Striving for semantic harmony across datasets, communities, and real-world data

Chris Stoeckert, Ph.D.

Department of Genetics, Institute for Biomedical Informatics

Perelman School of Medicine, University of Pennsylvania

ICBO 2022

Michigan League, University of Michigan, Ann Arbor, MI, USA

CCBOT-2022

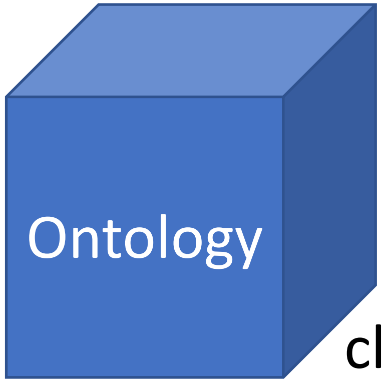
Penn Medicine

Semantic harmony

- Consistent representation of data and what the data is about using ontology terms.
- Consistent development, management and application of ontologies.
- Balance of “we need it now” pragmatism and “do it correctly” formalism.

Cocoa and Emma as yin-yang

Ontologies* can support different aspects / multiple dimensions of standardization at the same time



Instances in triples (ABox)

classes with relations (TBox)

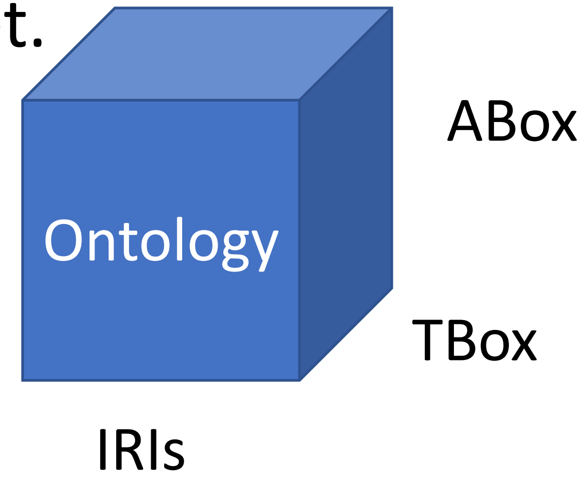
defined terms with IRIs

- *Not everyone needs to work in all dimensions but a better understanding of each will make for better usage.*

* See Rector et al.: On beyond Gruber: “Ontologies” in today’s biomedical information systems and the limits of OWL. J Biomed Inform. 2019;100S:100002.

