EOSC Technical Specification

Common Services

Metadata Cataloguing and Indexing

Version:	1
Status:	EOSC-hub Proposal
Dissemination Level:	Public
Document Link:	

Abstract

The EOSC Metadata Cataloguing and Indexing service comprises the management of metadata in the whole life cycle from generation up to uploading and indexing metadata in a searchable catalogue.

OSC-HUB RECEIVES FUNDING FROM THE EUROPEAN UNION'S HORIZON 2020 RESEARCH AND INNOVATION PROGRAMME UNDER RANT AGREEMENT NO. 777536.

COPYRIGHT NOTICE

DELIVERY SLIP

Date	***Name***	***Partner/Activity***
From:	Heinrich Widmann	DKRZ
Reviewed by:	Bartosz Kryza Sara Ramezani	CYFRONET SURFSARA

DOCUMENT LOG

Issue	***Date***	***Comment***	***Author***
v.1	24/01/2020	First release ready for public consultation	Heinrich Widmann (DKRZ)

TERMINOLOGY

https://wiki.eosc-hub.eu/display/EOSC/EOSC-hub+Glossary

Terminology/Acronym	***Definition***

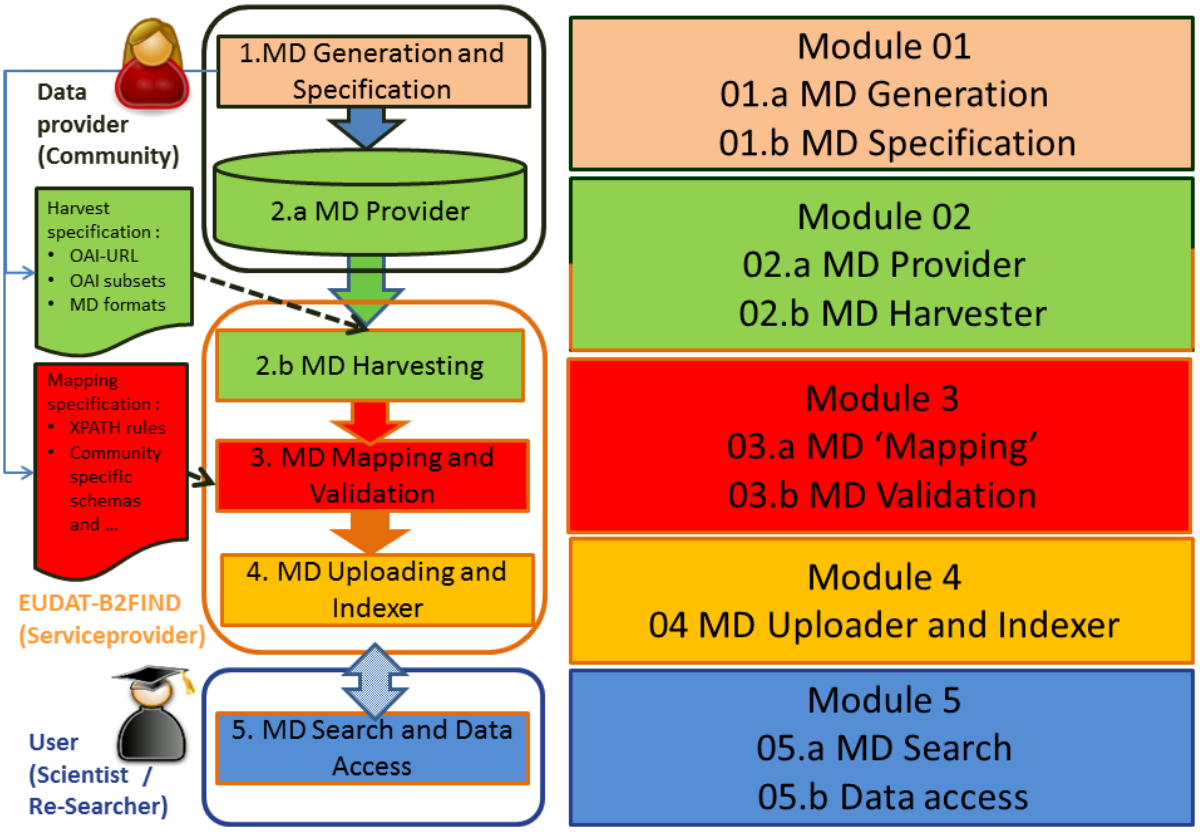
Contents

1 Introduction

Metadata Cataloguing and Indexing comprises the entire metadata ingestion workflow, i.e.

