Polar Data Ecosystem Mapping Data Model

Date: 2024-02-15

Version 1.1, 2024-02-15: Use Cases for application added

Versions

Version	Date	Comment	Responsible
1.1	2024-02-15	Added use cases for development of application	Øystein Godøy, Jan Rene Larsen, …
1.0	2023-09-24	First public version	Øystein Godøy
0.1	2023-02-16	First consolidated draft.	Øystein Godøy, Brendan Billingsley
0.0	2022-10-18	First draft for internal discussion.	Øystein Godøy

Table of Contents

1. Introduction

1.1. Background

The Mapping the Polar Data Ecosystem Project grew from the Mapping the Arctic Data Ecosystem Project that was initiated by the Arctic Data Committee in November of 2014. Since then, the project has eveloved to include multiple themes using a series of tools. One of those themes is focused on understanding how the community shares metadata. During the Polar Data Planning Summit [https://arcticdc.org/meetings/conferences/polar-data-planning-summit] which was arranged May 22 - 24, 2018 in Boulder, Colorado, USA, an effort was initiated to map the ecosystem of Polar data centres and how these interacted exchanging discovery metadata. This effort was further elaborated in the Polar Data and Systems Architecture Workshop [https://arcticdc.org/meetings/conferences/polar-data-architecture-workshop] which was arranged November 28-30, 2018 in Geneva, Switzerland. These two meetings resulted in a spreadsheet describing metadata sharing interactions between data centres (i.e. who was harvesting information from whom), information which interoperability standards that were used etc.

This spreadsheet has since been developed in the joint working group on Federated Search for Polar Regions [https://arcticdc.org/activities/core-projects/federated-search] and in joint the Polar Semantics Working Group [https://arcticdc.org/activities/core-projects/vocabularies-and-semantics-wg]. Different aspects have been emphasised in these groups.

In order to simplify analysis of the content of the spreadsheet, it was ingested into a graph database and an application for analysis of relationships was developed (Figure 1). This application has evolved through a number of projects and now is in the progress of supporting multiple use cases on interactions between data centres in the Polar regions. As the number of use cases have increased and the spreadsheet has grown in size it has become evident that managing this information in a spreadsheet is not sustainable. In order to have a solid basis for further development of the mapping activity and furthermore to move maintenance of information into a suitable system, it was decided to develop a data model for the application. This data model should support a wide variety of use cases as well as also support efficient and transparent maintenance of the information collected. In the process of establishing this data model it has been decided that the content of the mapping activity also shall be available under a CC-BY-4.0 license.

Figure 1. Screenshot of the second version of Mapping the Polar Data Ecosystem (MPDE) application. The third version is now available at https://mpde.gcrc.carleton.ca

1.2. Scope

This document describes the data model and provides explanations to the various elements.

1.3. Intended audience

The intended audience is people interested in the mapping activity and development of the federated search/unified data catalogue for polar data.

2. Data model

2.1. Introduction

The purpose of the data model is to ensure that the necessary information to establish a proper understanding of relationships between data centres, interoperability technologies used and the evolution of relationships is captured in a consistent manner. The data model should help the community to answer questions like (non exhaustive list):

Which data centres manages polar data?

Where are polar relevant data managed and how are these data made available to the community?

Which data centres acts as aggregators?

Some data centres aggregates information from other data centres to provide services to their user communities. Which are these data centres and how do they interact?

Which interoperability technologies are used?

Which technologies are actively used to exchange discovery information, but also to convey information about the data.

Which information containers are used for discovery level information??

Exchange of discovery metadata is the core functionality supporting federated search or unified virtual data catalogues. Which information containers are used as well as which semantic annotation is used to describe the contents of data (e.g. variables) and exchange mechanisms (e.g. purpose of URL's in discovery metadata).

2.2. Diagram

The data model for the polar data ecosystem mapping activity is outlined in Figure 2.

analysisPurpose

The core element in the data model is the "analysisPurpose". This should indicate that the underlying data are suitable for answering a number of use cases. Each `analysisPurpose` will contain 1 or more `mappingElement`. Currently the following keywords have been defined:

Active data centres

These are the data centres managing polar data.

Interaction between data centres

These data centres acts as aggregators in the ecosystem of data centres.