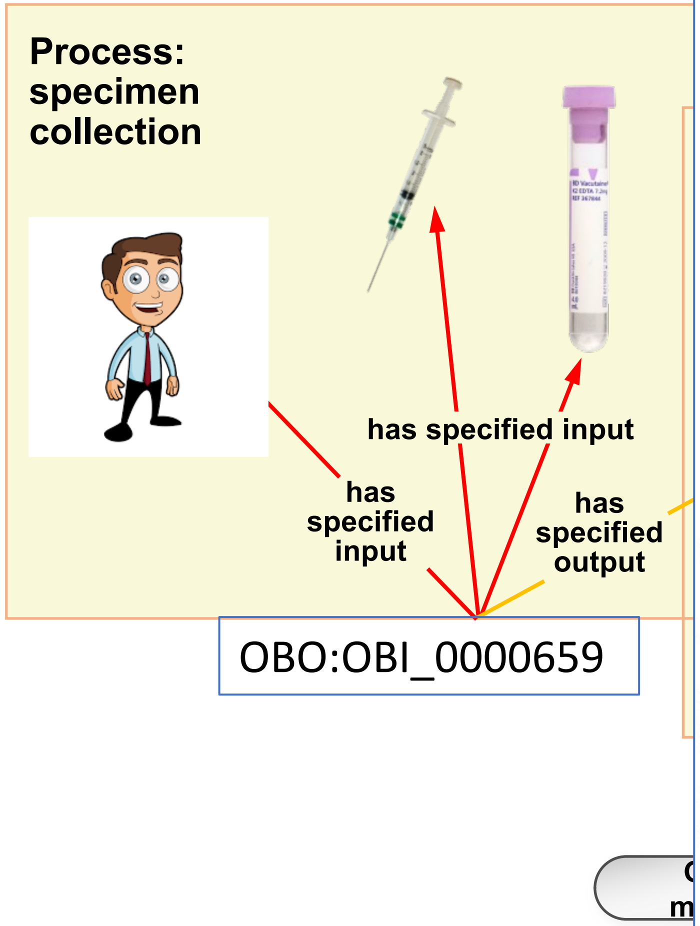
Definitions to get us on the same page

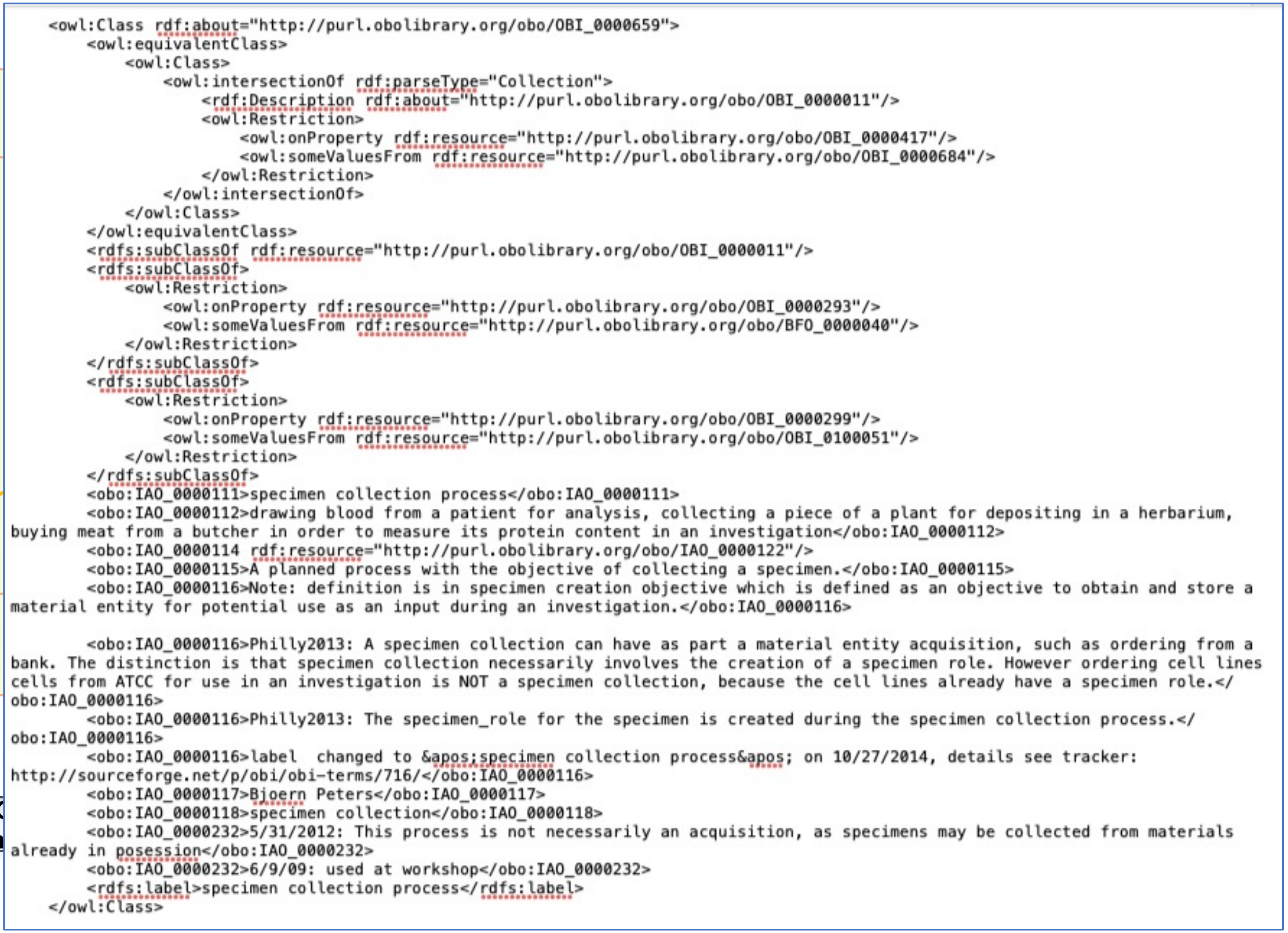
- Internationalized Resource Identifier (IRI). A Universal Resource Identifier or URI with an expanded character set.
  - Identifiers for finding things on the internet.
  - e.g. http://purl.obolibrary.org/obo/OBI_0000070 or OBO: OBI_0000070
  - Ontology terms are more than just the label (like “assay”)
- TBox (terminology): Expressions that contain universal statements about classes *
- ABox (assertion): Expressions that contain assertions about instances *



