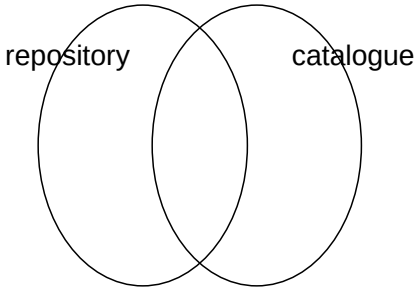
Deliver guidelines and demonstrators

The outcome of our work will be a report but also a set of guidelines, an architecture proposal and demonstrators applying our recommendations and showing the feasibility of our proposed strategy to make FAIR data findable, accessible and reusable in EOSC.

Group 2

People

- Pronk Tessa
- Carole Goble
- Rafa Jimenez
- Marta Hoffman-Sommer
- Hienola
- Ignasi Labastida
- Athanasia Iona
- Jan Wiebelitz
- Remedios Melero, CSIC

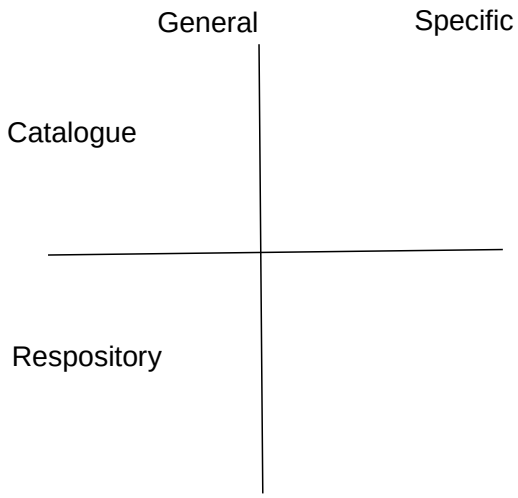


-
-
-

Formatted: Don't add space between paragraphs of the same style, Line spacing: Multiple 1,15 li, Outline numbered + Level: 1 + Numbering Style: Bullet + Aligned at: 0,63 cm + Indent at: 1,27 cm

http://tinyurl.com/osf-eosc-datacat

http://tinyurl.com/osf-eosc-datacat-materials



Notes

- Balance between mandating and being flexible
- EOSC should expect the repositories to have to make some changes
- Should have to set expectations of quality
- EOSC incentives
- Validator for the minimum metadata properties *for EOSC* not for the community
- MoSCoW in the metadata
- Has to be tensioned against the intents of the data catalogue - so what are these as a capability tiers like the DANS FAIR stars system?
- What is the metadata that helps people find things and for domain agnostic services to use it
- Understanding the intentions

Single repository - is that a catalogue?
A catalogue at the national or institutional level
Index vs Store -> see venn diagram above

FOR the PILOT

Questions.

- Which are the priority principles?
- What are the top 3 obstacles?
- What are the top 3 items to put on the roadmap to address them?

Reuse: Leverage the rich legacy of Research Infrastructures

- Making data FAIR is the responsibility of the Research Infrastructures and their data repositories

    - Agree.
        - Other groups working on this topic. We should rely on their work. Ie. RDA, GO-FAIR, ELIXIR, … No just RI, aslo e-infrastructure.
- We must rely on research infrastructure data catalogues
    - Agree
        - Amplification impact effect and “point of invention” impact
        - Datacite a catalogue for EOSC?
        - Zenodo is a EOSC data repository
        - Need to define relationship between generic and domain specific data catalogues and data repositories
        - Is it a quadrant
        - Perfect is the enemy of the good
        - Layered apps - a market for deduplication ad
        - Push and pull
        - Sensitive data in the data
        - Terms of use. And also licenses and beacons

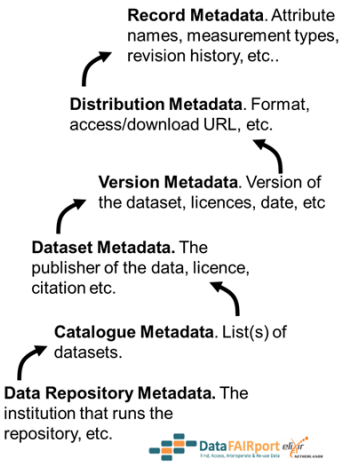
- We must support an ecosystem of catalogues
    - Agree
- We should provide metadata quality recommendations to feedback to RIs
    - Agree

