EOSC-IF

Interoperability Guideline: Access to content via PID

The EOSC Future project is co-funded by the European Union Horizon Programme call INFRAEOSC-03-2020, Grant Agreement number 101017536

Version 2.0
August 2023

EOSC-IF / Interoperability Guideline: Access to content via PID

Lead by **CNR**

Authored by the EOSC Future Working Group on Research Product Publishing: Alessia Bardi (CNR), Paolo Manghi (OpenAIRE), Jose Benito Gonzalez Lopez (CERN), Chris Ariyo (EUDAT), Andreas Czerniak (Bielefeld University Library), Paul Gondim van Dongen (SURF), Georgios Kakaletris (NKUA), Raul Palma (PSNC), Silvio Peroni (University of Bologna), Hans van Piggelen (SURF), Mark van de Sanden (EUDAT), Diego Scardaci (EGI), Jochen Schirrwagen (Bielefeld University Library), Debora Testi (CINECA), Raphaël Tournoy (CNRS), Irena Vipavc (Social Science Data Archive, University of Ljubljana), Deborah Grbac (Library of Università Cattolica del Sacro Cuore di Milano), Carl-Fredrik Enell (EISCAT Scientific Association), Guido Aben (CS3MESH4EOSC), Ivan Heibi (University of Bologna), Jorik van Kemenade (SURF)

Reviewed by Michelle Williams (GÉANT)

Dissemination Level of the Document

Public

Abstract

An important aspect of Open Science is the possibility to re-use existing research products (e.g. research data), deposited in repositories and accessible via their persistent identifiers (e.g. handle, DOI, ARK). However, there is no standard way a service can access the actual content behind persistent identifiers, as these typically resolve to the landing pages of the research products.

The lack of standard for accessing the actual content identified by persistent identifiers makes the automatic consumption of research products hardly implementable and, when possible, limited to the persistent identifiers issued by a specific repository (e.g. the first prototype of the EGI Data Transfer Service integrated in the EOSC EXPLORE portal supported only DOIs from Zenodo).

The EOSC Future Working Group on Research Product Publishing proposes the adoption of the Publication Boundary Pattern of the SignPosting protocol and recomends it for inclusion as interoperability guideline in the EOSC IF.

Version History

Version	Date	Authors/Contributors	Description
V0.1	08/05/2023	Alessia Bardi (CNR)	Initiation
V1.0	27/06/2023	Alessia Bardi (CNR)	Final version for submission
V2.0	30/08/2023	Alessia Bardi (CNR)	Updates based on the suggestions from EIAC members
V2.1	05/09/2023	Alessia Bardi (CNR)	Table 1 fits to page

Copyright Notice

Table of Contents

Glossary

EOSC Future project Glossary is incorporated by reference: https://wiki.eoscfuture.eu/x/JQCK

List of Abbreviations

Acronym	Definition
CRIS	Current Research Information System
DOI	Digital Object Identifier
EOSC	European Open Science Cloud
OJS	Open Journal System
PID	Persistent identifier
RDF	Resource Description Framework
URL	Uniform Resource Locator
WG	Working Group

1 Intended Audience

- Data sources hosting payloads of research products of any type. Examples: data archives, data repositories, thematic repositories, institutional repositories, pre-print repositories, journal publishing platforms
- Services that would like to access the payloads of a research product given its persistent identifier

2 Description and main features

Persistent Identifiers (PIDs) are unique identifiers of resources that are usually resolvable URLs (e.g. DOI[1], Handle[2]).

Given a PID URL, we can perform an HTTP request to get the resource. Typically, the returned content is the metadata of the resource (a description of it). Metadata can be returned in different format, depending on the server and on the specific HTTP request: an HTML landing page, json-ld, and RDF are some examples. The feature is supported by the concept of "content-negotiation", but there is no standard mechanism to get directly to the actual resource, bypassing the landing page/metadata.

Therefore, software programs must implement specific strategies for specific servers, or crawl the landing pages to identify the URL from which the resource can be downloaded.

With this guideline, we suggest standard protocols that can be used to support software programs at consuming the resources identified by PIDs in a consistent way across servers and regardless the specific type of the PID.

3 Response to Community Need

A researcher uses a thematic service to run analysis on a dataset available on a repository. Instead of downloading the dataset files from the repository and uploading them to a storage resource of the e-infrastructure, the researcher gives the dataset's PID as input to the thematic service, which can get the files and store them where they can be analysed.

