panosc
photon and neutron
open science cloud

Harvesting by EOSC repos, OAI-PMH, extended schemas

Gareth Murphy

European Spallation Source

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 823852

Harvesting vs authenticated search

- 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

http://doi.org/10.17616/R31NJMKO

SciCat

re3data.org REGISTRY OF RESEARCH DATA REPOSITORIES

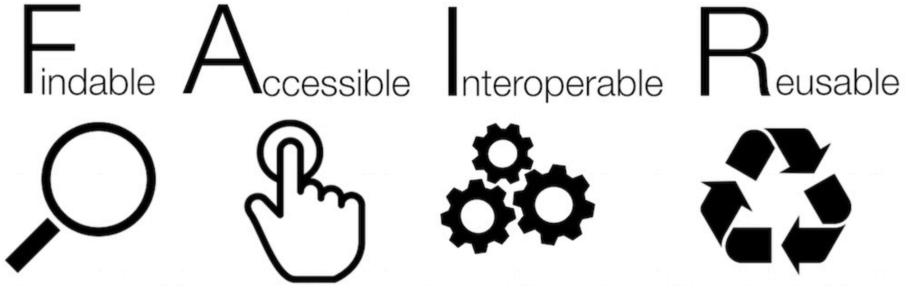
http://doi.org/10.17616/R33H18

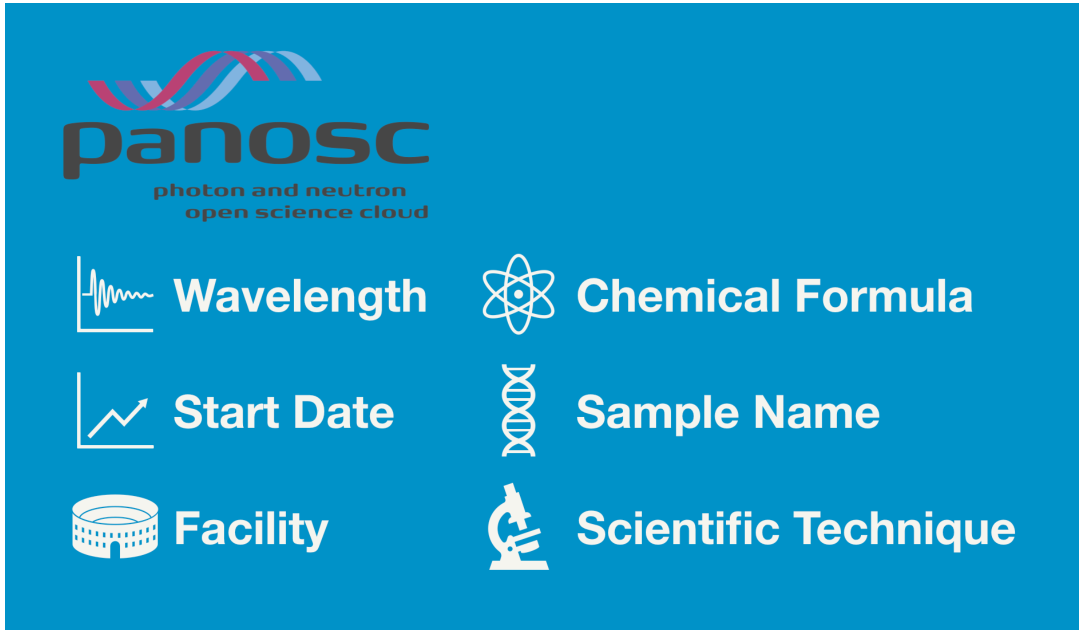
ILL Data Portal

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No.823852

Metadata standard format(s)

