|  |  |
| --- | --- |
| **Author** | **Dena Tahvildari** |
| **status** | **proposed** |
| **summary** | **A modularization approach to represent the Health Research Metadata Knowledge graph** |
| **participants** | **Dena, (and shortly this proposl is discussed with Luiz Bonino and got some feedback in December 2023)** |
| **Decision date** |  |

**Why do we want to modularize?**

Modularization is a generic concept that is intuitively understood as referring to a situation where simultaneously a thing (e.g. an ontology, metadata schema) exists but can also be seen as a set of parts (the modules).

So, in our context which we are building a semantic layer for onboarding research data to a national catalogue, we have a big metadata schema which enables and standardise this integration. But this big schema is a living thing, and it evolves and scale over time and contain various requirements (functional, domain and system specific). So, it needs to be decomposed in separate, interlinked modules.

**Why are the benefits of modularization?**

* **Scalability for querying data**
* **Complexity management**
* **Understandability**
* **Reuse**

**And so on**

For this, Dena argues that each part of the health-ri metadata Knowledge graph **(HRI-MKG)** should be organized as a module.

This means the part of the health-KG in question has its status, plans, and ownership documented according to a set template.

A module is:

* self-contained
* can develop on its own
* serving at least one-use case (one domain)
* can be a knowledge domain.
* or an organizational/technical unit

A module has:

* one ontology for defining the domain (as a good practice this one better be in rdf/owl
* one or more SHACL (Shapes Constraints Language) definitions for defining the constraints and rules that must be applied to the definition above.
* 0 to many taxonomies (defining range constraints)
* 0 to many other manually defined instances (controlled vocabularies)

**For example**

* + **Imaging module contains:**
    - **Imaging.ttl --> containing domain model and representation**
    - **ImagingRules.ttl.shape --> containing rules and constrains for validation and integration purposes**
    - **Vocabularies (conceptSchemes)**
      * **Bodyparts.ttl**
      * **ImageModality.ttl**
      * **TreatmentDoagnosis.ttl**

**Or Omics Module**

**On Git: A module is reflected as a folder in the Health-RI git (to do later)**

**Naming and IRI convention of a module:**

* + **The knowledge area/metadata module name is included in the IRI**
  + **Each module has their own prefix**
  + **Each module is either an extension to the ontology, a taxonomy, part of the data graph, or other instances.**
  + **To be able to navigate to HKG well also when it grows, distinct parts of it will have a technical identifier that is also reflected in the global part of the IRI**

**Below is a list of existing modules (not a Complete list)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **MODULES NAME** | **DESCRIPTION** | **WORKING GROUPS (PEOPLE INVOVLED)** | **URI namespace** | **prefix** | **Related shacls** | **Related taxonomies** |
| **Core** |  |  |  |  |  |  |
| **health** |  |  |  |  |  |  |
| **Imaging** |  |  |  |  |  |  |