HuBMAP Ontology Design: Data and Terminology Structures, Crosswalks, and Ontology Creation Software

Samuel Friedman, Ellen Quardokus, and Paul Macklin

May 28, 2019

1Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

Scope and Focus

This document contains the descriptions of various ontologies that were examined; reasons for selecting a subset of these ontologies, and an argumentation and documentation of how these ontologies were combined to create a HuBMAP Ontology. Please note that much of the discussion below is focused on kidney anatomy, given that we are initially prototyping the CCF on existing and expected kidney data. This specific focus should also make the discussion below more concrete.

Introduction

To determine how to create a HuBMAP ontology, we examined pre-existing ontologies as well as methodologies for creating ontologies. We decided to work with pre-existing ontologies as the basis for our ontology and developed a rationale for what pre-existing ontologies we want to use. Furthermore, we classified the ontologies and ranked them so that it becomes clear about how we select terms from the ontologies. Because we have multiple uses for the ontologies, we extended prior methodologies to create single "slim" ontologies that bundle subsets of large, complex ontologies (e.g, (Davis, Sehgal, and Ragan 2010)), to develop a new methodology to combine subsets of multiple ontologies into "slim" ontologies. With these methods, we also created a release of the ontology for v0.1.0. Going into the future, we developed a roadmap for the tasks necessary to reach v0.5.0 for the end of Y1 of HuBMAP.

2Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

Rationale for selecting ontologies

In attempting to create an ontology, we found that we had to reconcile multiple goals that were not always in full alignment. We tried to determine what makes a "good" ontology that could satisfy identified needs, including:

1. **Ease of navigation:** Some ontologies have additional layers of hierarchy that are difficult to navigate. We may want to implement "views" that act as fast shortcuts to critical elements of the ontology or data attributes. We leverage existing ontologies (or parts thereof) to serve as these shortcuts
    a. MeSH: /Urogenital System/Urinary Tract/Kidney
    b. NCIT: /Anatomic Structure, System, or Substance/Organ/Kidney
    c. FMA: /Anatomical Entity/Physical anatomical entity/Material anatomical entity/Anatomical Structure/Postnatal anatomical structure/Organ/Solid organ/Parenchymatous organ/Corticomedullary organ/Kidney
    d. BRENDA (BTO): /tissue, cell types, and enzyme sources/animal/whole body/gland/excretory gland/kidney
    e. UBERON: /BFO_0000002/BFO_0000004/anatomical entity/material anatomical entity/anatomical structure/multicellular anatomical structure/organ/compound organ/cavitated compound organ/kidney
    f. KTAO: /entity/continuant/independent continuant/kidney
2. **A balance of visual clarity and detail:** We do not want two items ("left kidney" and "right kidney") that clutter data organization and browsing, nor do we want a listing of

3Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

every fact about a kidney. We do not want to flatten our information, nor do we want every possibility of "adjective + noun" for a given noun.

3. **Completeness:** Recording of all relevant biological knowledge for the data HuBMAP expects to obtain
    a. We want to have all the terms written down and without uncertainty on the potentially missing terms. Some ontologies only cover particular organs and may omit some relevant scientific knowledge.
    b. Of MeSH, NCIT, FMA, BRENDA, UBERON, and KTAO, only MeSH and BRENDA (BTO) had finer-scale details besides "adult mammalian kidney", "left kidney", "mesonephros", "metanephros", "pronephros", and "right kidney".
4. **Crosswalks, interlinking, or cross-references with other ontologies:**
    a. Not all ontologies record their "crosswalks" between equivalent terms in different ontologies, but UBERON provides some excellent crosswalks.
    b. We need to ensure that each term that we add has the relevant crosswalks between the ontologies that we are using.
5. **Widespread prior adoption:** Ideally, we adapt ontologies that are widely used by the community so as to ensure the acceptance of the HuBMAP ontology by the wider community.
6. **Permissive licensing:** We anticipate that the HuBMAP data and ontologies will be used in third-party and derivative projects. Therefore, we must prioritize ontologies and technologies with permissive, non-cumbersome reuse policies.
    a. Open source is preferred (for software and ontologies), and permissive data licenses (e.g., CC0 or CC-BY) is preferred for data.

4Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

    b. Software license proliferation is a risk: if the HuBMAP ontology is built from components with too many licenses, it can become impossible for end users to completely determine their rights and restrictions, which hinders adoption and community contributions. It also increases the risk that some components of the CCF are legally incompatible with other components of the CCF.
    c. Ontologies with restrictive and confusing reuse criteria should be only used as a last resort, or with direct negotiations of a license exception. This includes scenarios where an open source ontology / software component is less feature complete than an encumbered one, but the open choice can be readily extended to recover the missing features.
7. **Active maintenance:** We seek ontologies that have not been abandoned, but rather are under active development and maintenance, because:
    a. We expect new terms to come along as scientific knowledge advances, thus requiring updates, and
    b. We need to have software to create the ontologies instead of editing manually to ensure that new terms from the dependent ontologies become integrated into the HuBMAP ontology
8. **Mature ontologies:** On the other hand, as consensus emerges with a community and software bugs are eliminated, we expect development and maintenance activity to slow. Thus, it is important not to dismiss ontologies with low activity and assume community abandonment, but recognize that this may rather be a sign of maturity**.**

5Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

Proposed Ontologies

We have the following list of proposed ontologies to use, along with their key roles within the HuBMAP CCF ontology:

1. Anatomic/Phenotypic
    a. Uber-anatomy Ontology (UBERON) to describe anatomy (Mungall et al. 2012)
    b. Phenotype And Trait Ontology (PATO) to describe large scale biological phenotypes (G. V. Gkoutos, Schofield, and Hoehndorf 2018)
    c. Foundational Model of Anatomy (FMA) as a subset of UBERON (Rosse and Mejino 2003)
    d. Human Phenotype Ontology (HPO) so we can know about sample origin (Robinson et al. 2008)
2. (Sub-)Cellular
    a. Cell Ontology (CL) to describe cellular properties (Bard, Rhee, and Ashburner 2005)
    b. Gene Ontology (GO) to describe genes and many subcellular properties (Ashburner et al. 2000)
    c. Chemical Entities of Biological Interest (ChEBI) to describe chemical properties (Degtyarenko et al. 2008)
    d. RNA Ontology (RNAO) to describe RNA (Hoehndorf et al. 2011)
    e. Protein Ontology (PR) to describe proteins (Natale et al. 2011)
    f. Cell Behavior Ontology (CBO) to describe how cells are operating at the cellular level and up (Sluka et al. 2014)

6Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

3. Metadata
    a. Basic Formal Ontology (BFO) because we need some base for all the ontologies (Arp, Smith, and Spear 2015)
    b. Information Artifact Ontology (IAO) because we need metadata (Smith and Ceusters 2015)
    c. Ontology of units of Measure (OM) (Rijgersberg, Assem, and Top 2013) because we need units and it has more features than the Units Ontology (UO) (Georgios V Gkoutos, Schofield, and Hoehndorf 2012)
4. Tissue/Sample Collection
    a. Biological Spatial Ontology (BSPO) so we can know about how samples were sliced (Dahdul et al. 2014)
    b. Ontology of Biomedical Investigations (OBI) to describe how the data was collected (Bandrowski et al. 2016)
    c. EDAM because it covers many bioinformatics terms (Ison et al. 2013)
5. Miscellaneous (these terminologies cover many ideas not necessarily covered elsewhere)
    a. Medical Subject Headings (MeSH) to cover many labels not otherwise easily covered (Rogers 1963)
    b. NCI Thesaurus (NCIT) to cover many labels not otherwise easily covered (Fragoso et al. 2004)

Methodology for Ontology Creation

Given that we have multiple requirements for an ontology, we propose creating “slim” ontologies like those done by the Gene Ontology (GO) (e.g., see (Davis, Sehgal, and Ragan 2010)). GO

7Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

creates subsets of the overall ontology to make it easier to use due to the fact that the end user is not necessarily interested in the entire ontology. This leads us to a new methodology of creating multiple ontologies for the HuBMAP CCF. Based upon user surveys of the funded TMCs and current kidney pilot data, we compiled an initial list of terms that need to go into a base ontology, those terms are found in pre-existing ontologies, and then merged together into a base ontology. We note that this list of terms will be grown based upon data. Year 2 will likely need to coordinate with the tools team to allow TMC end users to report and request missing terms.

Based upon usability testing and development needs of PI Börner's team, a second list of terms was created for those terms that are needed for a particular view of the data, e.g., for a given zoom level, for given cell type(s), or for given chemicals/RNA. This second list of terms then gets utilized to create a slim HuBMAP ontology.