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





Wavelength

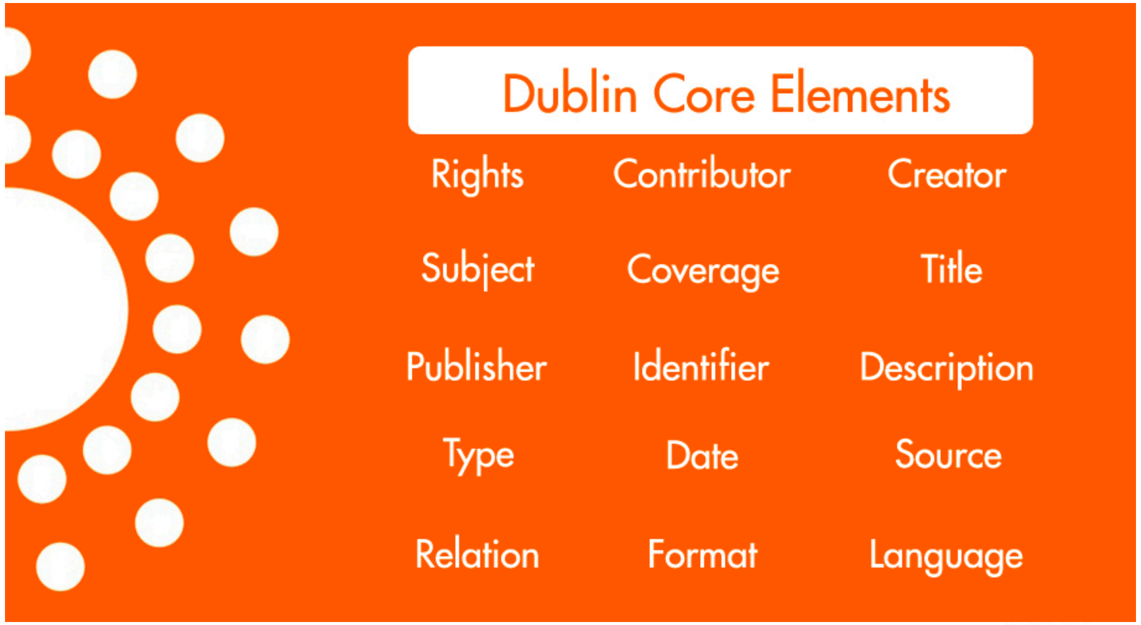
Chemical Formula

Start Date

Sample Name

Facility

Scientific Technique