1. Metadata harvesting from community repositories
2. Metadata mapping on common schema including curation and validation and
3. Uploading and indexing of metadata records in the metadata catalogue, to enable Data Discovery and Access , see related macro feature

2 High-level Service Architecture



The technical implementation of metadata cataloguing usually comprises five modules as shown in the figure below:

1. In the (Meta)data Provider Module, metadata must be available and harvestable in a known metadata schema and format. It also should be harvestable and accessible by a standardised transfer protocol (e.g. OAI-PMH).
2. For sustainable metadata ingestion synchronous and incremental harvesting should be set up on the service provider site.
3. On the service provider site, normalisation, homogenisation and mapping of the specific community standards onto a generic, common and unified metadata schema should be performed. The metadata mapping should be adopted to the needs of data provider and should include metadata validation and curation.
4. Finally, the mapped records are uploaded into the central metadata catalogue and indexed to allow faceted search in the discovery portal.
5. This enables now end users to search and filter datasets via the GUI or by using a command line tool and then access the found data resources.

3 Adopted standards

Standard	Short description	References
Community specific metadata schemas and standards	Central, cross-domain Metadata aggregators collect community specific formatted metadata. For instance B2FIND supports harvesting of multiple metadata formats (as XML,MarcXML, JSON) and schemas (e.g. DataCite, Dublin Core, ISO 19115, CMDI, DDI and others).	A list of some domain specific metadata standards can be found at http://b2find.eudat.eu/guidelines/providing.html#mdformats
DataCite Metadata Schema 4.1.	Common and widely-used Metadata Schema, on which e.g. OpenAire and EUDAT- B2FIND is based.	https://schema.datacite.org/meta/kernel-4.1/ http://b2find.eudat.eu/guidelines/mapping.html#b2fmdschema
Controlled Vocabulries	E.g. ISO 639-1 codes are a standardized nomenclature used to classify languages, or EUDAT-B2FIND develops a standardised taxonomy for 'Resaerch Disciplines' , which specifies the research disciplines	https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes https://cryptpad.fr/pad/#/1/edit/KDecbjauKCtZclOmZAbbWg/L4aEiGrzJlSbRSXrFutOb0Cd/ http://clara.science/

Protocol/API	Short description	References
OAI-PMH	The Open Archives Initiative Protocol for Metadata Harvesting provides an application-independent interoperability application to collect metadata from repositories.	http://www.openarchives.org/OAI/openarchivesprotocol.html
ResourceSync	This ResourceSync specification describes a synchronization framework for the web consisting of various capabilities that allow third-party systems to remain synchronized with a server's evolving resources.	http://www.openarchives.org/rs/1.1/resourcesync
REST API's	A full REST API is used to collect metadata formatted as JSON, e.g. the referenced REST API is used to 'harvest' from Herbadrop's repository	https://helpdesk.eudat.eu/Ticket/Attachment/122586/63597/RESTAPI_HowTo_SearchUserGuide_V3.pdf
CSW / OGC	Catalogue Service for the Web (CSW) is used to collect metadata from OpenGeoSpatial atalogues (OGC)	http://www.opengeospatial.org/standards/cat

4 Interoperability guidelines

In general the preconditions to publish metadata should be clearly described and stated by the discovery service provider in 'Guidelines for data providers', as in e.g. guidelines of OpenAire (https://guidelines.openaire.eu/en/latest/data/index.html) or of EUDAT-B2FIND (see http://b2find.eudat.eu/guidelines). This allows not only research communities, but also generic data storage repositories and metadata aggregators to make their data searchable in a simple way by following the guidelines.

In particular, EUDAT-B2FIND follows the FAIR principles and the rules of 'Open Science' to promote interoperability in research data management.

5 Examples of solutions implementing this specification

Examples of cross-domain discovery services using this approach are

- GoogleDataset Search (https://toolbox.google.com/datasetsearch), which crawls mainly schema.org , but supports no specific (meta)data curation and validation and does not consider on open data access (so 'dark data' is not necessarily excluded and it does not conform to FAIR data principles)
- EUDAT-B2FIND (http://b2find.eudat.eu/), the central indexer of EOSC-hub, provides an interdisciplinary discovery portal for research data with faceted search and comprises extensive meta(data) mapping, validation and curation in a FAIR manner

5.1 Procedure to integrate a service with the EOSC Hub Metadata Cataloguing and Indexing

To provide metadata to the Metadata Cataloguing and Indexing service, the following preconditions must be fulfilled:

1. provider server must be set up (e.g. OAI-PMH provider)
2. Metadata must be provided in a standardised format and schema and made available and accessible for harvest requests and some mandatory fields (e.g a title and data identifier) must be provided.
3. In the next stage, refinement and enrichment of the metadata is done iteratively.