Harvesting by EOSC repos, OAI-PMH, extended schemas

Gareth Murphy

European Spallation Source

Harvesting vs authenticated search

EOSCpilot
The European Open Science Cloud for Research Pilot Project

- For EOSC/OpenAIRE/ B2FIND, we need to provide all our metadata to be “harvested”. No login, anyone can access, truly open data
- For analysis/WP4 users, can query for their own (embargoed) metadata. Requires login, authentication, securing data and metadata

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
i 🔒 © pi ● §
http://doi.org/10.17616/R31NJMKO
SciCat

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES
i 🔒 © pi ● §
http://doi.org/10.17616/R33H18
ILL Data Portal

panosc

European Open Science Cloud (EOSC)

- Still a work in progress
- We want to provide our data and metadata to EOSC
- We don’t know what EOSC will be like in its final form

OpenAIRE, B2FIND

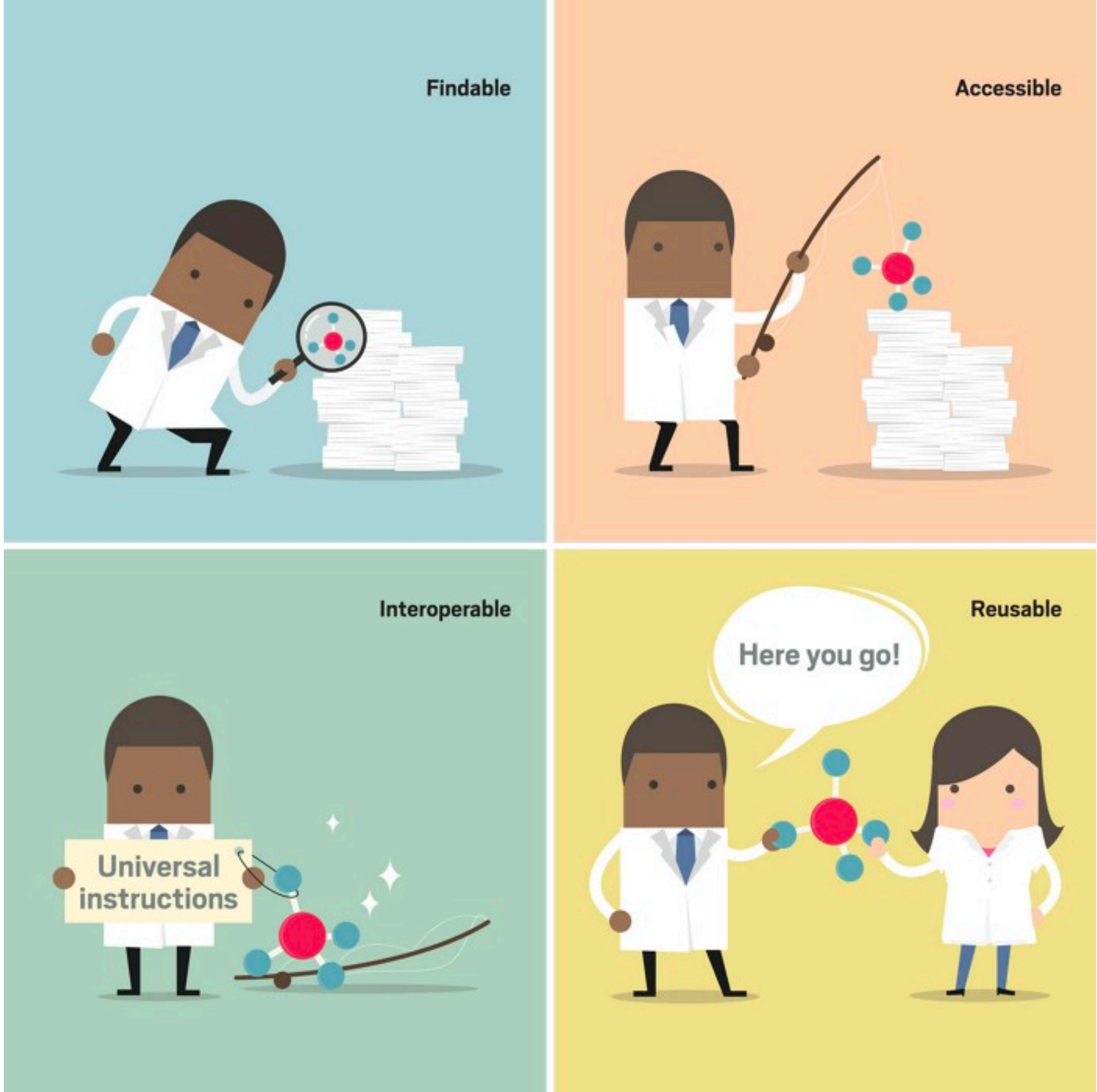
- Open access infrastructure for research in Europe
- We join them and they will connect to EOSC

B2FIND

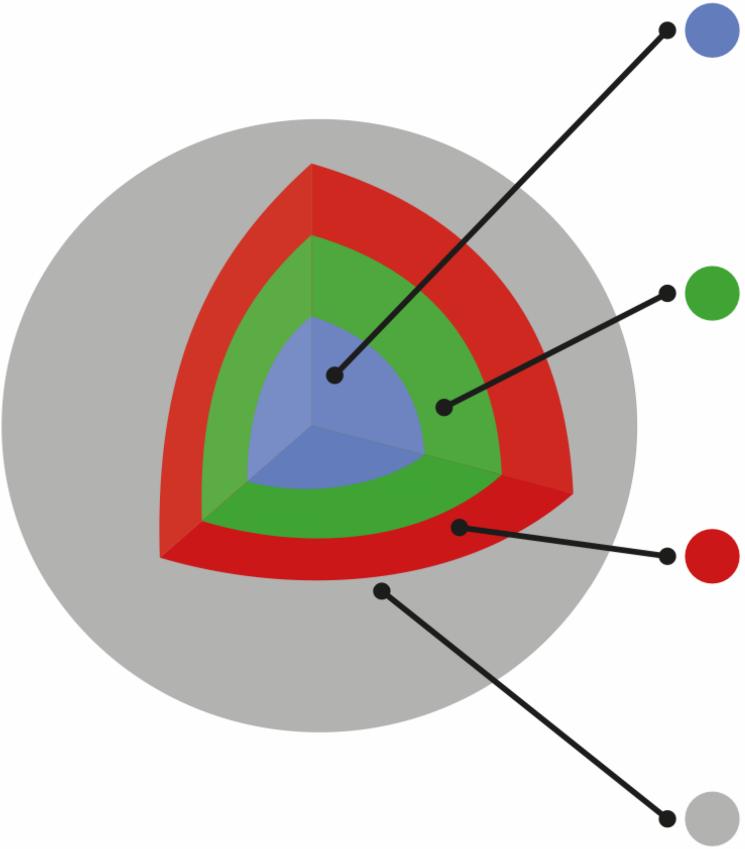
OpenAIRE

JOIN
THE
OpenAIRE
B2FIND
AND
SEE
THE
EOSC
The European Open Science Cloud
Principal Recruiting Offices:
LONDON: Central London Recruiting Depot, Great Scotland Yard, S.W.1

Findable Accessible Interoperable Reusable



FAIR Digital Object



DIGITAL OBJECT

Data, code and other research outputs

At its most basic level, data or code is a bitstream or binary sequence. For this to have meaning and to be FAIR, it needs to be represented in standard formats and be accompanied by Persistent Identifiers (PIDs), metadata and documentation. These layers of meaning enrich the object and enable reuse.

IDENTIFIERS

Persistent and unique (PIDs)

Digital Objects should be assigned a unique and persistent identifier such as a DOI or URN. This enables stable links to the object and supports citation and reuse to be tracked. Identifiers should also be applied to other related concepts such as the data authors (ORCIDs), projects (RAIDs), funders and associated research resources (RRIDs).

doi

ORCID Connecting Research and Researchers

re3data.org

REGISTRY OF RESEARCH DATA REPOSITORIES

http://doi.org/10.17616/R31NJMKO

SciCat

STANDARDS & CODE

Open, documented formats

Digital Objects should be represented in common and ideally open file formats. This enables others to reuse them as the format is in widespread use and software is available to read the files. Open and well-documented formats are easier to preserve. Data also need to be accompanied by the code use to process and analyse the data.

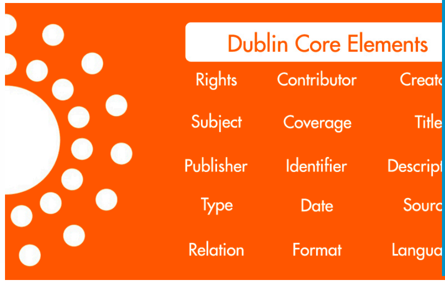
NeXus

HDF

METADATA

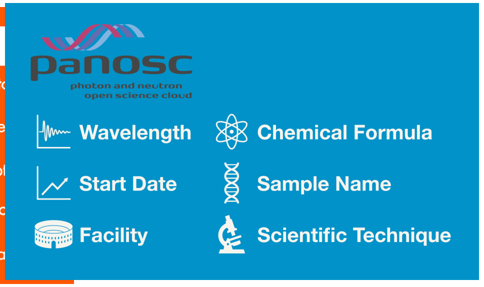
Contextual documentation

In order for Digital Objects to be assessable and reusable, they should be accompanied by sufficient metadata and documentation. Basic metadata will enable data discovery, but much richer information and provenance is required to understand how, why, when and by whom the objects were created. To enable the broadest reuse, they should be accompanied by a plurality of relevant attributes and a clear and accessible usage license.