4 Licensing Information

[1] https://www.doi.org/the-identifier/what-is-a-doi/

[2] https://www.handle.net/index.html

5 Related Standards

Table 1 Related Standards

Title	Short Description	relatedStandardIdentifier
SignPosting Publication Boundary Pattern	Landing pages support humans that interact with scholarly objects on the web, providing descriptive metadata and links to content. These pages are not optimized for use by machine agents that navigate the scholarly web. For example, how can a robot determine which links on the myriad of landing pages lead to content and which to metadata? Signposting caters to machine agents by providing this information, and more, in a standards-based way.	https://signposting.org/publication_boundary/

6 Integration Options

The SignPosting Publication Boundary Pattern[3] uses typed links to help machines find the resources that make up a digital object available online. The relation type that is relevant for this guideline is "item" and it is used to link to the actual payload(s) of the digital object.

The pattern can be implemented in two ways:

- via Linkset as json document (according to the proposed standard RFC9264[4])
- via HTTP link header

1. Via Linkset as json document

The typed links are available in a dedicated resource whose URL is discoverable using the HTTP protocol.

We can issue an HTTP HEAD request and find the Link header with rel="linkset" to find the URL from which we can access a file that contains information about the digital object, including the items it is composed of.

The Link header has the form

```
Link: <URL to the linkset>; rel="linkset"; type="application/linkset+json"
```

The linkset document must comply with the specification RFC9264 (see details in section 4.2[5]) and include at least a link target object with relation type "item" as in the follwing example:

```
{ "linkset":
  [
    { "anchor": "https://example.net/bar",
      "item": [
        {"href": "https://example.com/foo1.data"},
      ]
    }
  ]
}
```

[3] https://signposting.org/publication_boundary/

[4] https://www.rfc-editor.org/info/rfc9264

[5] https://www.rfc-editor.org/rfc/rfc9264.html#name-json-document-format-applic

To further specify the target resource, a “type” attribute can be added to specify the mime-type:

```
{ "linkset":
  [
    { "anchor": "https://example.net/bar",
      "item": [
        {"href": “https://example.com/foo1.data”,
         "type": “mime-type”}
      ]
    }
  ]
}
```

See section 8 for an example of usage of the pattern with linkset as json document.

2. Via HTTP Link header

The typed links are available in the Link HTTP header.

We can issue an HTTP HEAD request and find the Link in the response header. It contains relationships to other resources and to the items the digital object is composed of.

The Link header has the form of a comma separated list of target objects, each identified by a URL, a semantic relationships, and a type:

```
Link: <URL1>;rel="link semantics";type="mime-type", <URL2>;rel="link semantics2";type="mime-type2". . .
```

Among the target objects, there shall be one with relation type “item” as in the follwing example:

```
Link: <URL_to_jsonLD_metadata>;rel="describedby";type="application/ld+json",
<URL_to_bibtex_citation>;rel="describedby";type="application/x-bibtex",
<URL_to_payload>; rel="item";type="application/zip",
```

See section 8 for examples of usage of the pattern with linkset in the HTTP headers.

7 Interoperability Guidelines

The Interoperability Guidelines are defined by the SignPosting protocol, Publication Boundary Pattern

8 Examples of solutions implementing this specification

The page https://signposting.org/adopters/ lists the known adopters. However, not all adopters implement the Publication Boundary Pattern that is relevant for our context. Those that do implement the pattern are listed below.
Examples are given using the curl command (https://curl.se/docs/manpage.html). In particular, the following options of curl are used, also in combination:

- -I (or --head) to execute an HTTP HEAD request and get only the header of the response from the server
- -L (or --location) to automatically follow redirects (upon 3xx HTTP response of the serve)

Open Journal System (OJS)
Open journal system is a platform for the management of research journals.
We count about 1K journals using OJS contributing to the OpenAIRE Graph.
Type of implementation: via linkset

Example:

1. Get the URL to the linkset

```
curl -IL https://doi.org/10.4401/ag-7507
```

Response includes the URL to the linkset in the header element "Link" with rel="linkset"

```
HTTP/2 302
date: Wed, 30 Aug 2023 11:53:22 GMT
content-type: text/html;charset=utf-8
content-length: 219
location: http://www.annalsofgeophysics.eu/index.php/annals/article/view/7507
vary: Accept
expires: Wed, 30 Aug 2023 12:10:12 GMT
permissions-policy: interest-cohort=(),browsing-topics=()
cf-cache-status: DYNAMIC
report-to:
{"endpoints":[{"url":"https:\/\/a.nel.cloudflare.com\/report\/v3?s=DSCy%2BMPMKuP7LG4jBAB71OAosxo1
J42Ean0TQUc90Rvb9ESad4mDnCBpGEklhRrukpYWOC9Wzhn%2FWgVKkoNw5lxfQN0DCDf0S5D22Oh3wBAa4FtFX5RLK2S9bGm
6rzhGDHqM4y8%3D"}],"group":"cf-nel","max_age":604800}
nel: {"success_fraction":0,"report_to":"cf-nel","max_age":604800}
strict-transport-security: max-age=31536000; includeSubDomains; preload
server: cloudflare
cf-ray: 7fecd53acb4b3757-MXP
alt-svc: h3=":443"; ma=86400

HTTP/1.1 301 Moved Permanently
Date: Wed, 30 Aug 2023 11:53:22 GMT
Server: Apache
X-Frame-Options: SAMEORIGIN
Location: https://www.annalsofgeophysics.eu/index.php/annals/article/view/7507
Content-Type: text/html; charset=iso-8859-1

HTTP/1.1 200 OK
Date: Wed, 30 Aug 2023 11:53:22 GMT
Server: Apache
Link: <https://www.annalsofgeophysics.eu/index.php/annals/sp-linkset/article/7507>;
rel="linkset"; type="application/linkset+json"
Cache-Control: no-store
Set-Cookie: OJSSID=aqfrf3qh3ta31pmk67tmorju0u; path=/; domain=www.annalsofgeophysics.eu
X-Frame-Options: SAMEORIGIN
Content-Type: text/html; charset=utf-8
```

2. Get the json file of the linkset

```
curl https://www.annalsofgeophysics.eu/index.php/annals/sp-linkset/article/7507
```

The response is a json file that contains the link to the PDF in the "item" element (part of the json is omitted below for the sake of readability):

```
{
    "linkset": {
        "anchor": "https://www.annalsofgeophysics.eu/index.php/annals/article/view/7507",
        "author": [
            {
                "href": "https://orcid.org/0000-0002-4311-0897"
            }
        ],
. . .
        "item": [
            {
                "href":
"https://www.annalsofgeophysics.eu/index.php/annals/article/download/7507/6808",
                "type": "application/pdf"
            }
        ]
    }
```

Pangaea

Pangaea is a data repository hosting more than 400K research data. It is a repository registered in the EOSC Marketplace and its research products are available in EOSC EXPLORE.

Type of implementation: via HTTP link headers

Example:

Get the link information in the headers. We can find it in the entry with 'rel="item"

```
curl -IL https://doi.org/10.1594/PANGAEA.954506
```

