

New version of the DMP Common Standard Ontology

João Cardoso (RDA Portugal), Fajar J. Ekaputra (TU Wien), Leyla Garcia (ZB MED Information Centre for Life Sciences), Christine Jacquemot (DMP OPIDoR / Inist-CNRS France)

The development of the [DMP Common Standard Ontology \(DCSO\)](#) is an ongoing effort intended to provide semantic technology representation of the DCS application profile.

The objective of this Issue was to create a new version of the DCSO that complies with the best practices in the community. For this hackathon there were 3 main sub-issues to tackle: (1) Third-party concept integration; (2) controlled vocabularies; and (3) constraint representation.

Third party concept integration

The DCS application profile contains terms originated from the following three ontologies: (1) the [W3C DCAT specification](#); (2) the [DCMI Metadata Terms](#); and (3) [FOAF Vocabulary Specification 0.99](#). When the first version of DCSO was developed there was a deliberate choice not to establish import relations or reuse terms from these ontologies. This is however not a best practice and discourages linking with other ontologies. Therefore one of the sub-issues was to correct this by adding the necessary imports, and propagate any changes that needed to be made as a consequence of that.

The outcome of this sub-issue was the creation of the **DCSO Core**, as seen in Figure 1. The decision to create the DCSO Core is heavily linked to the other two sub-issues. The intention was having an ontology that would solely represent the core elements of the DCS application profile, whilst reusing appropriate classes and properties from existing ontologies.