Dublin Core Elements

Rights	Contributor	Creat
Subject	Coverage	Title
Publisher	Identifier	Descrip
Type	Date	Sourc
Relation	Format	Langua



panosc

photon and neutron open science cloud

Wavelength	Chemical Formula
Start Date	Sample Name
Facility	Scientific Technique

Findable Accessible Interoperable Reusable

according to emerging standards for trustworthiness and FAIR. The overall system and interactions between components and stakeholders are driven by metrics, incentives, investment and skills. In a European context, this FAIR ecosystem should be delivered primarily via the EOSC.

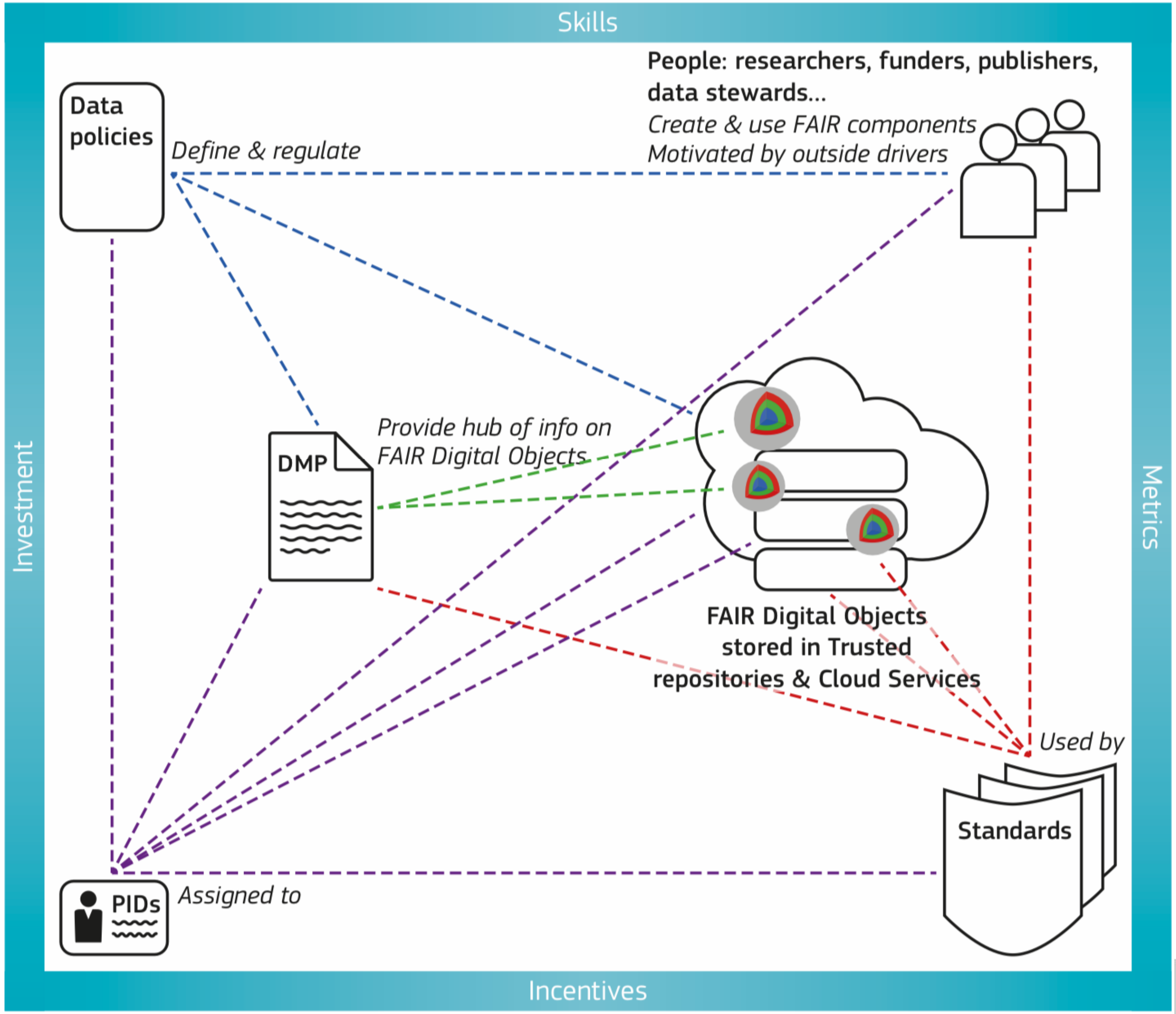

Skills
Investment
Metrics
Incentives
Data policies
Define & regulate
People: researchers, funders, publishers, data stewards...
Create & use FAIR components
Motivated by outside drivers
DMP
Provide hub of info on FAIR Digital Objects
FAIR Digital Objects stored in Trusted repositories & Cloud Services
Used by
Standards
PIDs
Assigned to

Figure 9. The interactions between components in the FAIR data ecosystem. Notes on this figure:

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

http://doi.org/10.17616/R31NJMNL

European XFEL Data Portal

doi

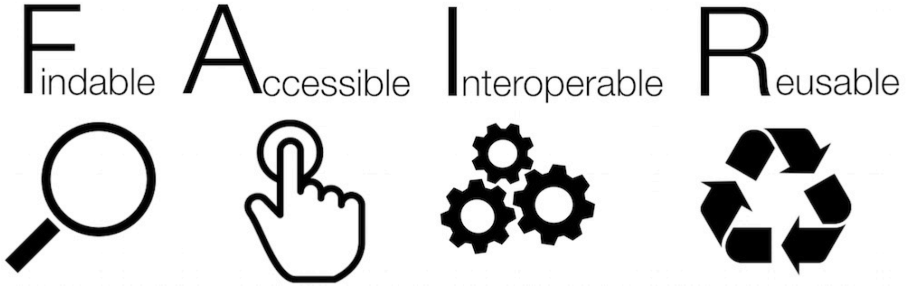
HDF

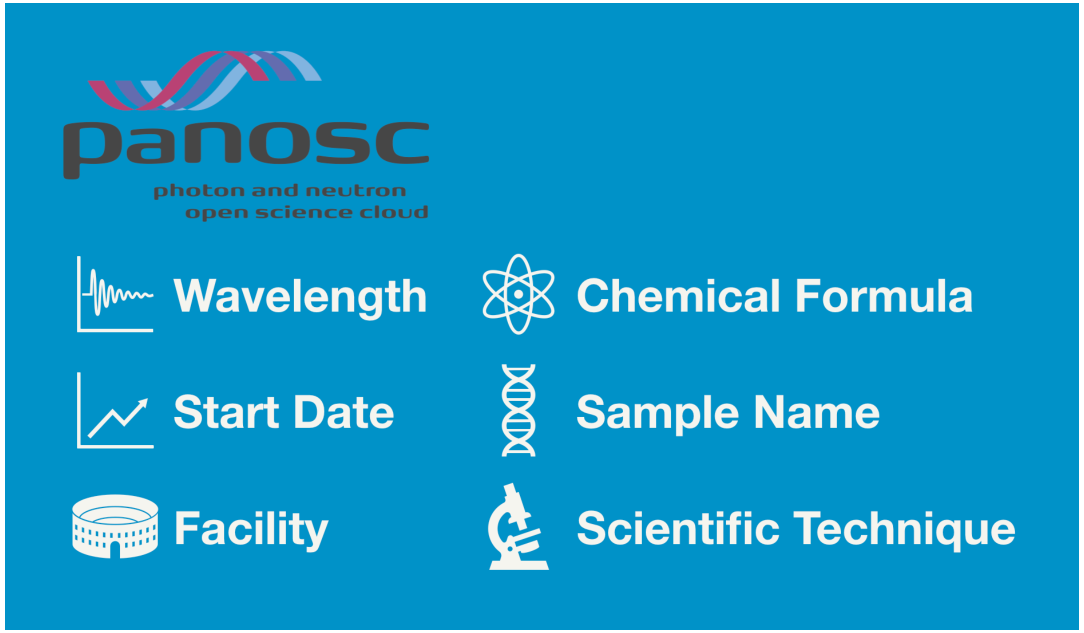
NeXus

panosc

Extending the schema

- Supported formats, (OAI) Dublin Core and PaN format





panosc
photon and neutron
open science cloud

Wavelength

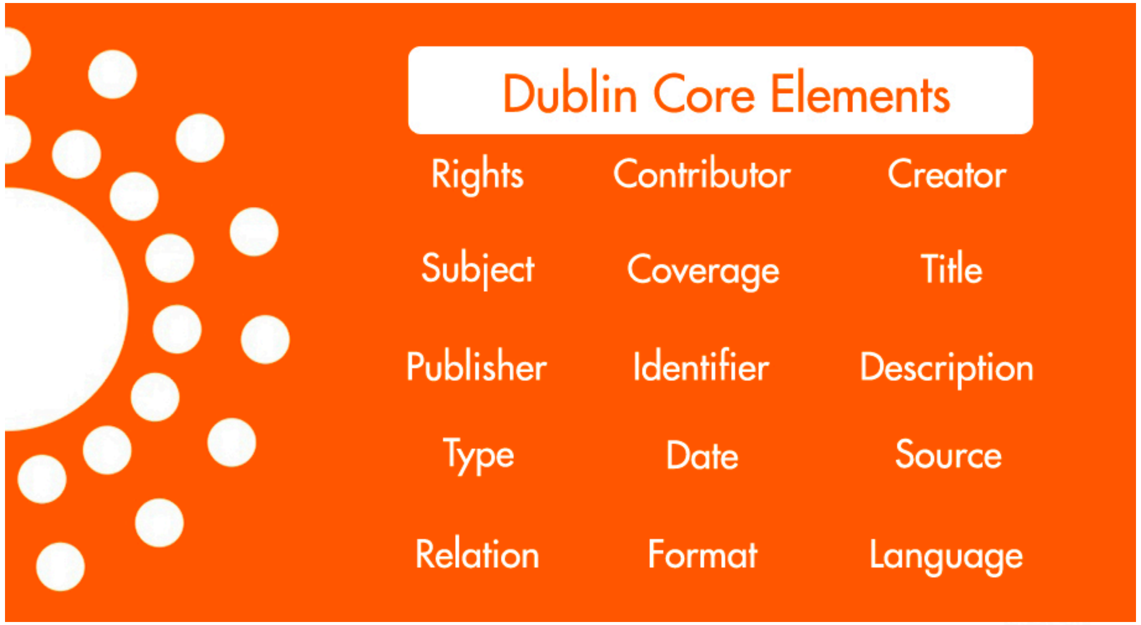
Chemical Formula

Start Date

Sample Name

Facility

Scientific Technique

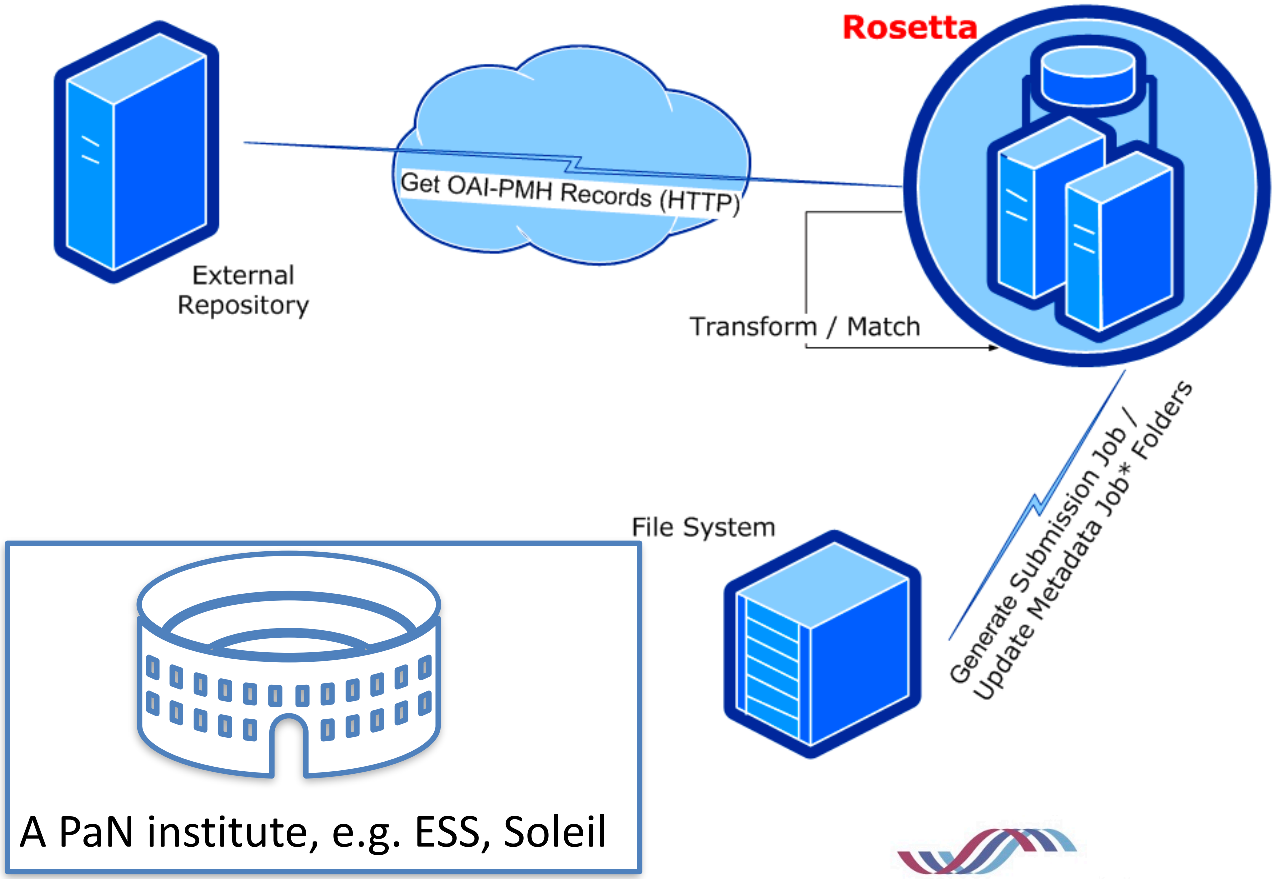


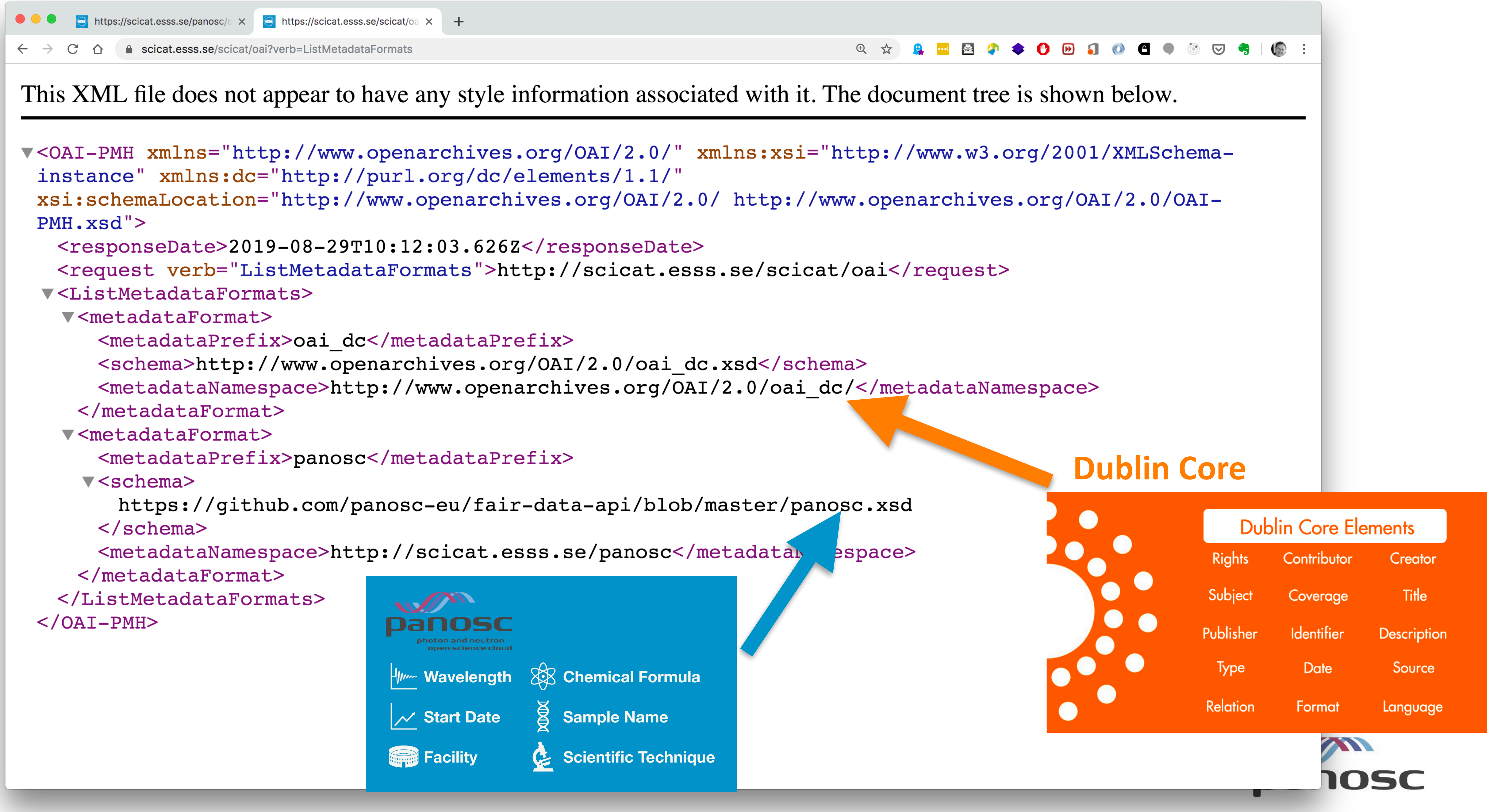
Dublin Core Elements

Rights	Contributor	Creator
Subject	Coverage	Title
Publisher	Identifier	Description
Type	Date	Source
Relation	Format	Language