Response:

```
HTTP/2 302
date: Wed, 30 Aug 2023 11:55:16 GMT
content-type: text/html;charset=utf-8
content-length: 175
location: https://doi.pangaea.de/10.1594/PANGAEA.954506
vary: Accept
expires: Wed, 30 Aug 2023 11:56:44 GMT
permissions-policy: interest-cohort=(),browsing-topics=()
cf-cache-status: DYNAMIC
report-to:
{"endpoints":[{"url":"https:\/\/a.nel.cloudflare.com\/report\/v3?s=8IsMyL1Txq6g6XbTU2XRP51mdrIWPP
YOtOAfx4MWtmKEllmB23fAE1mU6KRu4DJLr3cxVsfEtOQRzzZhvLkqMNUueXIGZ3Wj0fDpo5dDMA36CFpFfjT7BXZ6pHUDIhk
h9EnbYEI%3D"}],"group":"cf-nel","max_age":604800}
nel: {"success_fraction":0,"report_to":"cf-nel","max_age":604800}
strict-transport-security: max-age=31536000; includeSubDomains; preload
server: cloudflare
cf-ray: 7fecd8026c2bbab8-MXP
alt-svc: h3=":443"; ma=86400

HTTP/2 200
server: nginx/1.25.2
date: Wed, 30 Aug 2023 11:55:16 GMT
content-type: text/html;charset=utf-8
content-length: 61905
vary: Origin, Cookie, Authorization, Accept
set-cookie: pansessid=91f9fad04b7bb5e15513e7a677219dd6; Path=/; Domain=.pangaea.de; Secure;
HttpOnly
cache-control: public
x-cid: 91f9fad04b7bb5e15513e7a677219dd6
```

```
link: <https://doi.org/10.1594/PANGAEA.954506>;rel="cite-as",
<https://doi.pangaea.de/10.1594/PANGAEA.954506?format=citation_ris>;rel="describedby";type="appli
cation/x-research-info-systems",
<https://doi.pangaea.de/10.1594/PANGAEA.954506?format=metadata_datacite4>;rel="describedby";type=
"application/vnd.datacite.datacite+xml",
<https://doi.pangaea.de/10.1594/PANGAEA.954506?format=metadata_jsonld>;rel="describedby";type="ap
plication/ld+json",
<https://doi.pangaea.de/10.1594/PANGAEA.954506?format=metadata_panmd>;rel="describedby";type="app
lication/vnd.pangaea.metadata+xml",
<https://doi.pangaea.de/10.1594/PANGAEA.954506?format=metadata_dif>;rel="describedby";type="appli
cation/vnd.nasa.dif-metadata+xml",
<https://doi.pangaea.de/10.1594/PANGAEA.954506?format=metadata_iso19139>;rel="describedby";type="
application/vnd.iso19139.metadata+xml",
<https://doi.pangaea.de/10.1594/PANGAEA.954506?format=citation_text>;rel="describedby";type="text
/x-bibliography",
<https://doi.pangaea.de/10.1594/PANGAEA.954506?format=citation_bibtex>;rel="describedby";type="ap
plication/x-bibtex",
<https://doi.pangaea.de/10.1594/PANGAEA.954506?format=zip>;rel="item";type="application/zip",
```

```
<https://orcid.org/0000-0002-2078-0361>;rel="author", <https://orcid.org/0000-0001-7313-
100X>;rel="author", <https://orcid.org/0000-0002-4493-1734>;rel="author",
<https://orcid.org/0000-0003-4207-0309>;rel="author"
x-robots-tag: index,follow,archive
alt-svc: h3=":443"; ma=2592000
strict-transport-security: max-age=31536000
```

```
x-ua-compatible: IE=Edge
x-content-type-options: nosniff
x-frame-options: SAMEORIGIN
```

DSpace CRIS

According to https://signposting.org/adopters/, starting with version 5.8.2, the open source DSpace-CRIS system has built-in support for the Publication Boundary pattern.
About 20 CRIS systems are currently contributing to the OpenAIRE Graph, but some of them might not support the Publication Boundary (it depends on the specific platform they use and, for those that use DSpace CRIS, the specific version).

Type of implementation: via HTTP link headers

Example from IZTECH GCRIS:
Get the link information in the headers.

```
curl -IL https://hdl.handle.net/11147/13225
```

Response includes the link information it in the “Link” header with ‘rel=”item”:

```
HTTP/2 302
location: https://gcris.iyte.edu.tr/handle/11147/13225
expires: Wed, 30 Aug 2023 12:00:07 GMT
content-type: text/html;charset=utf-8
content-length: 173
date: Wed, 30 Aug 2023 11:58:27 GMT

HTTP/1.1 200
Server: nginx
Date: Wed, 30 Aug 2023 11:58:27 GMT
Content-Type: text/html;charset=UTF-8
Connection: keep-alive
Set-Cookie: JSESSIONID=1929CB9255203036DBCE49C754B4669B; Path=/; HttpOnly
Last-Modified: Sun, 16 Jul 2023 19:40:28 GMT
Link: https://doi.org/10.1080/15567036.2023.2171512; rel="cite-as"
Link:
https://gcris.iyte.edu.tr/bitstream/11147/13225/1/Air%20density%20calculation%20at%20high%20altitude.pdf; rel="item"; type="application/pdf"
Content-Language: en
Cache-Control: no-cache
Permission-Policy: push=deny
Cross-Origin-Resource-Policy: same-site
Cross-Origin-Opener-Policy: same-origin-allow-popups
Strict-Transport-Security: includeSubDomains
```

More information about the work of the EOSC Future Working Group on Research Product Publishing can be found in the wiki page of the WG.