Human Cell Atlas Ontology

Danielle Welter, David Osumi-Sutherland, Simon Jupp
Samples, Phenotypes & Ontologies Team
EMBL-EBI
Hinxton, Cambridge

***Abstract*—The Human Cell Atlas (HCA) project aims to build comprehensive reference maps of Human cells that will further our understanding of biological processes and diagnosing and treating disease. We present the Human Cell Atlas Ontology (HCAO), an application ontology that builds and extends from existing ontology standards to provide terminology standards with well-defined semantics for describing experimental data coming into the HCA to ensure the data is interoperable and amenable to integrative analysis.**

Keywords—Human Cell Atlas; application ontology; data interoperability

I. Introduction

While there are estimated to be in excess of 37 trillion cells in the adult human body, conventional understanding from microscopy and morphology studies has until recently put the number of different cell types at only 200 to 300. Emerging evidence from studies using new single-cell sequencing technologies however suggests that this may be a gross underestimate. The Human Cell Atlas (HCA) proposes building a systematic, data-driven atlas of human cells, redefining human anatomy and building a periodic table and classification of cell types.

Development of this atlas requires the collection of high quality metadata for each dataset following a standard that allows interoperability and reuse. But this presents a major challenge. Defining an exhaustive yet intuitive set of metadata standards in a fast-moving field such as single-cell sequencing presents a number of inherent difficulties, including harmonising conflicting field-specific conventions and anticipating the metadata requirements of future analyses. In order to ensure semantic interoperability within and beyond HCA data, we have built an application ontology that will provide a standard terminology for data submitted to the HCA.

II. HCA Ontology

The HCA Ontology (HCAO) is a dedicated application ontology developed for the annotation of HCA metadata with terms from relevant domain ontologies. We have developed a pipeline for the automated construction of the HCAO that builds from reference source ontologies to provide a core subset of terms relevant for HCA data. The HCAO is designed to provide a simplified view over a number of ontologies to ease navigation, search and visualisation for the submitters and data curators.

HCAO imports elements of Uberon, GO, CL and Human Development Stages (HSAPDV), as well as other dependent ontologies such as the PRotein Ontology (PRO) and the Phenotype And Trait Ontology (PATO). The HCA-Uberon import is largely limited to human anatomy terms, determined through cross-references between Uberon terms and FMA terms. FMA cross references are also used to extract a subset of human cells from the cell type ontology (CL). The HCA Ontology was created using the ontology starter kit[1] and the pipeline utilises ROBOT[2] to automatically construct the ontology.

In addition to the HCAO, we also use an HCA-specific slim of the Experimental Factor Ontology (EFO). This slim includes a wide range of concepts determined to be relevant to obtain good annotation coverage of HCA data, including experimental processes, instruments, biological macromolecules, organisms/strains, protocols and units. The choice of a separate EFO slim rather than further imports into HCAO was made to avoid overloading the scope of HCAO and limit accidental clashes between direct imports into HCAO and derived imports into EFO, which uses similar source ontologies to HCAO. Disease terms in HCA are mapped to the Monarch Disease Ontology (MONDO).

The entire set of ontologies are deployed via a dedicated instance of the EMBL-EBI ontology lookup service (OLS) that can be run inside a Docker container. OLS provides a search interface and RESTful API to the ontologies that will be used to access the ontologies by the HCA infrastructure. By deploying these in a Docker container we are able to manage stable releases of the ontologies and deploy these to the cloud-based infrastructure. A development instance of OLS hosting HCAO can be viewed at http://ontology.dev.data.humancellatlas.org

III. Data Validation

The initial role of the HCAO is to annotate experimental data submitted to the HCA. The HCA provides a metadata schema defined in JSON schema, a JSON based format for defining the structure of JSON data. We have extended the

[1] https://github.com/INCATools/ontology-starter-kit
[2] http://robot.obolibrary.org

This work was funded by Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation Year 1 Grant #: 2017-171304 (5022) / Year 2 Grant #: 2017-171668 (5022)

JSON schema syntax with modules to capture ontology values by including a “graph restriction” block that allows schema developers to specify what constitutes a valid ontology term for a given field in a form that can be checked programmatically. For example, the biomaterial schema[3] has an ontology module[4] for capturing the organ part using a subset of UBERON anatomy terms imported into the HCAO. The graph restriction block allows values to be restricted to a subset of terms from one or more ontologies. Valid terms from these ontologies are declared by specifying one or more starting nodes and a set of edges types to follow. Edge types can include subClassOf restrictions and simple existential restrictions using a single object property (e.g. part of). There are also options to include or exclude the starting node and limit subsets to direct children only or include all descendants. As an example, annotating the organ field with “neoplasm” (MONDO:0005070) would fail validation as this field is restricted to subclasses of organ (UBERON:0000062) and haemolymphatic fluid (UBERON:0000179). We have developed a custom JSON schema validation application that performs additional ontology-based validation against the HCAO as deployed in the OLS docker instance (https://github.com/HumanCellAtlas/ingest-validator).

The HCAO ontology is available from https://github.com/HumanCellAtlas/ontology.

IV. Conclusion

The HCAO aims to address data annotation and harmonisation challenges in the Human Cell Atlas by providing a focused and simplified aggregation of relevant source ontologies. We have also extended the HCA metadata JSON schema to include ontology-based restrictions to facilitate metadata validation and ensure ontology annotations are restricted to a set of relevant and appropriate terms.

Acknowledgment

The authors would like to acknowledge all members of the Human Cell Atlas Data Coordination Platform (DCP) at EMBL-EBI, the Broad Institute, the Chan Zuckerberg Initiative and the University of California Santa Cruz, with particular thanks to the HCA metadata team.

[3] https://schema.humancellatlas.org/type/biomaterial/5.1.1/specimen_from_organism

[4] https://schema.humancellatlas.org/module/ontology/5.1.0/organ_ontology