OAI-PMH

- OAI-PMH (Open Archives Initiative - Protocol for Metadata Harvesting)
- 6 verbs
  - Identify
  - ListMetadataFormats
  - ListRecords
  - GetRecord
  - ListSets
  - ListIdentifiers
- https://www.openarchives.org/pmh/


Rosetta
Get OAI-PMH Records (HTTP)
External Repository
Transform / Match
Generate Submission Job / Update Metadata Job* Folders
File System
A PaN institute, e.g. ESS, Soleil


https://scicat.esss.se/panosc/c
https://scicat.esss.se/scicat/oa
scicat.esss.se/scicat/oai?verb=ListMetadataFormats

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
 instance" xmlns:dc="http://purl.org/dc/elements/1.1/"
 xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-
 PMH.xsd">
   <responseDate>2019-08-29T10:12:03.626Z</responseDate>
   <request verb="ListMetadataFormats">http://scicat.esss.se/scicat/oai</request>
  ▼<ListMetadataFormats>
    ▼<metadataFormat>
       <metadataPrefix>oai_dc</metadataPrefix>
       <schema>http://www.openarchives.org/OAI/2.0/oai_dc.xsd</schema>
       <metadataNamespace>http://www.openarchives.org/OAI/2.0/oai_dc/</metadataNamespace>
     </metadataFormat>
    ▼<metadataFormat>
       <metadataPrefix>panosc</metadataPrefix>
      ▼<schema>
         https://github.com/panosc-eu/fair-data-api/blob/master/panosc.xsd
       </schema>
       <metadataNamespace>http://scicat.esss.se/panosc</metadataNamespace>
     </metadataFormat>
   </ListMetadataFormats>
 </OAI-PMH>
```

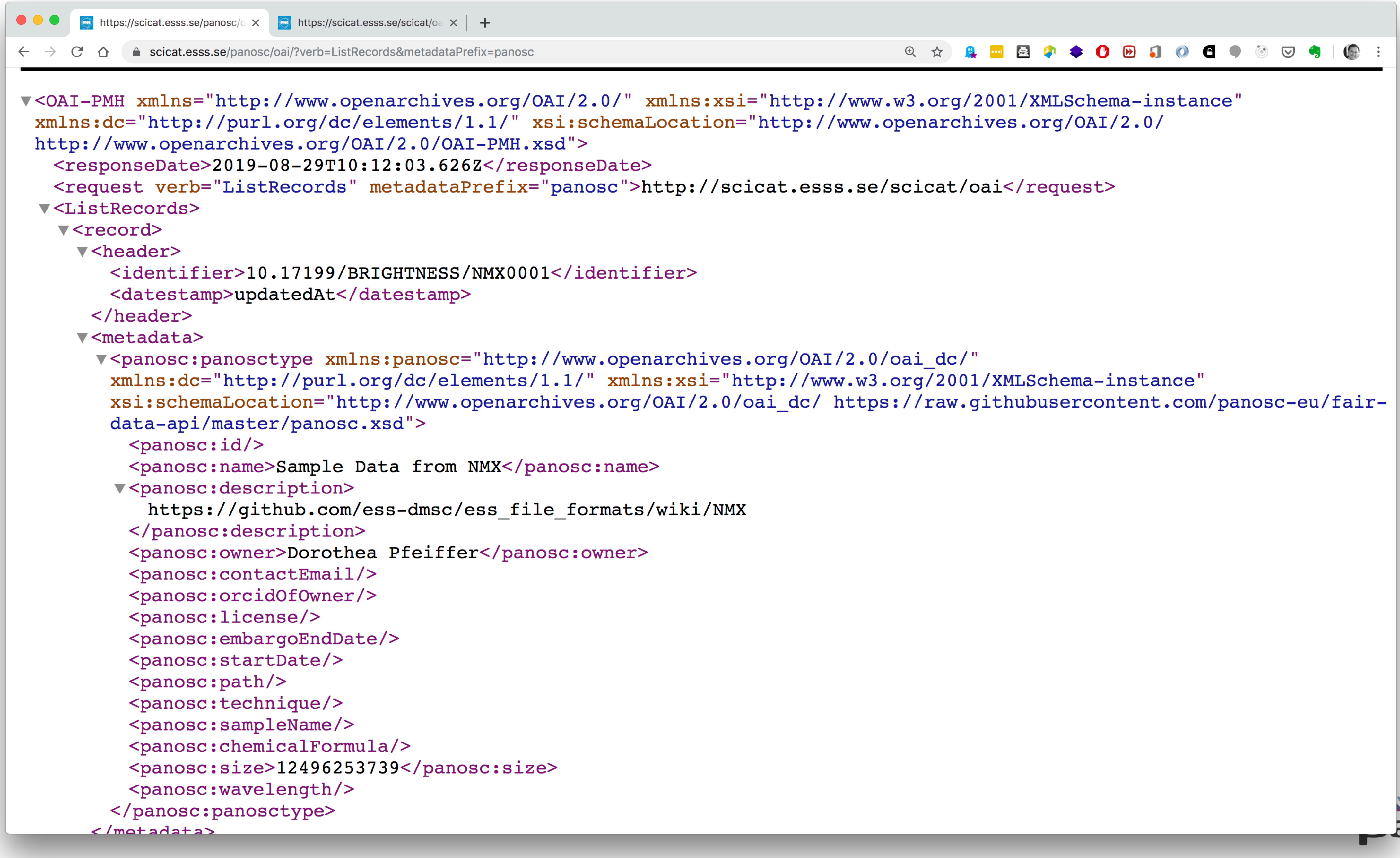
Dublin Core

Dublin Core Elements

Rights	Contributor	Creator
Subject	Coverage	Title
Publisher	Identifier	Description
Type	Date	Source
Relation	Format	Language

panosc
photon and neutron open science cloud

Wavelength	Chemical Formula
Start Date	Sample Name
Facility	Scientific Technique


https://scicat.esss.se/panosc/c
https://scicat.esss.se/scicat/oa
scicat.esss.se/panosc/oai/?verb=ListRecords&metadataPrefix=panosc

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:dc="http://purl.org/dc/elements/1.1/" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2019-08-29T10:12:03.626Z</responseDate>
  <request verb="ListRecords" metadataPrefix="panosc">http://scicat.esss.se/scicat/oai</request>
  <ListRecords>
    <record>
      <header>
        <identifier>10.17199/BRIGHTNESS/NMX0001</identifier>
        <datestamp>updatedAt</datestamp>
      </header>
      <metadata>
        <panosc:panosctype xmlns:panosc="http://www.openarchives.org/OAI/2.0/oai_dc/"
        xmlns:dc="http://purl.org/dc/elements/1.1/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ https://raw.githubusercontent.com/panosc-eu/fair-
        data-api/master/panosc.xsd">
          <panosc:id/>
          <panosc:name>Sample Data from NMX</panosc:name>
          <panosc:description>
            https://github.com/ess-dmsc/ess_file_formats/wiki/NMX
          </panosc:description>
          <panosc:owner>Dorothea Pfeiffer</panosc:owner>
          <panosc:contactEmail/>
          <panosc:orcidOfOwner/>
          <panosc:license/>
          <panosc:embargoEndDate/>
          <panosc:startDate/>
          <panosc:path/>
          <panosc:technique/>
          <panosc:sampleName/>
          <panosc:chemicalFormula/>
          <panosc:size>12496253739</panosc:size>
          <panosc:wavelength/>
        </panosc:panosctype>
      </metadata>
```

OAI-PMH pros and cons

Pros:

✅Well supported by e.g. OpenAire

✅Lots of versions available

✅Little implementation work required

Cons:

❌ OAI-PMH doesn’t scale

❌ In order to change some entries, harvesters have to harvest everything again

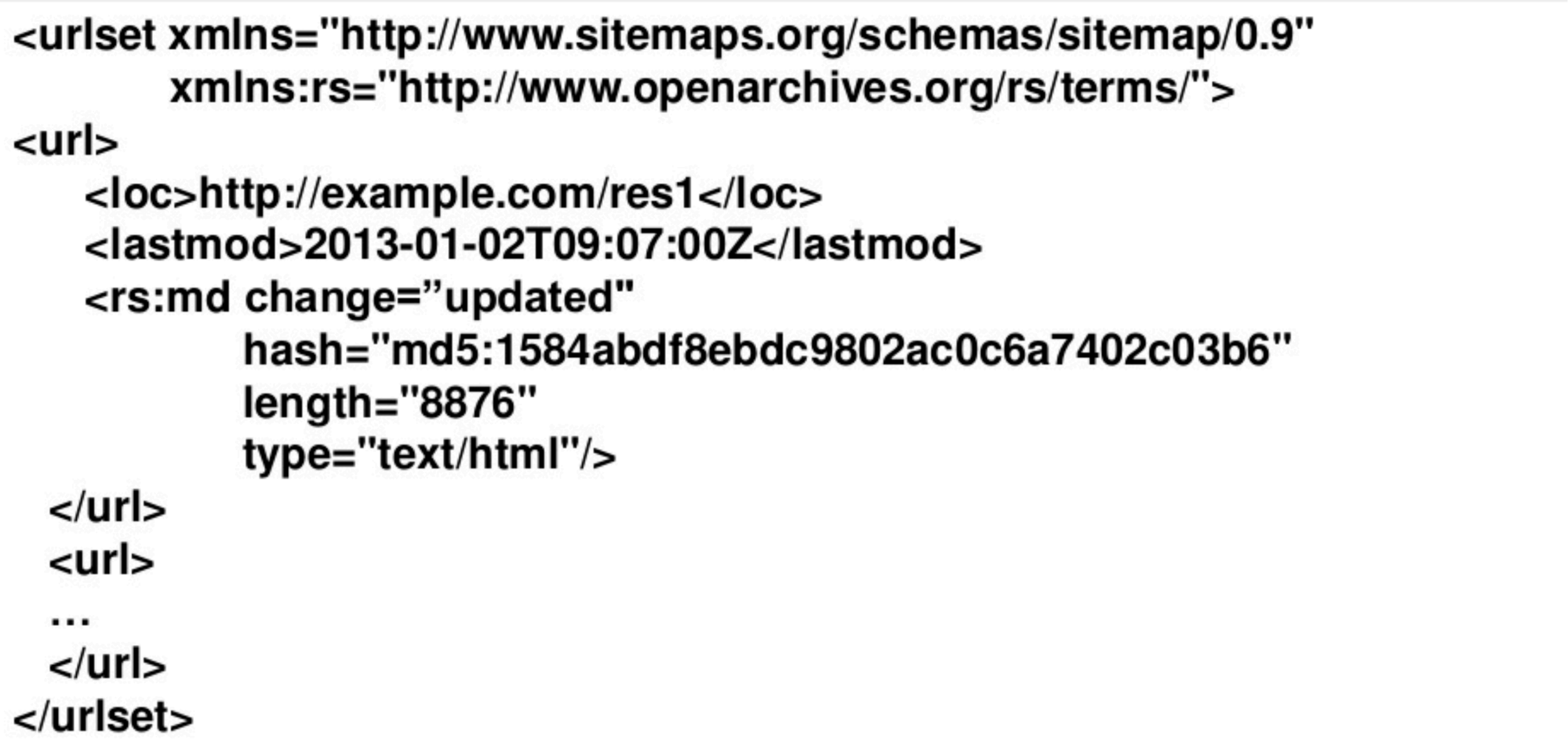
❌ Change not supported

panosc

ResourceSync

- Upgrade/rewrite of OAI-PMH
- Developed to address missing parts of PMH
- Supports Change List, Change Dump, Versioning,
- Sitemap technology so XML files broken into 50 MB chunks
- http://www.openarchives.org/rs/toc
- Works for any search engine

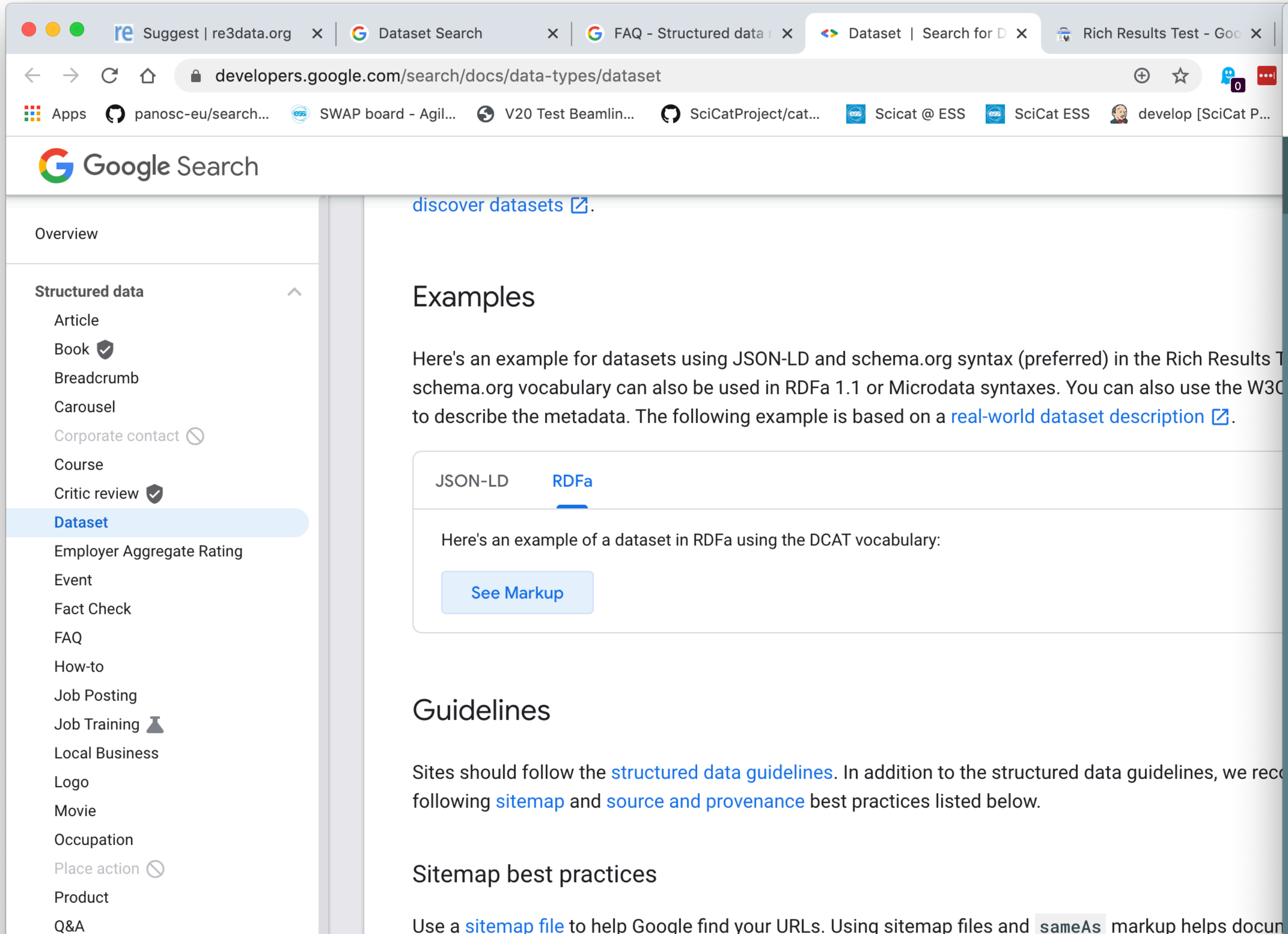
ResourceSync



```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
            xmlns:rs="http://www.openarchives.org/rs/terms/">
<url>
      <loc>http://example.com/res1</loc>
      <lastmod>2013-01-02T09:07:00Z</lastmod>
      <rs:md change=”updated"
                   hash="md5:1584abdf8ebdc9802ac0c6a7402c03b6"
                   length="8876"
                   type="text/html"/>
   </url>
   <url>
   …
   </url>
</urlset>
```

Google Dataset Search

- They will scrape your metadata if it is in JSON-LD or RDF-a format
- https://datasetsearch.research.google.com/



discover datasets.

Examples

Here's an example for datasets using JSON-LD and schema.org syntax (preferred) in the Rich Results T schema.org vocabulary can also be used in RDFa 1.1 or Microdata syntaxes. You can also use the W3C to describe the metadata. The following example is based on a real-world dataset description.

