



HAL
open science

Providing DDI metadata for OAI-PMH harvesting a Dataverse repository

Geneviève Michaud, Baptiste Rouxel

► **To cite this version:**

Geneviève Michaud, Baptiste Rouxel. Providing DDI metadata for OAI-PMH harvesting a Dataverse repository. 14th European DDI Users conference (EDDI 2022), EDDI, Nov 2022, Paris, France. 10.5281/zenodo.7529240 . hal-03936032

HAL Id: hal-03936032

<https://hal.science/hal-03936032>

Submitted on 12 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Providing DDI metadata for OAI-PMH harvesting a Dataverse repository

Geneviève Michaud, Baptiste Rouxel both CDSP (SciencesPo, CNRS)
14th European DDI Users conference, SciencesPo, Paris
28/12/2022

[DOI: 10.5281/zenodo.7529240](https://doi.org/10.5281/zenodo.7529240)

CDSP's research data repositories

The CDSP, a support unit for SSH researchers since 2005. As a CESSDA/Progedo partner, used Nesstar until 2020.

Responsible for the research data repository data.sciencespo (Dataverse) that serves:

- an institutional self-deposit collection
- the CDSP's Data Bank.

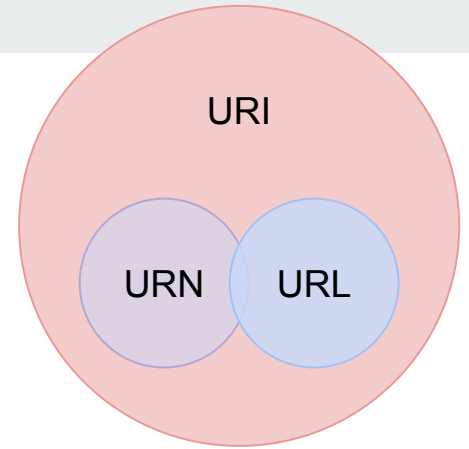
What is OAI-PMH?

- Open Archives Initiative **Protocol** for Metadata Harvesting
 - Generic and lightweight protocol
 - Only metadata are exchanged
- Standard for **interoperability**
 - Say-it-once approach
 - Several metadata harvesters (or reusers) can rely on a single data source
- A **wrapper** around a **metadata** <insert your **standard** name>
 - At the CDSP we use DDI, a rich and widely used standard, since 2006
 - We focused on OAI-PMH as a wrapper for DDI 2.5 metadata provided by Dataverse

Harvesting projects at the CDSP

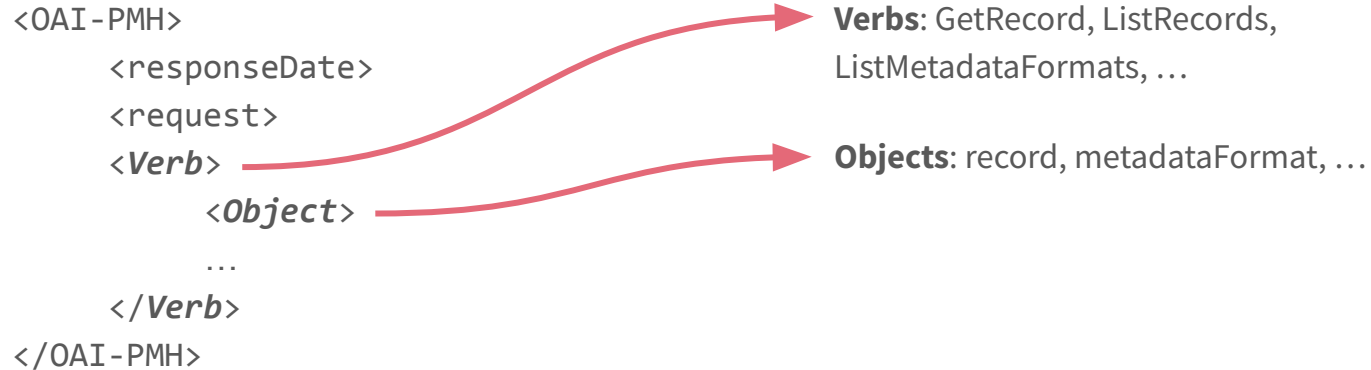
	Support Institution / Entity	Scope	Domains	Language(s*)
OpenAire	European Commission	all records	all	all
CESSDA DC	CESSDA	CDSP DataBank	social sciences	english
Isidore	RI* Huma-Num	CDSP DataBank	social sciences	french
Recherche Data Gouv	french ministry for higher education and research	all records	all	english, french