When reviewing ontologies, we found that ontologies covered different domains and also had some convergent evolution towards similar solutions to commonly encountered problems. Thus, multiple ontologies partially but not fully address the CCF needs, and it is more productive to adapt a "nearly perfect" ontology than it is to additionally search for a (likely non-existent) perfect ontology. Indeed, it is infeasible to find and review all bio ontologies at a fine level of detail. Instead, we supplemented our combined 10+ years of domain expertise with six months of additional searching to find and adopt widely-used ontologies that cover most of the terms that we need to cover, and then extended them as needed. We plan to report missing terms from our extensions back to these ontologies in the future, to help maintain compatibility with community standards. With this short list of ontologies, we could rank them to determine which ontologies

8Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

we should look at first. We created this ranking based on the presence of crosswalks between ontologies and sufficient specification of terms.
Since we will provide multiple views, we can easily provide multiple lists and thus generate multiple ontologies. As we can translate the base ontology into a DAG (directed acyclic graph) (Peng, Jiang, and Radivojac 2018), we can easily remove edges to create an ordered tree or forest. We adopted Python as the programming language of choice because of strong graph theory libraries, ease of prototyping, and well-developed OWL libraries. To create the base CCF ontology, we went through the following procedure:

1. We used the Ontospy Python package to load an ontology (e.g., UBERON) and sort it into different classes (each ontological term has its own class) and keep track of the triples associated with each class (Pasin n.d.). We picked Ontospy because it interfaced well with RDFLib.
2. We created a spreadsheet with a list of desired ontological terms with their corresponding ID numbers. The mapping between each term and its ID was purposely done manually: automated methods (such as BioPortal's annotator) returned too many results or incorrect results. (A pending need would be an ontological annotator that understands aspects of the terms' author's background.) Each term also indicated whether descendant terms of that term should be included in the ontology. The spreadsheet was then exported into CSV to be read. We used a spreadsheet to make term entry easy for domain experts.
3. Using OWL elements, we created a directed graph (though not necessarily acyclic) where the edges:
    a. Created from "X rdfs:subClassOf Y" representing a directed edge from Y to X
    b. Created from "X BFO:partOf Y" representing a directed edge from Y to X, although we needed to go through triples that involved "node owl:OnProperty

9Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

BFO:partOf", "node owl:someValuesFrom Y", and "X rdfs:subClassOf node" where node was usually an unidentified/unnamed element.

c. Later on, we did not create and then removed edges representing from "X owl:equivalentClass Y" as Protégé will normally create an edge between these nodes as this dramatically decreased navigability.
d. We created these edges so as to have a more complete graph.

4. We checked the list of input terms and compared that against the list of terms in our ontologies.
    a. We output as a warning those list of nodes that cannot be used. We add this to a list of terms to ignore.
    b. The user needs to determine what needs to be done: Either edit the input list or somehow add the term to an ontology that we can read in.
5. We assigned weights to the edges. The weights went up for increasing order of importance. Arbitrarily picked factors of 5-20 were involved in scaling up to emphasize certain policies.
    a. We assigned to each edge a numerical weight to favor certain edges over other edges.
        i. Edges created from a subClassOf element received a weight of 1
        ii. Edges created from a partOf element received a weight of 20
        iii. We assigned these weights to separate the values and to emphasize part-ness vs. subclass-ness.
    b. We added additional weight to certain edges to emphasize connections from particular nodes, namely those nodes that are are key "root nodes" of the ontology. This includes two of the major HuBMAP organs: kidney and heart.

10Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

i. Outward edges from key nodes (e.g. the node of a particular organ) received an additional weight of 100. Key nodes are specified in the list of input terms under the "Parent ID" term (and its associated "Parent Name" term).
ii. We computed simple paths from each requested node term to each emphasized node (if possible) and gave each edge in every simple path an additional weight of 500.

c. To create the "insertion" points for the various organs, we added even more weight to a single simple path connecting each of the key "root nodes" to the node "anatomical system".
i. The "anatomical system" node serves as the entry bode for the entire body.
ii. Since there can be multiple simple paths between each key "root node" and "anatomical system", we picked the shortest simple path between those two nodes. If multiple simple paths had the same shortest length, we picked the first one that was found.
iii. All edges along this single simple path received an addition weight of 5000.

6. We then created a subgraph of the original graph, only selecting those nodes that were either nodes in the original list of ontological terms, ancestors of those nodes, or descendants of those nodes. This creates a "slim" ontology that is relevant for HuBMAP.
7. Using the created weighted directed subgraph, we computed a maximum cost arborescence or a maximum cost branching depending on the nomenclature. Such a branching is critical for creating a navigable ontology.

11Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

8. We removed from the branching any terms that should be ignored. These terms can come either from user input or from previously identified terms to be ignored (such as invalid input choices).
9. We removed from the branching various edges:
    a. All incoming edges to the major organ nodes to ensure that we have these nodes be root nodes in the ontology. Using this branch, we computed the edges present in the original graph and not present in the modified branching to determine the list of edges that need to be removed from the graph representing the ontology. We wanted to remove these edges to improve navigability.
    b. We also removed edges from nodes that were not ancestors of the initial input terms when using the graph computed from maximum cost branching. We termed those nodes that are ancestors (in the branching) or the terms themselves the nodes that "support" the ontology.
10. We further removed the triples associated with all of the edges that needed to be removed and saved the resulting ontology to a file. Because we used Ontospy, which in turn uses RDFLib, we easily saved the ontology in the RDF/XML format. This can easily be changed in the input parameter file. We removed the triples to make the ontology nicer for display purposes.
11. We also created a partonomy where we essentially replace the subClassOf relationship with the part_of relationship. This becomes non-trivial due to how nested all the terms are.

The UBERON ontology also has triples that are of the format "X oboInOwl:hasDbXref Y" where X is the UBERON term and Y is a term in another ontology. Thus, we can easily create

12Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

graphs of other ontologies, add edges in the graph representing UBERON from UBERON nodes to nodes in the other ontologies.

Ontology Usage

To make use of the ontology, we recommend that users find the relevant term and then use identifiers from the ontology to perform semantic annotation. Users should use the rdfs:label as a human readable term and oboInOwl:id as a machine readable term. For navigation purposes, users should follow rdfs:subClassOf elements and BFO "part of" elements (http://purl.obolibrary.org/obo/BFO_0000050) as these are the edges of the graph between the different nodes of terms.

Inherited Hierarchies

Since we are using previously created ontologies, we inherit their hierarchies when we create the HuBMAP ontology. As these hierarchies take the form of directed graphs, we have some freedom to choose paths through the graph as path from an organ root node to a node that comes from the list of parts of an organ. For v0.1.0, we simply picked one path through the ontologies to create our edges for navigation purposes. We will investigate in the future if some paths are better than others.

13Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

Open Questions

We need to determine how to integrate in specific terms from individual ontologies. While more detailed, some of these ontologies are not fully open source (LungMAP) or have other challenges.

Ontologies for individual organs

We have identified and are assessing organ-specific ontologies for prototyping the kidney, but we still have open questions:

1. **Kidney** can use KTAO (He et al. 2018), and it has so far met the annotation needs communicated to the CCF team. We will want to consult domain experts to discover any missing terms.
2. We have not yet identified a **Heart** ontology satisfying the 8 rationale criteria above to date. We have created a prototype ontology (a short list of terms) as extracted from interactions with the Heart TMC. It will be a continuing challenge to select an appropriate ontology.

Ontologies for provenance

As we work on developing the data pipelines, we will need to work more on metadata and data provenance. We are looking at using these ontologies/data terminologies:

14Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

1. Software Ontology (SWO) (Malone et al. 2014)
2. EDAM (Ison et al. 2013)
3. VIVO (Ding, Mitchell, and Corson-Rikert 2011)
4. ORCID (Not an ontology per se, but useful)("ORCID" n.d.)
5. Dublin Core (Not an ontology per se, but useful) ("DCMI: Home" n.d.)
6. PAV Ontology (Ciccarese et al. 2013), see https://github.com/pav-ontology/pav/blob/2a7653affb2870576120b2425e8e0f93e0c41a44/pav.rdf for an example with ORCID.

Crosswalks / MetaData mapping

UBERON has numerous crosswalks already included, especially with FMA, which includes some terms not in UBERON. We would like to establish a database of these crosswalks so that they can be automatically added to the ontologies as terms are added. There are also prior methods that perform ontology alignment (such as LOOM) and this should be a productive avenue in the future. Furthermore, we investigated Machine learning (ML) and unsupervised ontology matching using Natural Language Processing (NLP) (Arguello Casteleiro et al. 2018; Kolyvakis, Kalousis, and Kiritsis 2018).