JSON-LD RDFa

Here's an example of a dataset in RDFa using the DCAT vocabulary:

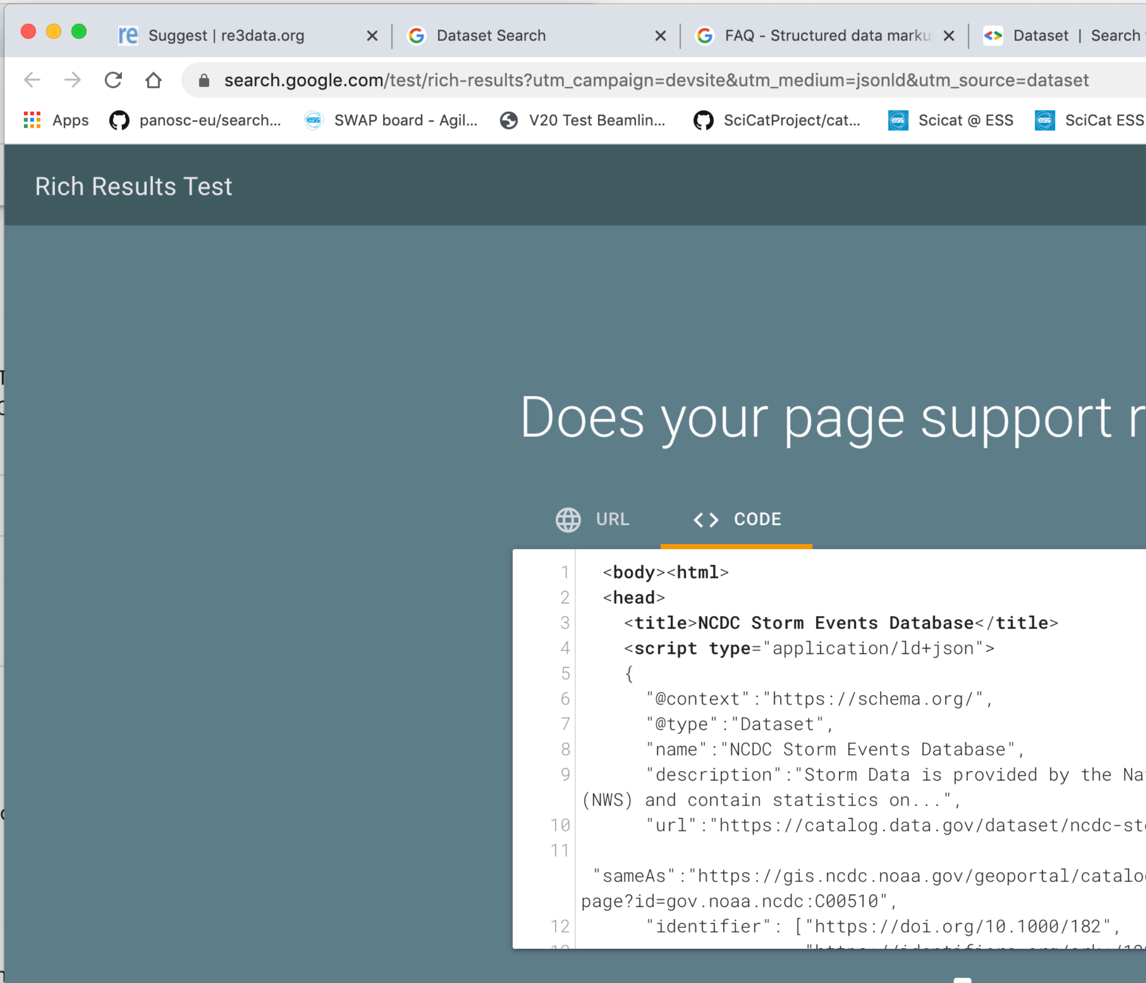
See Markup

Guidelines

Sites should follow the structured data guidelines. In addition to the structured data guidelines, we reco following sitemap and source and provenance best practices listed below.

Sitemap best practices

Use a sitemap file to help Google find your URLs. Using sitemap files and `sameAs` markup helps docum



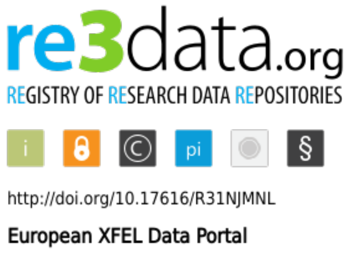
Does your page support r

```
<body><html>
<head>
  <title>NCDC Storm Events Database</title>
  <script type="application/ld+json">
  {
    "@context":"https://schema.org/",
    "@type":"Dataset",
    "name":"NCDC Storm Events Database",
    "description":"Storm Data is provided by the Na
(NWS) and contain statistics on...",
    "url":"https://catalog.data.gov/dataset/ncdc-st

 "sameAs":"https://gis.ncdc.noaa.gov/geoportal/catalo
page?id=gov.noaa.ncdc:C00510",
    "identifier": ["https://doi.org/10.1000/182",
```

Roadmap

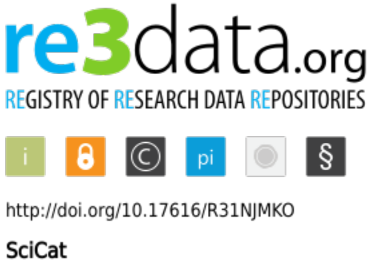
1. Deploy OAI-PMH at 6 PaNOSC institutes
2. Register at re3data
3. Provide data to B2FIND/ OpenAire
4. Upgrade to ResourceSync



re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

http://doi.org/10.17616/R31NJMNL

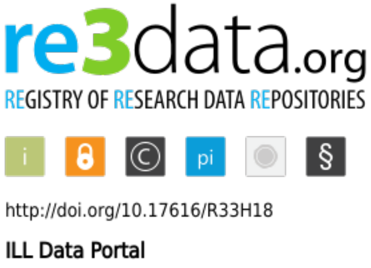
European XFEL Data Portal



re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

http://doi.org/10.17616/R31NJMKO

SciCat



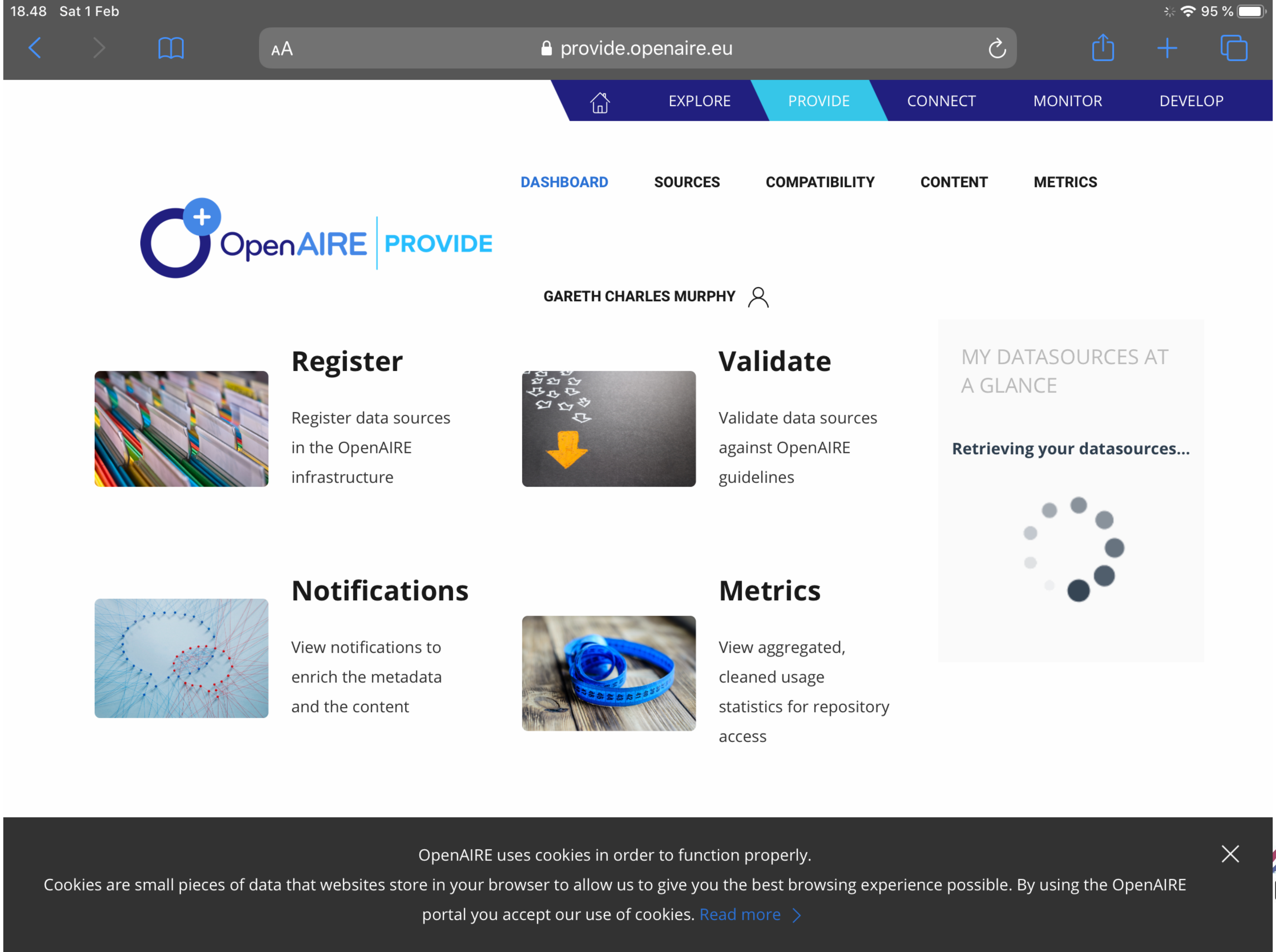
re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

http://doi.org/10.17616/R33H18

ILL Data Portal

Roadmap

	Task	Target	CERIC	ELI	ESRF	ESS	ILL	XFEL
1	Deploy OAI-PMH	March 2020	✅			✅	✅	
2	Re3data registration	March 2020	✅		✅	✅	✅	✅
3	OpenAIRE registration	March 2020	✅			✅	✅	
4	Upgrade to ResourceSync	2021						