OAI-PMH unique identifiers and DOIs



- Every record must have a UID (Unique Identifier)
- Must follow the URI (Uniform Resource Identifier) format
- DOIs are widely used PID (Persistent and unique identifiers) for DDI records (for datasets)
- Since a URL (Uniform Resource Locator) is an URI, we can use DOIs in an URL format, for instance:
<https://doi.org/10.21410/7E4/UQ55HB>
- What a DOI allows / provides:
 - Link to a web landing page
 - Persistence (with curation)
 - Metadata ([Datacite schema](#))
 - ... FAIR

The structure of OAI-PMH



The structure of OAI-PMH

<OAI-PMH>

...

<ListRecords>

<record>

<header>

<metadata>

<about>

</record>

...

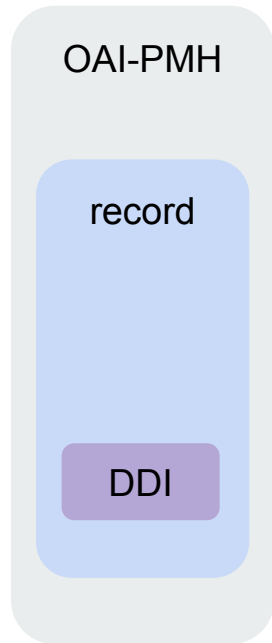
</ListRecords>

</OAI-PMH>

Contains information like the **unique identifier**.

Metadata can be of any format, generally Dublin Core. For example Dataverse supports multiple formats here, such as DC and **DDI**.





```

▼<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2022-11-22T13:36:16Z</responseDate>
  <request verb="GetRecord" identifier="oai:fsd.uta.fi:FSD0153"
    metadataPrefix="oai_ddi25">https://services.fsd.tuni.fi/v0/oai</request>
  ▼<GetRecord>
    ▼<record>
      ▼<header>
        <identifier>oai:fsd.uta.fi:FSD0153</identifier>
        <timestamp>2022-11-09T23:09:02Z</timestamp>
        <setSpec>language:en</setSpec>
        <setSpec>language:fi</setSpec>
        <setSpec>study_groups:wvs</setSpec>
        <setSpec>data_kind:Kvantitatiivinen</setSpec>
        <setSpec>data_kind:Quantitative</setSpec>
        <setSpec>openaire_data</setSpec>
      </header>
      ▼<metadata>
        ▶<codeBook xmlns="ddi:codebook:2_5" version="2.5" xsi:schemaLocation="ddi:codebook:2_5
          http://www.ddialliance.org/Specification/DDI-Codebook/2.5/XMLSchema/codebook.xsd">
          ...
          </codeBook>
        </metadata>
      </record>
    </GetRecord>
  </OAI-PMH>

```

OAI-PMH GetRecord sample from Aila (Finnish Social Science Data Archive):

https://services.fsd.tuni.fi/v0/oai?verb=GetRecord&identifier=oai%3Afsd.uta.fi%3AFSD0153&metadataPrefix=oai_ddi25

OAI-PMH DDI implementation in Dataverse

OAI-PMH and Dataverse

OAI-PMH server is available by default in Dataverse.

Out of the box feature that serves all records / datasets.

DDI metadata provided by Dataverse OAI-PMH incomplete and to a certain extent, flawed.