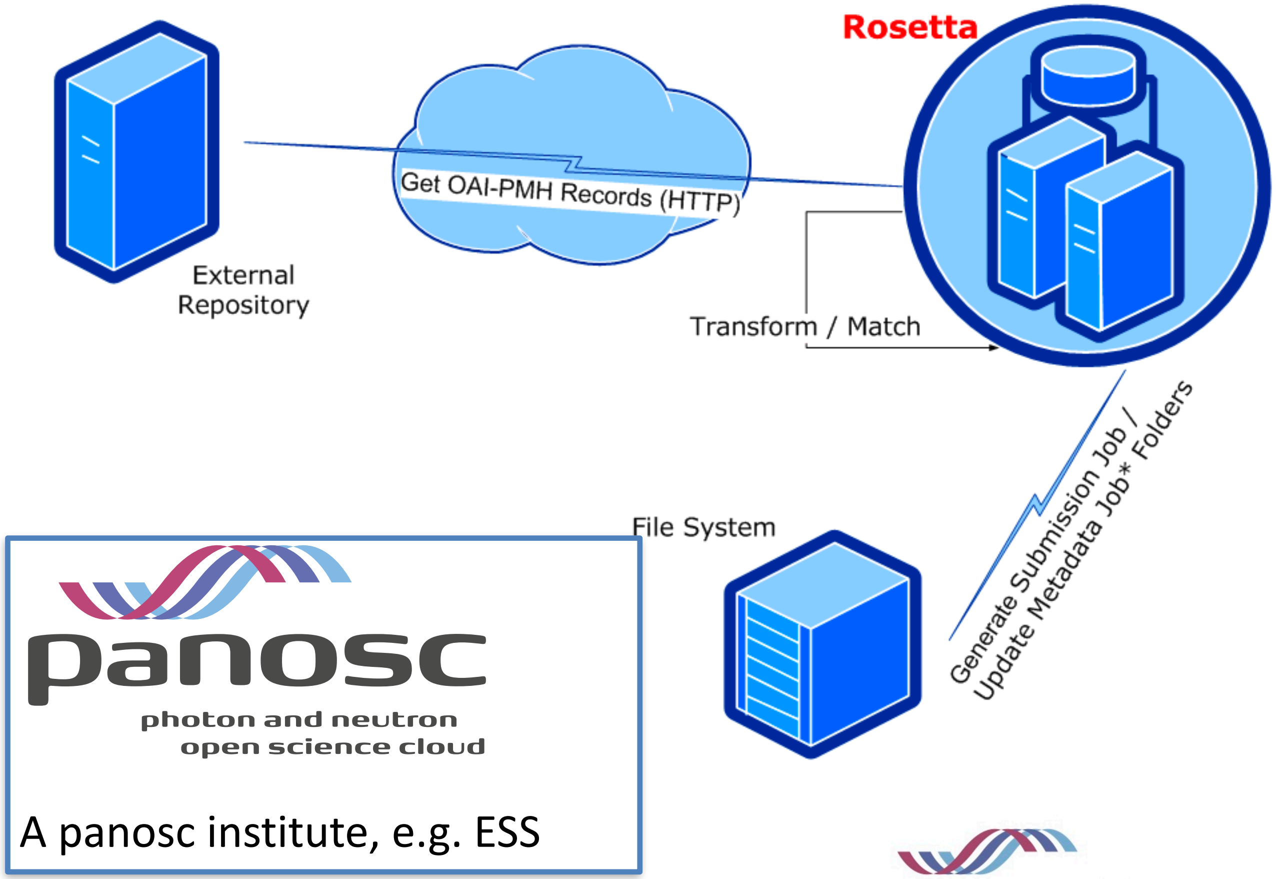
Dublin Core Elements

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

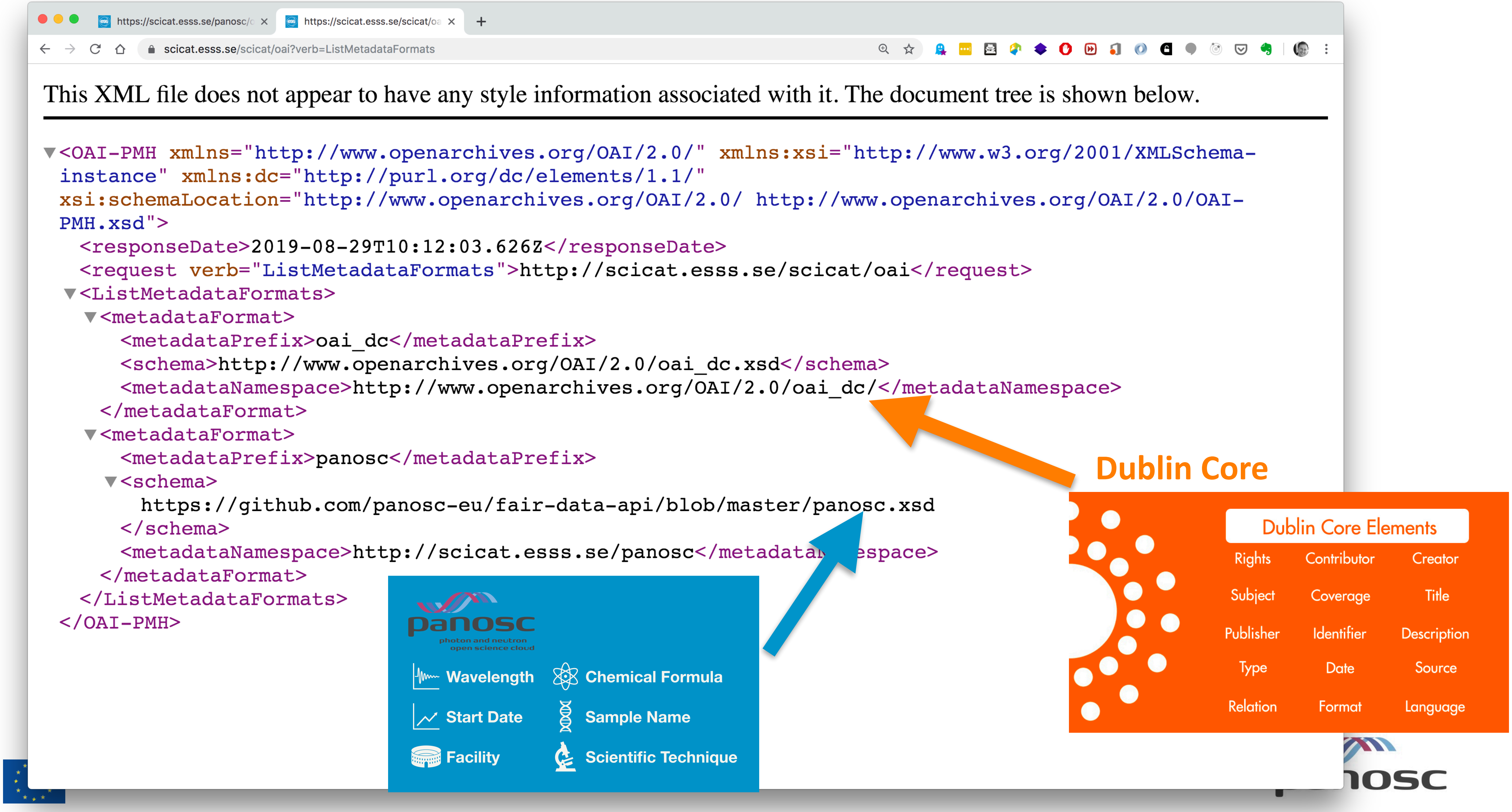
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

