# Goal and scope

This report presents the findings from the evaluation of the Core Metadata Schema within the scope of Plateau 1.

Evaluation method

The evaluation method involved a Microsoft form featuring one question per class, along with a thorough review of each property within these classes. This approach aimed to gather detailed feedback on the schema's applicability and identify areas for enhancement.

# Results

Responses were received from 7 entities, including the hub and 6 nodes, with absences from Maastricht UMC and other entities.

Only two nodes submitted their responses as a group, a collaborative approach recommended on the Co-Creation Day.

# Evaluators:

|  |
| --- |
| Mark Janse, Research Software Engineer Imaging, [mark.janse@health-ri.nl](mailto:mark.janse@health-ri.nl) |
| Jutta de Jong, Sarita Hartgring, Jet Zoon, UMC Utrecht, Prinses Maxima Center, [j.r.dejong-35@umcutrecht.nl](mailto:j.r.dejong-35@umcutrecht.nl), [s.a.y.hartgring@umcutrecht.nl](mailto:s.a.y.hartgring@umcutrecht.nl), [j.zoon-3@prinsesmaximacentrum.nl](mailto:j.zoon-3@prinsesmaximacentrum.nl) |
| Viola Woeckel, Erasmus MC, FAIR Coordinator, [v.woeckel@erasmusmc.nl](mailto:v.woeckel@erasmusmc.nl) |
| Joost Daams, AUMC, RDM consultant & developer, [j.g.daams@amsterdamumc.nl](mailto:j.g.daams@amsterdamumc.nl) |
| Rene Oostergo, [r.j.oostergo@umcg.nl](mailto:r.j.oostergo@umcg.nl) |
| Anna Niehues, LUMC (Advanced Data Management, Department of Biomedical Data Sciences), Senior Data Steward / FAIR coordinator LUMC Health-RI node, [a.niehues@lumc.nl](mailto:a.niehues@lumc.nl) |
| Inês de Oliveira Coelho Henriques (Rumc, Data Steward, [ines.oliveiracoelhohenriques@radboudumc.nl](mailto:ines.oliveiracoelhohenriques@radboudumc.nl)); Peter-Bram 't hoen (Rumc), Bruna dos Santos Vieira (Rumc, Data Steward), Pim Kamerling (Rumc), Alexandros Papadopoulos (Rumc, Data Steward), Milou de Jong (Rumc, Data steward), Daniela Vicentini Albring (Rumc, Data Steward), Junda Huang (Rumc), Xiaofeng Liao (Rumc) |

**Classes**

The feedback pointed to Catalog, Dataset, Distribution, and Project as the most relevant classes. While the usage of resource, dataseries and dataservice and agent remain unclear.

A graph with different colored bars

Description automatically generated

**Discussion**

Question 11 – What type of Research Resource do you work with (produce/reuse)?

A screenshot of a chat

Description automatically generated

General, for this question only one answer reflected on which types of resources the organisation was using or missing in the model.

Research resource is an abstract concept and not clearly defined for instance we consider any research object that is being reused or produced throughout the research cycle. A research resource can be a publication, dataset, sample, code (packages), images, ... .

To do: Update the documentation to address this and make it clear what is a research resource by adding more examples.

How to accommodate Resources that are not datasets: samples (unclear if physical samples are resources or are a theme of a dataset), patient registries, biobank registry... resource uses the term rdfs:resource is this true?

To-do: change rdfs:resource to dcat:Resource

To-do: add Research Resources:

core:PatientRegistry

core:BiobankRegistry

core: Sample

Todo: check for additional typos with capital letters across the confluence: e.g., dcat:dataseries, dcat:catalogue)

Question 12

A screenshot of a computer

Description automatically generated

Several Catalogues and systems were mentioned, Dataverse was mentioned twice.

No catalog

Domain catalog (xnat,

General catalogs (DANS, DataVerse)

Unclarity on DCAT vs FDP interface implementation perceived hierarchy between Catalog and Dataset. It is important to follow the DCAT standard relation of Catalog being a subtype of Dataset, independently from the user interface. Catalog is actually a dataset of datasets. We are currently unsure of the relevance and if this is automatically fixed by FDP implementation

Todo: ask Kees if this “hierarchy” is relevant when implementing or is something automated.

Question 13

A screenshot of a computer

Description automatically generated

To-do: decide how to use Theme vs Type vs Data type. Do we keep it general like FDP does, anything can be a Theme? And the rest can go under Type? Or we make it specific. Example:

Vasca registry (dcat:Dataset, core:PatientRegistry), has survey data and lab data (dcterms:type), has synthetic data (dcterms:type), the registry is about vascular anomalies (dcat:theme), the registry focuses on “lower limbs vascular anomalies” (dcat:theme), etc

To-do: We need to model the taxonomy (poly hierarchy) to define all these concepts. For this we need a focused group enabled with a tooling to be able to define all these hierarchies, vocabularies and link them to the classes in the core model.

Question 14

What type of distribution so you work with?

