

StOWG_RFC_6: Building Ontology Slims for CFDE

CFDE-Draft-# (RFC Admin Only):	CFDE-DRAFT-10		
CFDE-Draft-Title:	Building Ontology Slims for CFDE		
CFDE-Draft-type: (Right click the box and select checkmark ✓ CTRL - Z to undo)	Standard Implementation ✓	Design Principle ☐	Policy ☐
Point of Contact Name:	Michelle Giglio		
Point of Contact Email Address:	mgiglio@som.umaryland.edu		
End of Date Comments (MM/DD/YY):	12/23/2022		
Submitting Team/DCC Name:	Ontology Working Group		
CFDE-Draft-Status:	Pending ●	Active and Open for Comments ●	
URL Link to the document:	https://docs.google.com/document/d/1TedisnuIogbaOPLCLnC4q-OSvcHqGwoqMnD2EVr3e4/edit?usp=sharing		
License (Optional):			

Introduction

The Ontology Working Group has decided on several standard ontologies and vocabularies to be used for submission of metadata to the C2M2. These include: EDAM for data type and file format, Ontology for Biomedical Investigations (OBI) for assay types, Human Disease Ontology (DO) for diseases, PubChem for chemical compounds and substances, Uberon for anatomy, and NCBI Taxon for taxonomy. DCCs submitting metadata using these vocabularies use the most granular terms available that are appropriate for their assets. However, extensive granular data can be challenging to visualize in summary views or in comparisons. It is also useful to be able to search based on larger grouping categories. To meet this need, “slims” of ontologies are used. An ontology slim comprises a set of higher level, more general terms into which annotations to more specific terms can be binned based on the parent-child relationships between the ontology terms. This RFC provides a description of the processes used to generate slims for use in the CFDE Portal. These slims are not used for curating metadata directly. Each DCC will continue to use the full scope of the chosen vocabularies for curation.

CFDE Ontology Slim Generation

Identifying the slim terms

Both algorithmic and manual methods can be used to choose terms to be included in an ontology slim. Many community ontologies have pre-built slims or term subsets for various purposes. For the CFDE slim construction, in some cases we modified existing slims/subsets while in others we constructed slims from scratch manually.

- Slims modified from existing term subsets provided by the source ontology:
 - Disease Ontology (DO) - We started with the slim built by the DO for use with the Alliance of Genome Resources (AGR) called [DO_AGR_slim](#). As a group we reviewed this slim and then modified it to both add additional terms with more specificity and to make sure that all upper level branches of DO were included. Therefore any granular DO term that might get used would be able to be slimmed. This resulted in [Version 1 of the CFDE DO slim](#).
 - Uberon - We started with the Uberon tsv files containing [subsets of terms](#) under the various organ system terms in uberon (e.g. circulatory, digestive). The slim terms became the relevant uberon organ system term (e.g. 'circulatory system') and the subsets were used to generate the granular-term to slim-term mappings. We manually reviewed these sets and in some cases merged two of the uberon sets into one slim term or added additional slim terms. The uberon sets that we chose did not include mappings to all uberon granular terms. Therefore, we checked for granular terms in use in C2M2 that did not map to one of the uberon slim terms and manually mapped them to one of the slim terms or added additional slim terms to map them to as needed. The uberon slim was reviewed by the OWG. The result is [Version 1 of the CFDE Uberon slim](#).
- Slims generated manually:

When constructing a slim, one can use a top-down or bottom-up approach. A top-down approach involves inspecting the upper level structure of the ontology to see if you can select a set of terms that encompass categories at the level of specificity you are interested in. This works well for highly structured and well-organized ontologies. A bottom-up approach starts with the granular term annotations for a specific dataset and builds a slim of upper level terms that will encompass the totality of the terms in your dataset, but not necessarily the totality of all terms in the ontology.

For the below slims we used a combination of top-down and bottom-up approaches. We first selected upper level grouping terms that would function well as slim terms for the CFDE. For any terms that were used in CFDE datasets but did not slim to one of those terms, we chose additional slim terms to insure that all granular terms in our datasets had a term to slim to.

- [Version 1 - EDAM format](#)
- [Version 1 - EDAM data](#)
- [Version 1 - OBI assay](#)

Slim files

- Each slim mapping file (linked above) contains three tabs. The first is the list of slim terms. The second is a mapping between the slim terms and the terms used in the CFDE datasets. The third is a mapping of all terms in the ontology that can map to the slim terms. Some of the slims don't include terms to cover the entire ontology.
- Slims are submitted for use in the CFDE Portal in the form of tsv files that contain two columns, the first is granular ontology terms the second is the slim term they map to. The files are placed in the [CFDE Deriva GitHub](#) branch for further processing and incorporation into the CFDE Portal.

Decisions that were made on how the slims were constructed and used

- slim terms slim to themselves
- some slim terms may be children of other slim terms
 - When this is the case, we chose to handle the mappings such that granular mappings will map both to the intermediate parent and to the grandparent. For example: the DO term 'skin carcinoma' maps to both the slim term 'skin cancer' and the slim term 'cancer' which is a parent of 'skin cancer'. In this situation, the term 'skin carcinoma' will be mapped to both slim terms. When displayed in summary displays or comparative analyses the granular annotation to 'skin carcinoma' will be counted/displayed both in the 'skin cancer' and 'cancer' bins.

Proposed Actions

- Over time as new datasets are submitted to the CFDE the pool of granular terms used in annotations will grow. Therefore, after each data submission, checks will be run to identify terms used in the CFDE data that do not map to any slim term. We will generate new versions of the slim so as to encompass all of these in appropriate slim mappings. For ontologies in which the subset of slim terms is not inclusive enough such that all granular terms will map to a slim term, we will also include the root or the branch node for the terms in question (for example 'assay' in OBI).
 - We will explore the use of automated methods (for example: [PMID:20929524](#)) to help in this process.
 - We will continue the manual methods we have used so far.
- New slim mapping files will be versioned in their names and stored in the Ontology Working Group Google Drive area for slims.
- New slim versions will be submitted to GitHub for inclusion in the CFDE Portal via the process described above.
- Slims will be created for additional ontologies/vocabularies in use by the C2M2 and will include:
 - PubChem
 - NCBI Taxon
 - the phenotype ontology that gets chosen for use in C2M2
 - any other additional ontologies