Technology usage

These interoperability technologies are actively used for either discovery or access/retrieval purposes.

Other

Other use cases, not strictly specified.

mappingElement

Each `mappingElement` describes a relation between data centres. In order to track the information over time, the element has a number of attributes. A unique internal `identifier` which is used to separate `mappingElement` elements from each other. The time this information was last updated in `dateUpdated`, this information is captured following ISO8601. And name and email of the person providing the updated information in `authorName` and `authorEmail`. Each `mappingElement` contain 1 `aggregatorRepository` element.

aggregatorRepository

This describes a repository that is harvesting information from other data repositories. For each repository the name, url and country (respective attributes are `name`, `url` and `country`) is needed to support the top level use cases. The `url` should be the landing/front page of the repository. Each `aggregatorRepository` element will have 0 or elements of type `harvestProfile` that describes the relations and technologies used. Each `harvestProfile` of an `aggregatorRepository` is linked with a `sourceRepository`.

harvestProfile

harvestStatus

States whether the repository is actively harvested, has been in the past or is in planning. See Figure 2 for details.

harvestFrequency

How often are the information retrieved by the aggregator. See Figure 2 for details.

harvestEviction

Describes how often the aggregator evict all information harvested and do a clean full harvest again. See Figure 2 for details.

harvestProtocol

Identifies the interoperability protocol used to exchange information on discovery metadata between the source and the aggregator. The link between a aggregator and a source is only one protocol, but a source may be harvested using different protocols by different aggregators. See Figure 2 for details.

harvestMetadata

The discovery metadata standard the aggregator is retrieving from the source. See Figure 2 for details.

harvestContent

Indicates if incremental or full harvests are used. This should be interpreted in combination with `harvestFrequency`.

harvestNotes

Any comment added as free text supporting the understanding of the relation between the aggregator and the source.

dataAccess

Indicates which mechanisms an aggregator is using to integrate the data provided by the source in the service portfolio of the aggregator. Typically this can be multiple elements for each aggregator/source linkage. This will be a list.

lastValidated

Information on when the current content of the harvesting was confirmed by the aggregator. Also using ISO8601.

sourceRepository

This describes the data repository that is being harvested by the `aggregatorRepository`. The information provided for this element is the same as for the `aggregatorRepository` (i.e. name, url, country), but the information on metadata standards, protocols etc in `harvestProfile` reflects the services offered by the `sourceRepository` as they are actively utilised by the aggregator.
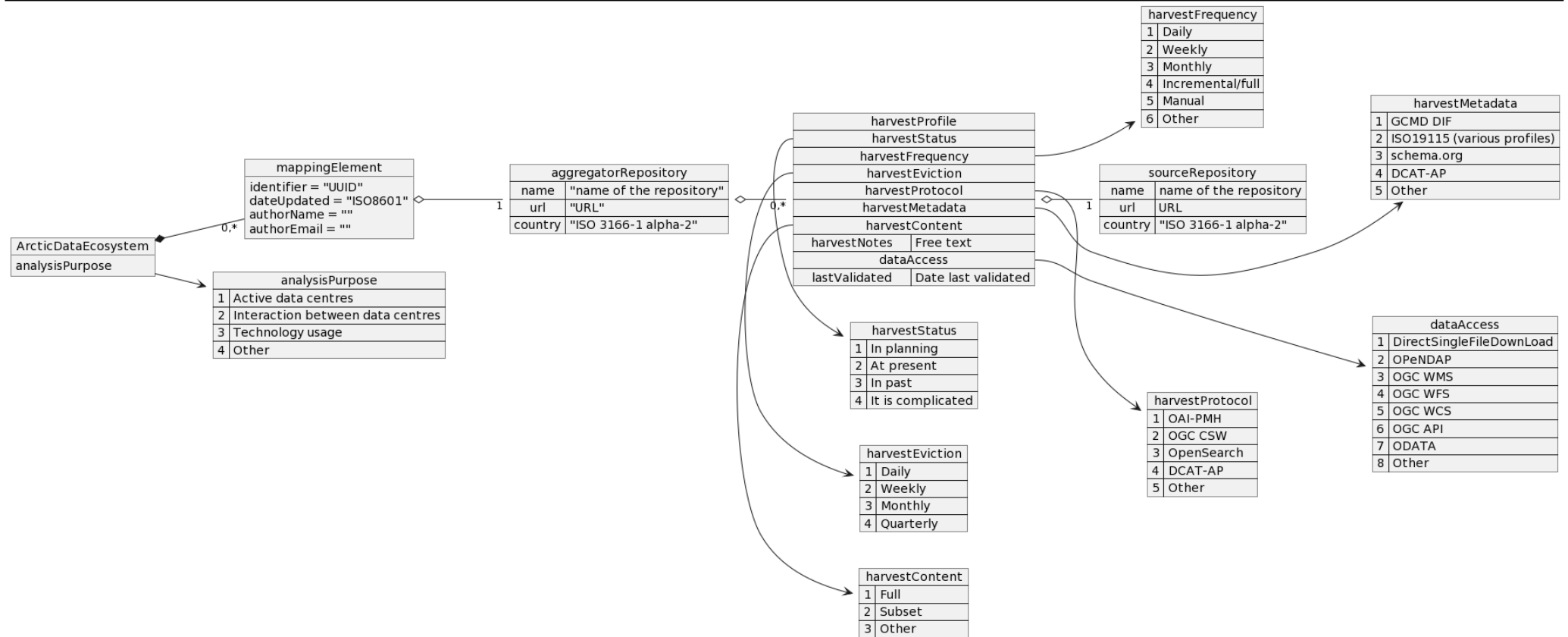
Figure 2. Data model for the polar data ecosystem mapping activity.