Register

Register data sources in the OpenAIRE infrastructure

Validate

Validate data sources against OpenAIRE guidelines

Notifications

View notifications to enrich the metadata and the content

Metrics

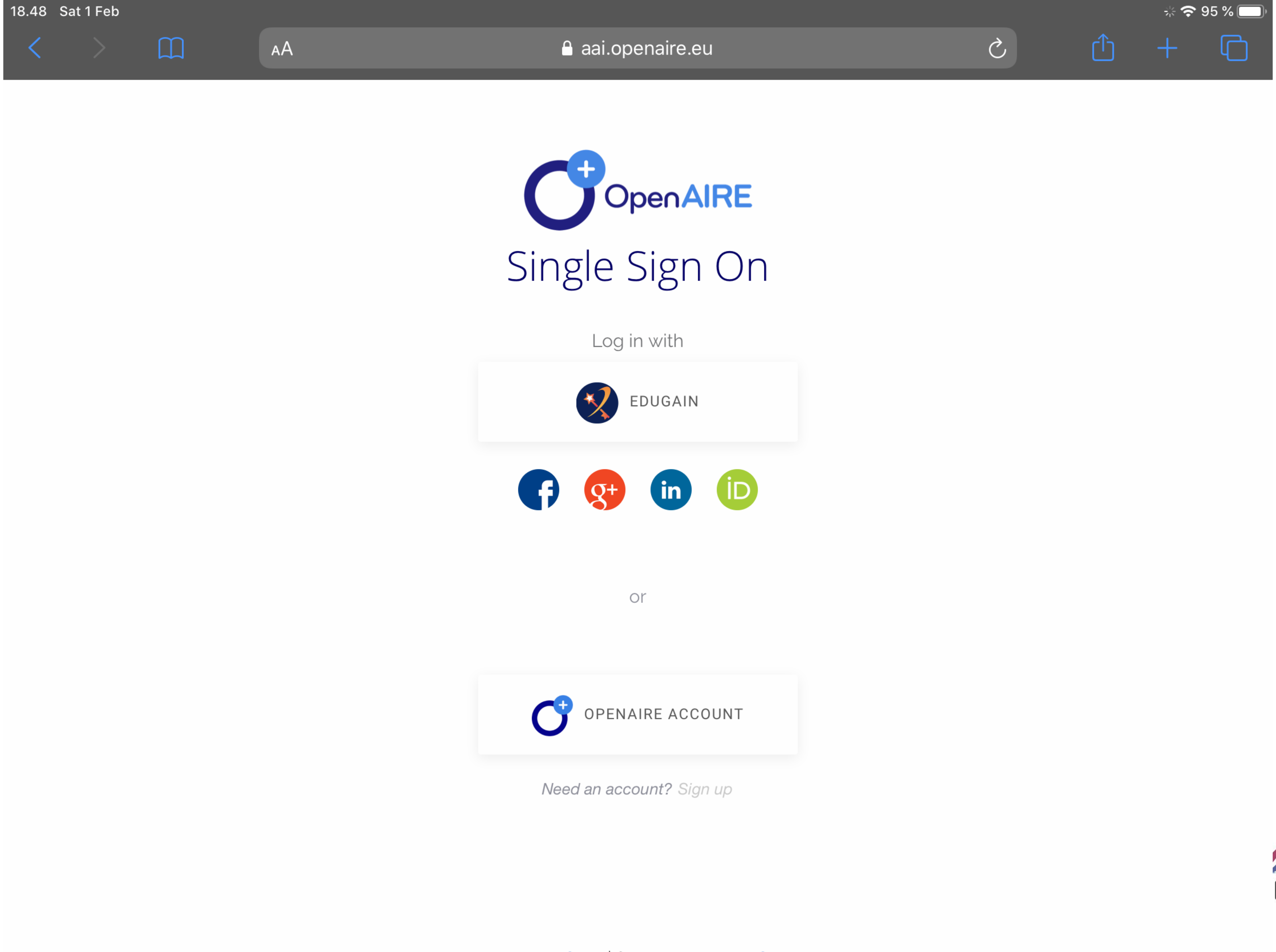
View aggregated, cleaned usage statistics for repository access

MY DATASOURCES AT A GLANCE

Retrieving your datasources...

OpenAIRE uses cookies in order to function properly.

Cookies are small pieces of data that websites store in your browser to allow us to give you the best browsing experience possible. By using the OpenAIRE portal you accept our use of cookies. Read more