Least: The least possible metadata for the most benefit

- Findability should come first
    - IR is much harder! Many RIs and organisations are addressing these. It is the responsibility of the providers and consumers.
    - Findability that is discriminatory to be useful, how do you rank and filter (Tessa)
        - Existence finding vs corpus finding (Marta)
        - Finding services
    - How to Access and How to reuse it is mandatory
    - Text mining finding use case
    - Drill down metadata down to the original data
- Common and minimum metadata
- Focus on common data types: datasets and data repositories
    - What is a dataset
    - What do we do about hybrid data repositories/catalogues?
    - Data FAIRPoint model ? Bioschemas.org
- Flexible metadata models to embrace domain specifics
- Service requirements and operational metadata first class citizen
    -

http://tinyurl.com/osf-eosc-datacat

http://tinyurl.com/osf-eosc-datacat-materials



Practical: Sustainable and pragmatic delivery

- Engage EOSC demonstrator data repositories
- Propose methods to expose metadata
- Simple to implement, easy to sustain
- Deliver guidelines and demonstrators

Group 1

People

- Donatella Castelli
- Keith Jefferrey
- Brian Matthews
- Penny Labopolou
- Massimiliano Assante
- ELLI Papadoupulo

Reuse: Leverage the rich legacy of Research Infrastructures

- Making data FAIR is the responsibility of the Research Infrastructures and their data repositories
- We must rely on research infrastructure data catalogues
- We must support an ecosystem of catalogues
- We should provide metadata quality recommendations to feedback to RIs

http://tinyurl.com/osf-eosc-datacat

http://tinyurl.com/osf-eosc-datacat-materials

Merging DCs leads to uncontrolled duplication, this is going to be an issue, when we will design the DC take into account the problem of de-duplication. OPENAire already has a deduplication service, so perhaps EOSC could re-use existing infrastructure service

Each RI is a EOSC in miniature.

Not a catalogue of catalogues, we should treat everything as first class objects. (Multi Entry to datasets via different catalogues and search (why a catalogue at all for finding :-))

It is true the we should rely on RIs but EOSC is sth different for example there is all these researchers that work across domains. The content of the catalogue can be reused, are the solutions implemented in each RI can be reused as they are?

EOSC has to provide a mechanism to make understand people the specific metadata, e.g. ask EPOS about this.

There is an RDA WG on FAIR Data. Metadata Interest Group https://www.rd-alliance.org/groups/metadata-ig.html Data Discovery Paradigms Interest Group https://www.rd-alliance.org/groups/data-discovery-paradigms-ig

Least: The least possible metadata for the most benefit

- Findability should come first
- Common and minimum metadata
- Focus on common data types: datasets and data repositories
- Flexible metadata models to embrace domain specifics
- Service requirements and operational metadata first class citizen

You need to collect the metadata from your data. This is easier if you have context.
Findability come with a richer metadata, perhaps controlled vocabulary is the best solution for FIndability.

There is a need to have the domain of the data, domain is crucial for users. So you could put the major domain in EOSC and have the RIs to side in subdomains. (This should be an option)

Perhaps, which metadata element helps what? And then you prioritize them.

When you try to find something by profiling the user you could rank higher the hits from her domain.

About license, it should be considered very well.

Practical: Sustainable and pragmatic delivery

- Engage EOSC demonstrator data repositories
- Propose methods to expose metadata
- Simple to implement, easy to sustain
- Deliver guidelines and demonstrators

http://tinyurl.com/osf-eosc-datacat

http://tinyurl.com/osf-eosc-datacat-materials

The first one (EOSC demonstrator) is happening isn't it?

Methods to expose metadata, exposing the metadata to a citizen scientist is different from exposing it to a domain scientist. Supposing that you have the broad domain you could expose the common metadata along with the domain specific one. Dynamicity is important

SImple to implement easy to sustain, generate the metadata as much as automatic as possible, validate them as much as automatic as possible. In this case deep learning algorithms may be of help.
The other thing that EOSC could do is let people think that metadata are not just library catalogue cards, not flat metadata.

We need a paradigm shift

- Pressing on with arguments of Fame, Money and Love do not work
- Instead focus on "by side effect" using modern infrastructure and knowledge practices - deep learning, search, crowdsourcing, social media, pay-as-you-go metadata accumulation
- Expose metadata to be consumed by finding services, not just catalogue

Metadata between communities often hits mismatches and different interpretations of scale.