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
File System
Generate Submission Job / Update Metadata Job* Folders
panosc
photon and neutron open science cloud
A panosc institute, e.g. ESS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852


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>
```

panosc
photon and neutron
open science cloud

Wavelength

Chemical Formula

Start Date

Sample Name

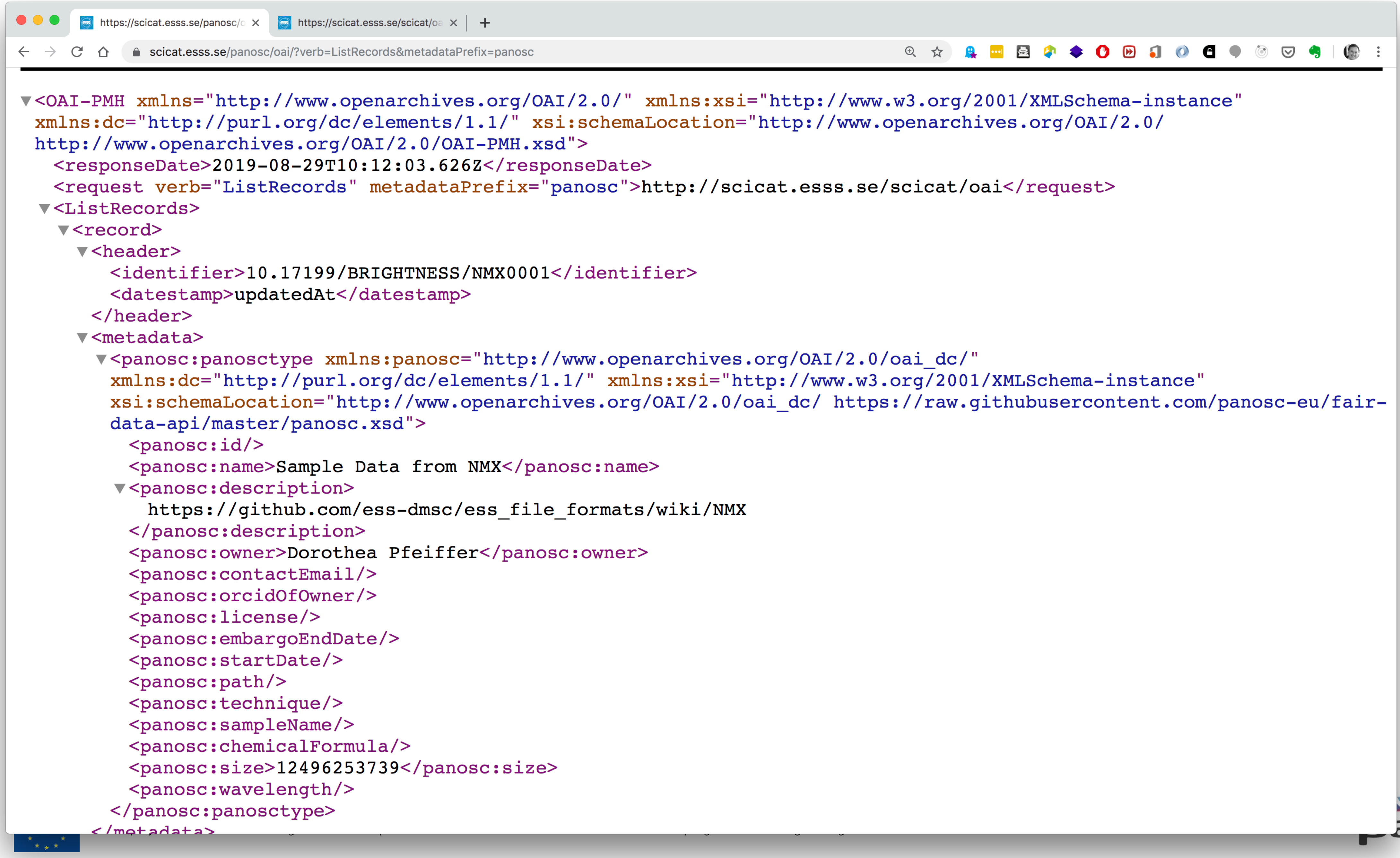
Facility

Scientific Technique

Dublin Core

Dublin Core Elements

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