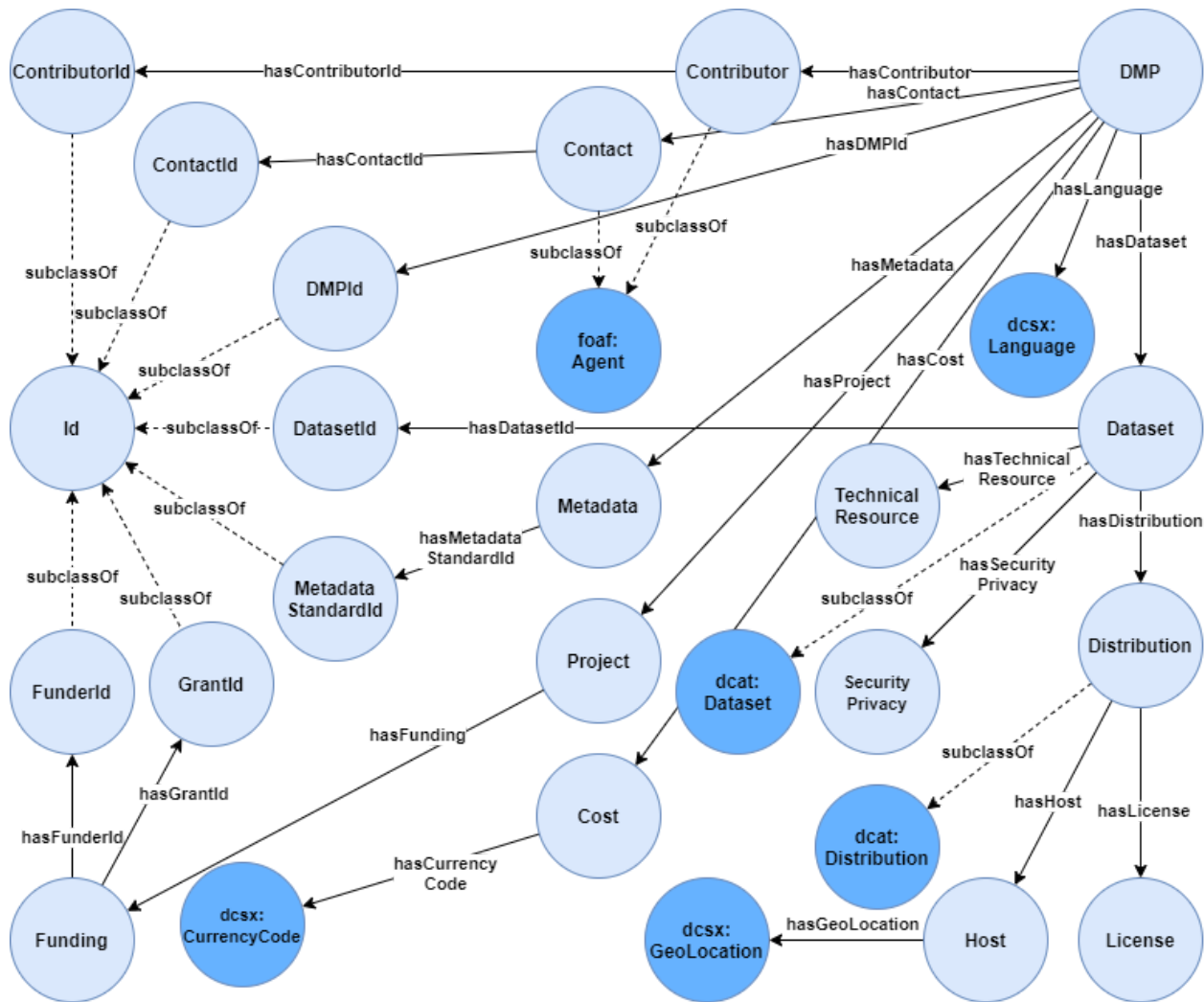


Figure 1 - DCSO Core Diagram

There were three key decisions that were made in the process of creating the DCSO core that are relevant to mention:

- The `dcso:Contact` and `dcso:Contributor` classes were considered as subclasses of the `foaf:Agent` class, as both of these classes can represent people.
- Similarly, all the Identifier classes, e.g., `dcso:ContributorId`, `dcso:DMPId`, `dcso:DatasetId`, etc., were considered subclasses of a newly created `dcso:Id`, for they all share the same data properties, i.e., identifier and type, with only slight differences in the values that are accepted for each data property and their property domains.
- Finally, it was decided that any data properties that would require controlled vocabularies would not feature in the DCSO Core.

Controlled vocabularies

Some of the terms in the DCS application profile must abide by controlled vocabularies. For example, in the Host class, the `geo_location` field describes the physical location of the data expressed using ISO 3166-1-alpha2 country code. The original solution for this issue was to create dedicated Datatypes in the DCSO. This solution was not ideal, for it left up to the developers of the DCSO the burden of updating the Data types, and it added complexity to the DCSO by having multiple custom defined Data types defined. A sub-issue was defined to provide an alternative solution in line with best practices.

Initially, we identified the fields that should be considered for a controlled vocabulary representation, opting for the inclusion only of those coming from standard controlled vocabularies. As such, the following fields were selected: [geo_location](#), [currency_code](#) and [language](#). Any other fields not compliant with a standardised controlled vocabulary will be handled by the constraint representation sub-issue.

The selected fields were represented as instances of one of the newly created ontology classes: `dcso:GeoLocation`, `dcso:CurrencyCode` and `dcso:Language`. These classes are not considered part of the DCSO core; therefore, a second ontology was created, namely [DCSO Extension \(DCSX\)](#). We introduced new object properties in DCSO to express relations between its classes and the classes in the DCSX, which became an imported ontology in DCSO.

Constraint representation

The DCS Application Profile provides a set of constraints to check whether an maDMP instance is compliant to the standard. We support such constraints in DCSO via a set of ontology shape constraints. There are currently two main languages to define shape constraints for ontology instances: [SHACL](#) and [ShEx](#). Validation approaches via ShEx or SHACL make it easier to identify possible errors on the way classes and properties from DCSO are used.

In this hackathon, we defined a set of basic ShEx constraints for a number of DCSO classes, including `dcso:DMP`, `dcso:Contact`, `dcso:Contributor`, `dcso:Cost`, `dcso:Project` and `dcso:Funding` based on the DCS Application Profile. We used tools such as [RDFShape](#) and the [Simple Online Validator](#) to test DMP examples with and without constraint errors to make sure that the defined constraints work as intended.

Conclusions and Future work

In addition to the initially defined issues, others were identified during the course of the hackathon. These were not added to the hackathon working plan, due to the short time frame available. The identified issues are as follows: (1) Conversion from a JSON representation of the DCS application profile to the DCSO; (2) Semantic validation using the ShEx constraints; (3) Using SPARQL queries to query the DCSO. SPARQL queries were built to answer questions provided through the [user stories](#) initially collected by the WG. However, as only one example

was available at the time. It has not been possible to test all the different queries. This issue is to be continued once the conversion of DMP from JSON to RDF is achieved. Thus a larger set of maDMP graphs will be available; (4) Analysing the DCS application profile for possible extensions to the DCSX controlled vocabularies; (5) Exploring the possibility of collaboration with Issue 6; and (6) Integration of the changes to the DCS application profile proposed by Issue 14.

The hackathon team members have already shown availability to address these issues in the coming months. Additionally the team members have also committed to submit a long paper to the [First Workshop on Research Data Management for Linked Open Science - DaMaLOS](#), that is to be held in association with the [19th International Semantic Web Conference \(ISWC 2020\)](#).