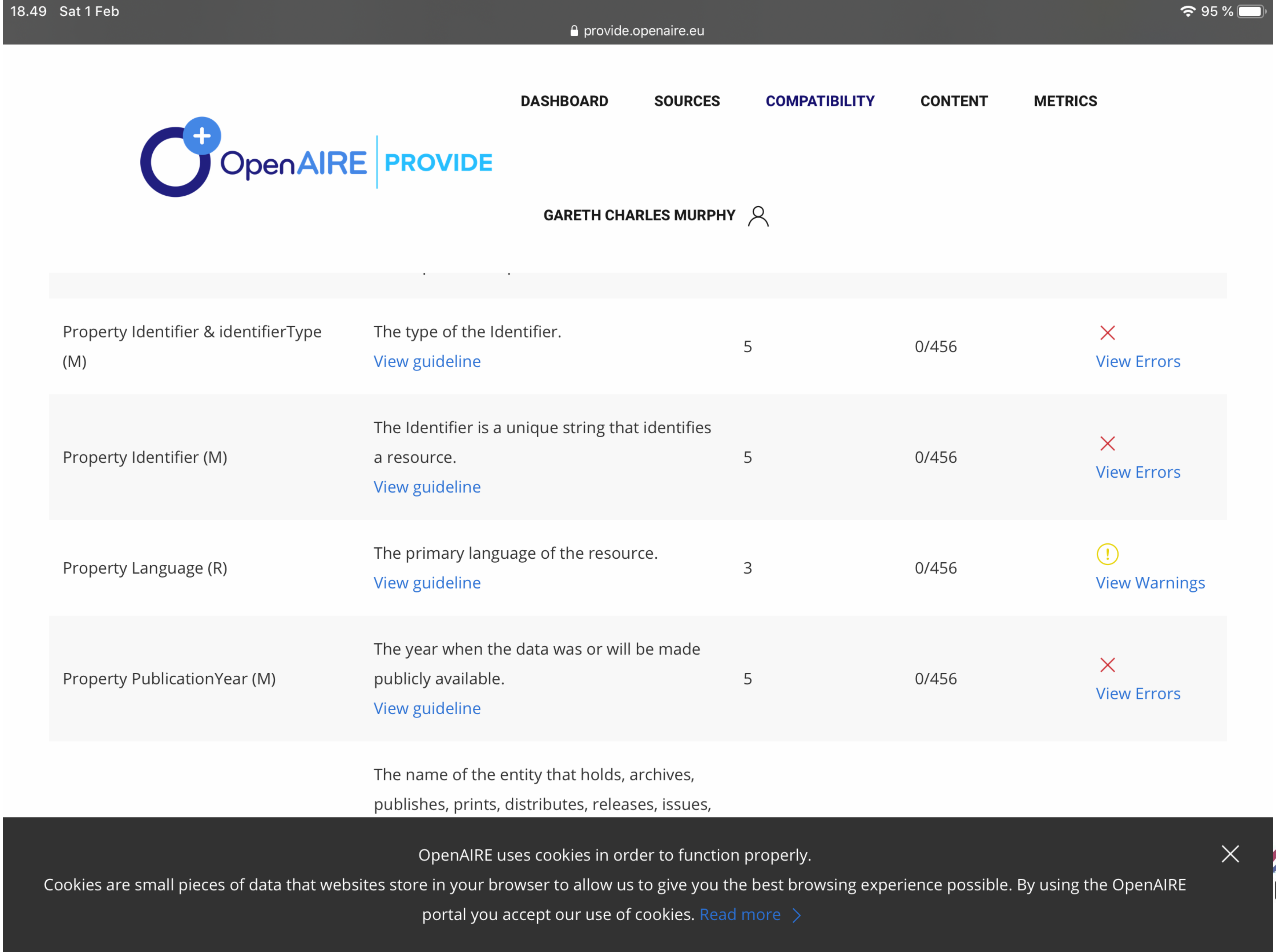


Single Sign On

Log in with

or

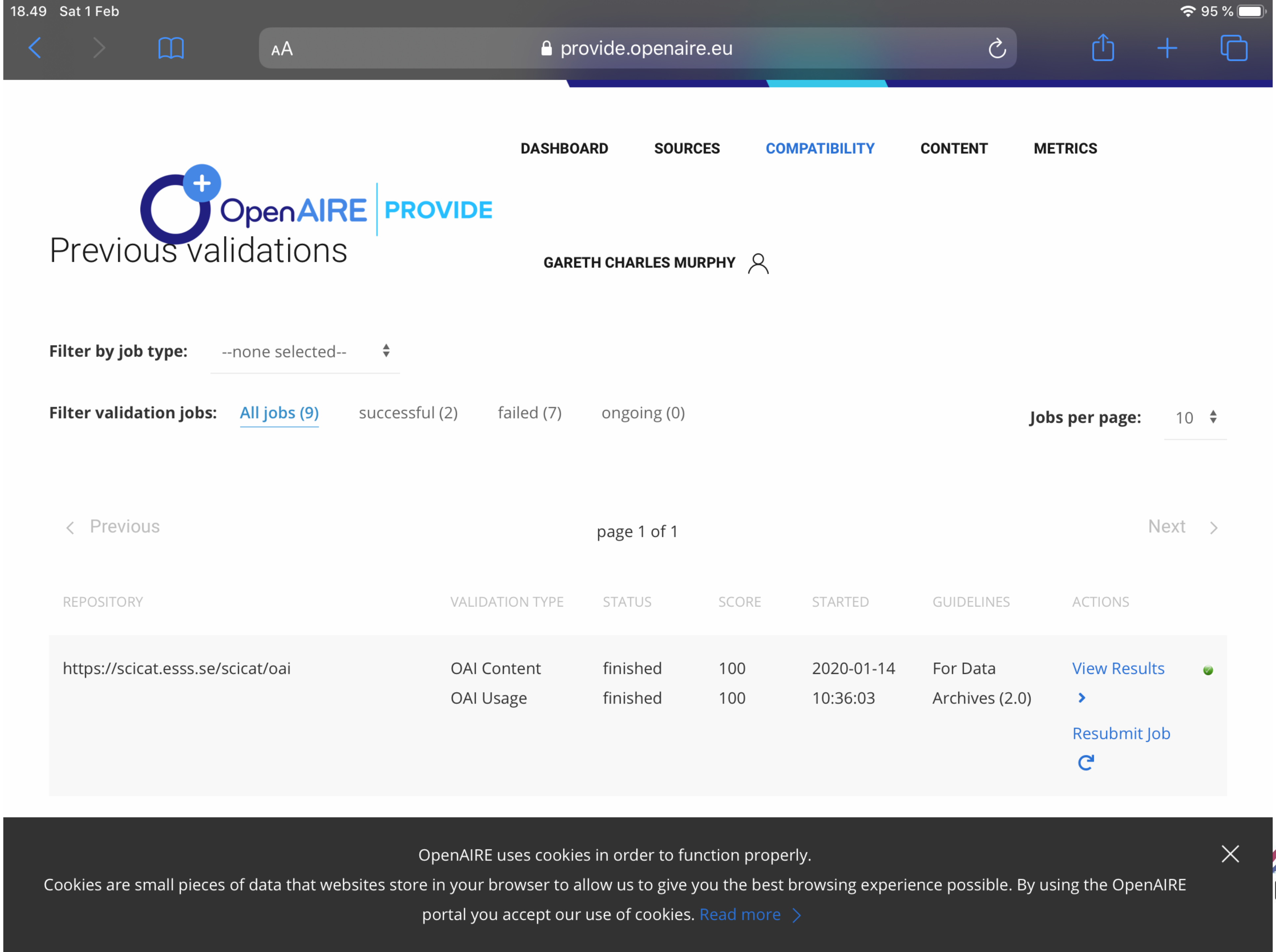
Need an account? Sign up



Property Identifier & identifierType (M)	The type of the Identifier. View guideline	5	0/456	✕ View Errors
Property Identifier (M)	The Identifier is a unique string that identifies a resource. View guideline	5	0/456	✕ View Errors
Property Language (R)	The primary language of the resource. View guideline	3	0/456	(!) View Warnings
Property PublicationYear (M)	The year when the data was or will be made publicly available. View guideline	5	0/456	✕ View Errors
	The name of the entity that holds, archives, publishes, prints, distributes, releases, issues,			

OpenAIRE uses cookies in order to function properly.

Cookies are small pieces of data that websites store in your browser to allow us to give you the best browsing experience possible. By using the OpenAIRE portal you accept our use of cookies. Read more >



Previous validations

Filter by job type: --none selected--

Filter validation jobs: All jobs (9) successful (2) failed (7) ongoing (0)

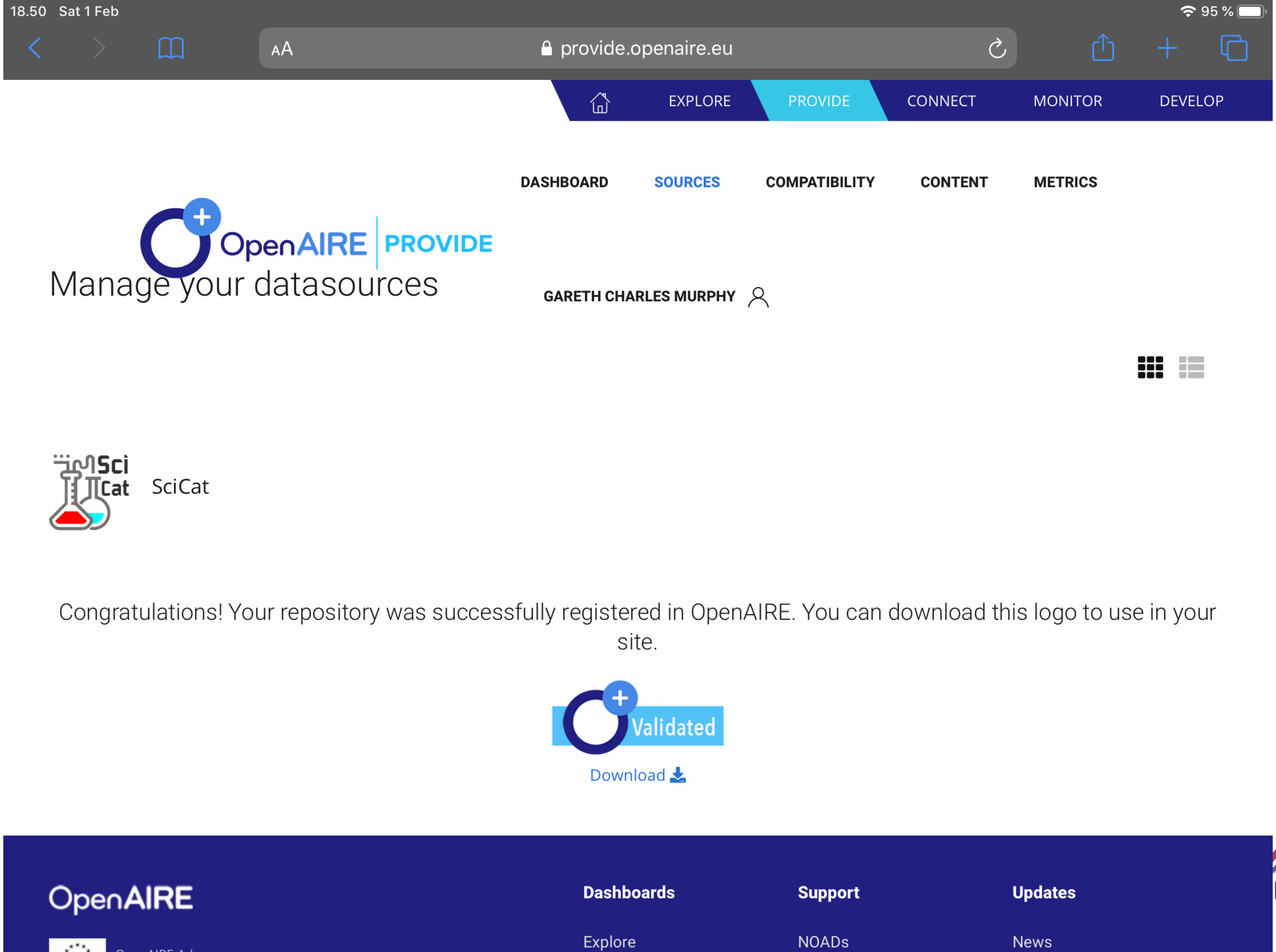
Jobs per page: 10

< Previous page 1 of 1 Next >

REPOSITORY	VALIDATION TYPE	STATUS	SCORE	STARTED	GUIDELINES	ACTIONS
https://scicat.esss.se/scicat/oai	OAI Content OAI Usage	finished finished	100 100	2020-01-14 10:36:03	For Data Archives (2.0)	View Results ● Resubmit Job

OpenAIRE uses cookies in order to function properly.

Cookies are small pieces of data that websites store in your browser to allow us to give you the best browsing experience possible. By using the OpenAIRE portal you accept our use of cookies. Read more >



Manage your datasources

Congratulations! Your repository was successfully registered in OpenAIRE. You can download this logo to use in your site.