Harvesting metadata with OAI-PMH

Repository (Server) = OAI-PMH Data Provider

1. Identify which records to "serve" for harvesting by clients
2. Create a **set** containing the metadata
3. Communicate the set URL to clients or publish it openly
4. Handle the requests

Harvester (Client) = OAI-PMH Service Provider

1. Identify which records set is to be harvested on the repository
2. Obtain the set URL
3. Send a request to the given URL, specifying a **set** and a **verb**
4. Handle the response

record - **A record is metadata in a specific metadata format.** A record is returned as an XML-encoded byte stream in response to a protocol request to disseminate a specific metadata format from a constituent item

(in "The Open Archives Initiative Protocol for Metadata Harvesting, Protocol Version 2.0 of 2002-06-14, Document Version 2015-01-08, <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>")

Using OAI-PMH with Dataverse

1. Create a **set** (in Dashboard → Harvesting servers). You can specify filter criterias to feed your set, or include all datasets.
2. Run the export (index datasets in the set).
3. Make a request to the set with its URL and by specifying a metadata "flavor". Example:
https://data.sciencespo.fr/oai?verb=ListRecords&metadataPrefix=oai_ddi&set=CDSP.

Flavors:

- oai_ddi (DDI 2.5)
- oai_dc (Dublin Core)
- oai_datacite (Datacite)
- oai_json (JSON)

set name

CDSP collection set on data.sciencespo:

Edit Harvesting Set

Define a set of local datasets available for harvesting to remote clients.

Definition Query * ⓘ subtreePaths:/7
Example query: authorName:king

Name/OAI setSpec CDSP
The name can not be changed once the set has been created.

> Next Cancel

```

➡ ▼ <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2022-11-21T14:14:29Z</responseDate>
  <request verb="ListRecords" metadataPrefix="oai_ddi" set="CDSP">https://data.sciencespo.fr/oai</request>
➡ ▼ <ListRecords>
➡ ▼ <record>
  ▼ <header>
  ➡ <identifier>doi:10.21410/7E4/00LYOG</identifier>
    <timestamp>2022-10-28T00:00:03Z</timestamp>
    <setSpec>CDSP</setSpec>
    <setSpec>ALL_SCPO</setSpec>
  </header>
  ▼ <metadata>
  ➡ ▶ <codeBook xmlns="ddi:codebook:2_5" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="ddi:codebook:2_5 https://ddialliance.org/Specification/DDI-
    Codebook/2.5/XMLSchema/codebook.xsd" version="2.5" xml:lang="fr">
    ...
    </codeBook>
  </metadata>
</record>
▶ <record>
  ...
</record>

```

OAI-PMH ListRecords sample from data.sciencespo:
https://data.sciencespo.fr/oai?verb=ListRecords&metadataPrefix=oai_ddi&set=CDSP

Issues with OAI-PMH implementation of DDI in DV

While using OAI-PMH on Dataverse, we contributed to some issues that are now solved.

- Distributor tag automatically added in the OAI-DDI and DDI export [#7387](#)
- Internationalization
 - Language attribute missing in the OAI-DDI [#7388](#)
 - Include translation of a controlled vocabulary in the oai_ddi [#6751](#)
- No URI in OAI-PMH records (with oai-ddi metadata prefix) [#7786](#)

Benefits

- For a repository maintainer
 - **Control:** no compromise on your documentation: harvesters will get your DDI metadata as you intended
 - **Efficiency:** say it once
 - **Consistency:** up-to-date metadata are disseminated at no cost
- For re-users
 - **Discoverability and Reusability**
 - PIDs included in DDI metadata: Cross-references between all research outputs
 - explorable research graph**
 - Data source still one click away



Shared responsibilities

- From the metadata "client" (harvester) side
 - The OAI-PMH **protocol** must be supported by the harvester,
 - The metadata **format** encapsulated in the record must also be supported,
 - The harvester should understand the metadata **version** used by the repository.
- From the metadata source
 - DDI metadata should be **valid!**
 - Useful tool provided by [CESSDA metadata validator](#)
 - Each harvester may bring additional requirements (i.e. [CESSDA metadata profiles](#))

Thank you

Get it touch: itcdsp-scpolst@sciencespo.fr