A screenshot of a computer

Description automatically generated

To-do: cater for Biobanks that are a Catalog/collection / bundle of dataset, distribution level appears to not be relevant, need to figure out how biobank search works. Refer biobank to EU standard of Distribution (add link) see if their distribution can be found on the list.

To-do: physical embodiment: fix documentation on Distribution definition. --> is fixed

Question 15

What type of Dataset Series do you work with (produce and/or reuse)?

A screenshot of a computer

Description automatically generated

Data series is still an abstract concept. Data series can be considered a bundle of dataset meaning a dataset can be comprised of multiple datatypes multimodal. Dataset is part of a study and a study can have mutually exclusive samples and observations.

Usage example from the EU:

https://joinup.ec.europa.eu/release/dcat-ap-how-model-dataset-series

To do:

Clarify what data series mean and provide examples.

The example that EU brings for the usage of the data series is simple and understandable but how can it be used for health domain. Question 16

Question 16

What type of Data Service do you work with?

Data service must be further specialised into several types of service.

To do: Model data service

Question 17

What type of Project do you work with?

A screenshot of a computer

Description automatically generated

Collect controlled vocabulary for types of projects at various levels of representations:

EU projects. National, international, clinical research, Panama registered project

Question 18:

What type of Agent do you work with?

A screenshot of a computer

Description automatically generated

To do:

Define diverse types of organizations vocabularies, agents

The open discussion is about Using Vcard for the contact point. NEEDS TO BE ADJUSTED IN THE MODEL.

Explain diverse types of roles that is modelled in the schema (publisher, creator)

**Attributes**

**Dataset**

|  |  |  |
| --- | --- | --- |
| Evaluated attributes | remarks | To do |
| relation | Clarity is needed | Change from mandatory to optional |
| type | What are vocabularies in this schema? | Make the definition explicit, |
| theme | What are the vocabularies in this scheme? | Make the definitions explicit, Collect vocabs |
| modified | Why is this mandatory? |  |
| Contact point | Dcat uses vcard,  Semantically maybe it is better to use foaf:agent, for better interoperability we shall follow dcat v3 | The range of contactPoint must change to Vcard |

**Data** **Service** --> for this we should align with the request requirements from Tasneem

|  |  |  |
| --- | --- | --- |
| Evaluated attributes | remarks | To do |
| Endpoint URL |  | Make it optional? |
|  |  |  |
|  |  |  |

**Distribution** -> for this we should align with the request requirements from Tasneem

|  |  |  |
| --- | --- | --- |
| Evaluated attributes | remarks | To do |
| **media type** | What vocabulary? |  |

**Project**

|  |  |  |
| --- | --- | --- |
| Evaluated attributes | remarks | To do |
| relation | It is not explicitly named and defined? | This is a relation from dataset to it project and vice versa, so let ‘s just make the predicate explicit in the naming Such as class (Project) Object Property(hasResource) class (Dataset) |
| Grant number | Is there a standard way of representing grant number? |  |
| Grant agreement | Model for this? |  |
| Project number |  |  |
| acronym |  |  |
| Call number (related to proposal) |  |  |
| Project website |  |  |
|  |  |  |

# Conclusion and Recommendations and tasks

Regarding collaborative modelling and tooling:

Tooling is missing – We do not have a place to define our model, and vocabularies (taxonomies) linked to it where we can collaboratively extend with our community.

See the vendor analysis here

[**https://healthri.sharepoint.com/:w:/s/hri-team012/EY0pHJYYRvVMmFnmVL7zWewBawxFig60sDq3wX0Jt4B0bw?e=JBecog**](https://healthri.sharepoint.com/:w:/s/hri-team012/EY0pHJYYRvVMmFnmVL7zWewBawxFig60sDq3wX0Jt4B0bw?e=JBecog)

**Modifications** **to** **the** **current** **model**:

|  |  |  |
| --- | --- | --- |
| Requirements | Plateau 1 | tasks |
| Resource types | P1 | Specify more research resources,  add properties like FDP-O <https://specs.fairdatapoint.org/fdp-specs-v1.2.html> |
| data service | NOT FOR P1 | Specialise it more Tasks(work with Tasneem data request requirement) |
| Data type | P1 (definitions and examples for P1 and document the decision log) | Define it explicitly and collect vocabularies (data type vs theme) |
| Data theme | P1 (definitions and examples for P1 and document the decision log) | Define it explicitly and collect vocabularies (data theme vs type) |
| contact point | P1 | Range points to vCard |
| Data series | Non P1 | Definition, example |
| Dataset | P1 | We made an issue out of this on git https://github.com/Health-RI/health-ri-metadata/issues/45  -- Further specify dataset subclasses, example:  core:patientRegistry  core:BiobankRegistry Model the sample and data (make the relation)      specify relation to project (rename it explicitly) |
| Project | P1 | Types(EU, National) , grant info, acronym, project number |

# TASKS: Collecting new Requirement

New requirements are grouped into p1 and non p1 categories to be address in line with plateau planning. From the form, the following requirements were identified:

Plateau 1 requirements / Modification

# Other tasks:

Update documentation