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

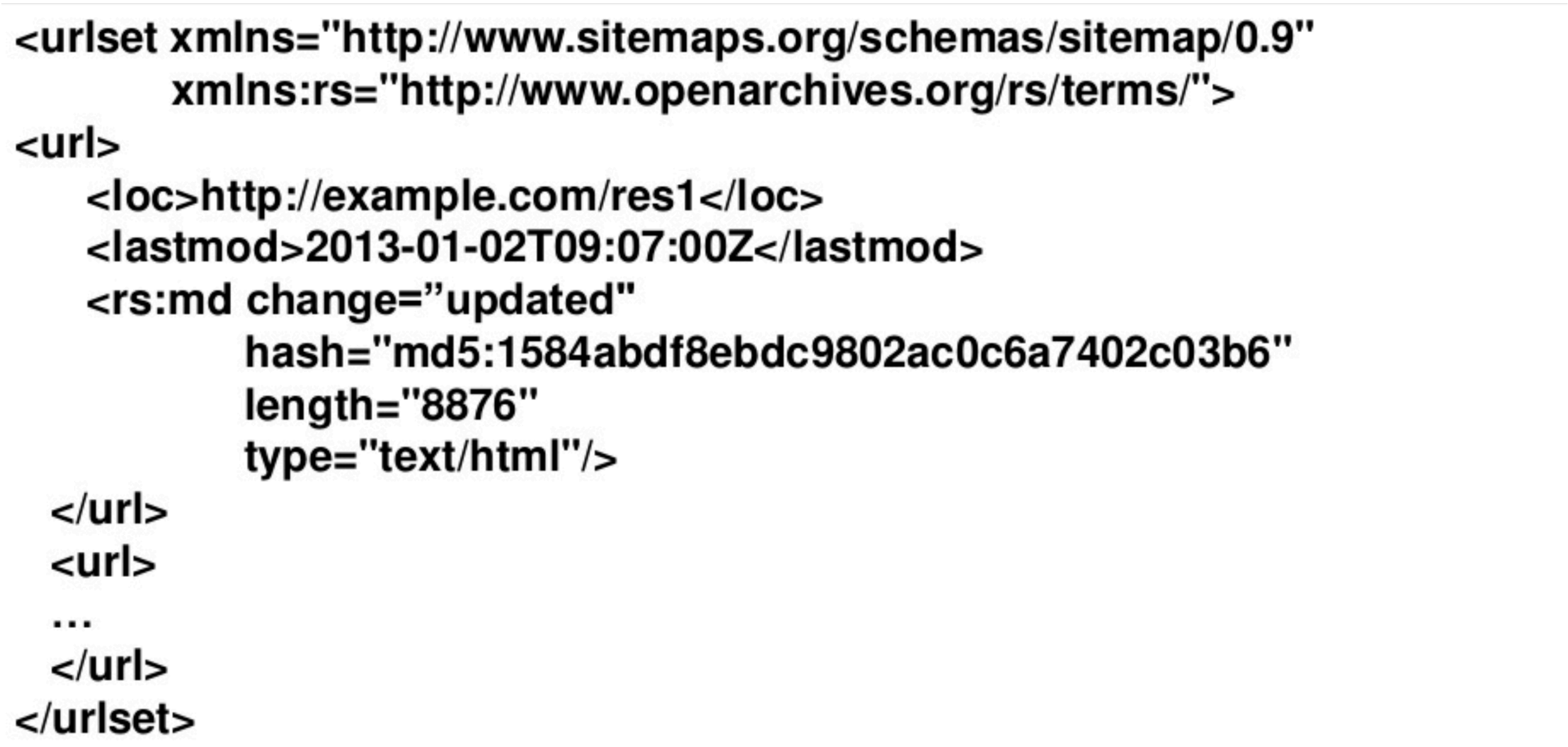
This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 823852

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

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

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>
```

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 823852

Roadmap

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

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

OpenAIRE

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 823852

panosc