References

Arguello Casteleiro, Mercedes, George Demetriou, Warren Read, Maria Jesus Fernandez Prieto, Nava Maroto, Diego Maseda Fernandez, Goran Nenadic, Julie Klein, John Keane, and

15Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

Robert Stevens. 2018. “Deep Learning Meets Ontologies: Experiments to Anchor the Cardiovascular Disease Ontology in the Biomedical Literature.” *Journal of Biomedical Semantics* 9 (1): 13–13. https://doi.org/10.1186/s13326-018-0181-1.

Arp, Robert, Barry Smith, and Andrew D. Spear. 2015. *Building Ontologies with Basic Formal Ontology*. The MIT Press.

Ashburner, M, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, et al. 2000. “Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.” *Nature Genetics* 25 (1): 25–29. https://doi.org/10.1038/75556.

Bandrowski, Anita, Ryan Brinkman, Mathias Brochhausen, Matthew H. Brush, Bill Bug, Marcus C. Chibucos, Kevin Clancy, et al. 2016. “The Ontology for Biomedical Investigations.” *PLOS ONE* 11 (4): e0154556. https://doi.org/10.1371/journal.pone.0154556.

Bard, Jonathan, Seung Y. Rhee, and Michael Ashburner. 2005. “An Ontology for Cell Types.” *Genome Biology* 6 (2): R21. https://doi.org/10.1186/gb-2005-6-2-r21.

Ciccarese, Paolo, Stian Soiland-Reyes, Khalid Belhajjame, Alasdair Jg Gray, Carole Goble, and Tim Clark. 2013. “PAV Ontology: Provenance, Authoring and Versioning.” *Journal of Biomedical Semantics* 4 (1): 37–37. https://doi.org/10.1186/2041-1480-4-37.

Dahdul, Wasila M, Hong Cui, Paula M Mabee, Christopher J Mungall, David Osumi-Sutherland, Ramona L Walls, and Melissa A Haendel. 2014. “Nose to Tail, Roots to Shoots: Spatial Descriptors for Phenotypic Diversity in the Biological Spatial Ontology.” *Journal of Biomedical Semantics* 5 (August): 34–34. https://doi.org/10.1186/2041-1480-5-34.

Davis, Melissa J, Muhammad Shoaib B Sehgal, and Mark A Ragan. 2010. “Automatic, Context-Specific Generation of Gene Ontology Slims.” *BMC Bioinformatics* 11 (October): 498–498. https://doi.org/10.1186/1471-2105-11-498.

“DCMI: Home.” n.d. Accessed March 22, 2019. http://www.dublincore.org/.

Degtyarenko, Kirill, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. 2008. “ChEBI: A Database and Ontology for Chemical Entities of Biological Interest.” *Nucleic Acids Research* 36 (Database issue): D344–50.

16Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

https://doi.org/10.1093/nar/gkm791.
Ding, Ying, Stella Mitchell, and Jon Corson-Rikert. 2011. “The Vivo Ontology: Enabling Networking of Scientists.” In .
Fragoso, Gilberto, Sherri de Coronado, Margaret Haber, Frank Hartel, and Larry Wright. 2004. “Overview and Utilization of the NCI Thesaurus.” *Comparative and Functional Genomics* 5 (8): 648–54. https://doi.org/10.1002/cfg.445.
Gkoutos, G. V., P. N. Schofield, and R. Hoehndorf. 2018. “The Anatomy of Phenotype Ontologies: Principles, Properties and Applications.” *Brief Bioinform* 19 (5): 1008–21. https://doi.org/10.1093/bib/bbx035.
Gkoutos, Georgios V, Paul N Schofield, and Robert Hoehndorf. 2012. “The Units Ontology: A Tool for Integrating Units of Measurement in Science.” *Database : The Journal of Biological Databases and Curation* 2012 (October): bas033–bas033. https://doi.org/10.1093/database/bas033.
He, Yongqun, Becky Steck, Edison Ong, Laura Mariani, Chrysta Lienczewski, Ulysses Balis, Matthias Kretzler, et al. 2018. “KTAO: A Kidney Tissue Atlas Ontology to Support Community-Based Kidney Knowledge Base Development and Data Integration,” 6.
Hoehndorf, Robert, Colin Batchelor, Thomas Bittner, Michel Dumontier, Karen Eilbeck, Rob Knight, Chris J. Mungall, et al. 2011. “The RNA Ontology (RNAO): An Ontology for Integrating RNA Sequence and Structure Data.” *Appl. Ontol.* 6 (1): 53–89.
Ison, Jon, Matús Kalas, Inge Jonassen, Dan Bolser, Mahmut Uludag, Hamish McWilliam, James Malone, Rodrigo Lopez, Steve Pettifer, and Peter Rice. 2013. “EDAM: An Ontology of Bioinformatics Operations, Types of Data and Identifiers, Topics and Formats.” *Bioinformatics (Oxford, England)* 29 (10): 1325–32. https://doi.org/10.1093/bioinformatics/btt113.
Kolyvakis, Prodromos, Alexandros Kalousis, and Dimitris Kiritsis. 2018. *DeepAlignment: Unsupervised Ontology Matching with Refined Word Vectors*. https://doi.org/10.18653/v1/N18-1072.
Malone, James, Andy Brown, Allyson L Lister, Jon Ison, Duncan Hull, Helen Parkinson, and

17Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

Robert Stevens. 2014. “The Software Ontology (SWO): A Resource for Reproducibility in Biomedical Data Analysis, Curation and Digital Preservation.” *Journal of Biomedical Semantics* 5 (June): 25–25. https://doi.org/10.1186/2041-1480-5-25.

Mungall, C. J., C. Torniai, G. V. Gkoutos, S. E. Lewis, and M. A. Haendel. 2012. “Uberon, an Integrative Multi-Species Anatomy Ontology.” *Genome Biol* 13 (1): R5. https://doi.org/10.1186/gb-2012-13-1-r5.

Natale, Darren A, Cecilia N Arighi, Winona C Barker, Judith A Blake, Carol J Bult, Michael Caudy, Harold J Drabkin, et al. 2011. “The Protein Ontology: A Structured Representation of Protein Forms and Complexes.” *Nucleic Acids Research* 39 (Database issue): D539–45. https://doi.org/10.1093/nar/gkq907.

“ORCID.” n.d. Accessed March 22, 2019. https://orcid.org/.

Pasin, Michele. n.d. *Ontospy: Query, Inspect and Visualize Knowledge Models Encoded as RDF/OWL Ontologies.* (version 1.9.8.2). Python. Accessed March 24, 2019. https://github.com/lambdamusic/ontospy.

Peng, Yisu, Yuxiang Jiang, and Predrag Radivojac. 2018. “Enumerating Consistent Sub-Graphs of Directed Acyclic Graphs: An Insight into Biomedical Ontologies.” *Bioinformatics (Oxford, England)* 34 (13): i313–22. https://doi.org/10.1093/bioinformatics/bty268.

Rijgersberg, Hajo, Mark van Assem, and Jan Top. 2013. “Ontology of Units of Measure and Related Concepts.” *Semant. Web* 4 (1): 3–13.

Robinson, Peter N, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. 2008. “The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease.” *American Journal of Human Genetics* 83 (5): 610–15. https://doi.org/10.1016/j.ajhg.2008.09.017.

Rogers, F B. 1963. “Medical Subject Headings.” *Bulletin of the Medical Library Association*, Fragoso, 51 (1): 114–16.

Rosse, Cornelius, and José L.V. Mejino. 2003. “A Reference Ontology for Biomedical Informatics: The Foundational Model of Anatomy.” *Unified Medical Language System* 36 (6): 478–500. https://doi.org/10.1016/j.jbi.2003.11.007.

18Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)

Sluka, James P, Abbas Shirinifard, Maciej Swat, Alin Cosmanescu, Randy W Heiland, and James A Glazier. 2014. “The Cell Behavior Ontology: Describing the Intrinsic Biological Behaviors of Real and Model Cells Seen as Active Agents.” *Bioinformatics (Oxford, England)* 30 (16): 2367–74. https://doi.org/10.1093/bioinformatics/btu210.

Smith, Barry, and Werner Ceusters. 2015. “Aboutness: Towards Foundations for the Information Artifact Ontology.” In *ICBO*.

19Thus, we have the following (rough) ranking, classified by each category:

1. Anatomic/Phenotypic
    a. UBERON
    b. Foundational Model of Anatomy (FMA)
    c. Human Phenotype Ontology (HPO)
    d. Phenotype and Trait Ontology (PATO)
    e. Organ specific
2. (Sub-)Cellular
    a. Cell Ontology (CL)
    b. Gene Ontology (GO)
    c. Chemical Entities of Biological interest (ChEBI)
    d. Protein Ontology (PR)
    e. RNA Ontology (RNAO)
    f. Cell Behavior Ontology (CBO)
3. Metadata
    a. Basic Formal Ontology (BFO)
    b. Relationship Ontology (RO)
    c. Information Artifact Ontology (IAO)
    d. Ontology of units of Measure (OM)