3. Use Cases

3.1. Introduction

The purpose of this section is to provide use cases that supports development of the Mapping the Polar Data Ecosystem [https://mdpe.gcrc.carleton.ca] (MPDE) application. Initially the focus is on use cases helping the community to maintain and improve the content of the application while in the longer term it will contain descriptions focusing on utilising the application for specific purposes.

3.2. Specific use cases

Table 1. How to manage the MPDE.

Name	Management of the MPDE recors and users.
Goal	To enable selected MPDE editors to manage users and data centres in the MPDE application. This has to be a controlled process in order to maintain a consistent list of data centres and their relations.
Actors	• MPDE editor • MPDE system manager • Data centre responsible
Pre-conditions	1. The MPDE system manager has access to configuration of the MPDE 2. The MPDE editor has access to configuration of users and data centres in the MPDE
Post-conditions	1. The MPDE editor has enabled a data centre responsible to edit records for a specific data centre. 2. The MPDE editor has approved/modified requests for new data centres to be added to the system.
Normal flow	1. A user requests to become responsible for a specific data centre. 2. The MPDE editor approves or rejects the request. 3. The data centre responsible requests a relation to a new data centre be established. This is approved or rejected by an MPDE editor.
Notes and issues	1. New data centres has to be approved by an MPDE editor. This also holds for data centres that are added by existing data centre responsible that need to set up new relations. 2. MPDE editors are approved by MPDE system manager.
Last updated	FIXME
Last updated by	FIXME

Table 2. How to manage information in MDPE.

Name	Management of information in the MPDE application

Goal	To enable responsible and authorised users to add and maintain information on data centers and their relations in the MPDE application. This implies adding a new data centre, setting up relations to existing data centres and adding new data centres previously not listed.
Actors	• Responsible for a data centre • MPDE editor
Pre-conditions	1. The data centre responsible have an account in the MPDE application. 2. The data centre responsible is authorised to edit records for a specific data centre.
Post-conditions	1. The data centre record is updated with existing and new relations and interfaces.
Normal flow	1. The data center responsible authenticates and is authorised to edit records for a specific data centre. 2. Existing relations are modified as necessary through modification of technical interfaces etc. 3. Potential new relations are added, but integration is pending approval of an MPDE editor. 4. Once new additions are approved the new records are fully integrated and visible for everyone.
Notes and issues	1. New data centres has to be approved by an MPDE editor to ensure consistent naming conventions/controlled vocabularies. 2. Relations between data centres should in the mid to long term be drag and drop relations, but in the short term a tabular overview/edit would be sufficient.
Last updated	FIXME
Last updated by	FIXME