* L. Vogt Journal Biomedical Semantics 2021 12:20

Referring to reality: IRIs and classes (TBox) and instances (ABox)





```
    <owl:Class rdf:about="http://purl.obolibrary.org/obo/OBI_0000659">
        <owl:equivalentClass>
            <owl:Class>
                <owl:intersectionOf rdf:parseType="Collection">
                    <rdf:Description rdf:about="http://purl.obolibrary.org/obo/OBI_0000011"/>
                    <owl:Restriction>
                        <owl:onProperty rdf:resource="http://purl.obolibrary.org/obo/OBI_0000417"/>
                        <owl:someValuesFrom rdf:resource="http://purl.obolibrary.org/obo/OBI_0000684"/>
                    </owl:Restriction>
                </owl:intersectionOf>
            </owl:Class>
        </owl:equivalentClass>
        <rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/obo/OBI_0000011"/>
        <rdfs:subClassOf>
            <owl:Restriction>
                <owl:onProperty rdf:resource="http://purl.obolibrary.org/obo/OBI_0000293"/>
                <owl:someValuesFrom rdf:resource="http://purl.obolibrary.org/obo/BFO_0000040"/>
            </owl:Restriction>
        </rdfs:subClassOf>
        <rdfs:subClassOf>
            <owl:Restriction>
                <owl:onProperty rdf:resource="http://purl.obolibrary.org/obo/OBI_0000299"/>
                <owl:someValuesFrom rdf:resource="http://purl.obolibrary.org/obo/OBI_0100051"/>
            </owl:Restriction>
        </rdfs:subClassOf>
        <obo:IAO_0000111>specimen collection process</obo:IAO_0000111>
        <obo:IAO_0000112>drawing blood from a patient for analysis, collecting a piece of a plant for depositing in a herbarium,
buying meat from a butcher in order to measure its protein content in an investigation</obo:IAO_0000112>
        <obo:IAO_0000114 rdf:resource="http://purl.obolibrary.org/obo/IAO_0000122"/>
        <obo:IAO_0000115>A planned process with the objective of collecting a specimen.</obo:IAO_0000115>
        <obo:IAO_0000116>Note: definition is in specimen creation objective which is defined as an objective to obtain and store a
material entity for potential use as an input during an investigation.</obo:IAO_0000116>

        <obo:IAO_0000116>Philly2013: A specimen collection can have as part a material entity acquisition, such as ordering from a
bank. The distinction is that specimen collection necessarily involves the creation of a specimen role. However ordering cell lines
cells from ATCC for use in an investigation is NOT a specimen collection, because the cell lines already have a specimen role.</
obo:IAO_0000116>
        <obo:IAO_0000116>Philly2013: The specimen_role for the specimen is created during the specimen collection process.</
obo:IAO_0000116>
        <obo:IAO_0000116>label  changed to &apos;specimen collection process&apos; on 10/27/2014, details see tracker:
http://sourceforge.net/p/obi/obi-terms/716/</obo:IAO_0000116>
        <obo:IAO_0000117>Bjoern Peters</obo:IAO_0000117>
        <obo:IAO_0000118>specimen collection</obo:IAO_0000118>
        <obo:IAO_0000232>5/31/2012: This process is not necessarily an acquisition, as specimens may be collected from materials
already in posession</obo:IAO_0000232>
        <obo:IAO_0000232>6/9/09: used at workshop</obo:IAO_0000232>
        <rdfs:label>specimen collection process</rdfs:label>
    </owl:Class>
```

Bandrowski et al. The Ontology for Biomedical Investigations. PLoS One 2016

Semantic harmony for databases, community standards, and graphs

- Striving for semantic harmony in a large project with a small ontology team
  - The VEuPathDB story (annotations, IRIs)

VEuPathDB
Eukaryotic Pathogen, Vector & Host Informatics Resources

- Aspiring to common terms and patterns across diverse domains and with diverse ontology backgrounds.
  - OBI/OBO Foundry (classes, TBox)

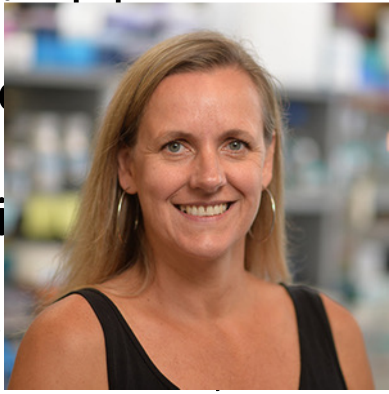
- Applying ontology-driven tools on real-world data
  - TURBO and semantic graphs (individuals, ABox triples)

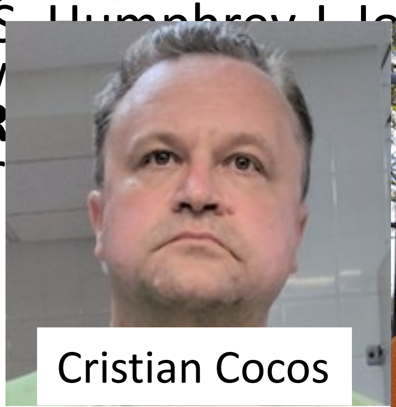
TURBO

Penn Medicine

VEuPathDB is a collection of eukaryotic pathogen, vector, and host informatics resources

- Primarily an **NIAID-supported Bioinformatics Resource Center**.
- Also funding from Wellcome Trust. Now an **Elixir-UK service** selected to ensure high service quality and match UK priorities as identified by its funders.
- Started in 2000 with the Plasmodium Genome Database - PlasmoDB (reflects a **long running project**) to provide researchers studying p omi
- Grew to include other ampicomplexans ar the
  ssful scalable infrastructure
- coechea C, Barba M, Barret levi
  stelli J, Brunk BP, Caddick M,
  GK, Crouch K, Davis K, DeBa D,
  Gajria B, Giraldo-Calderón GI, Harb OS, Harper E Hic
  Hu S, Humphrey J, Iodice J, Jones A, **Judkins** J, Ke C, Kwon DK,
  Law J, MacCallu **McDowell**
  J, **R** **Shanmugas** V, Spruill D,
  cker eltz S, Wiec Xu L, **Zheng**
  J. VEuPathDB: the host bioinf center.
  Nucleic Acids Res.

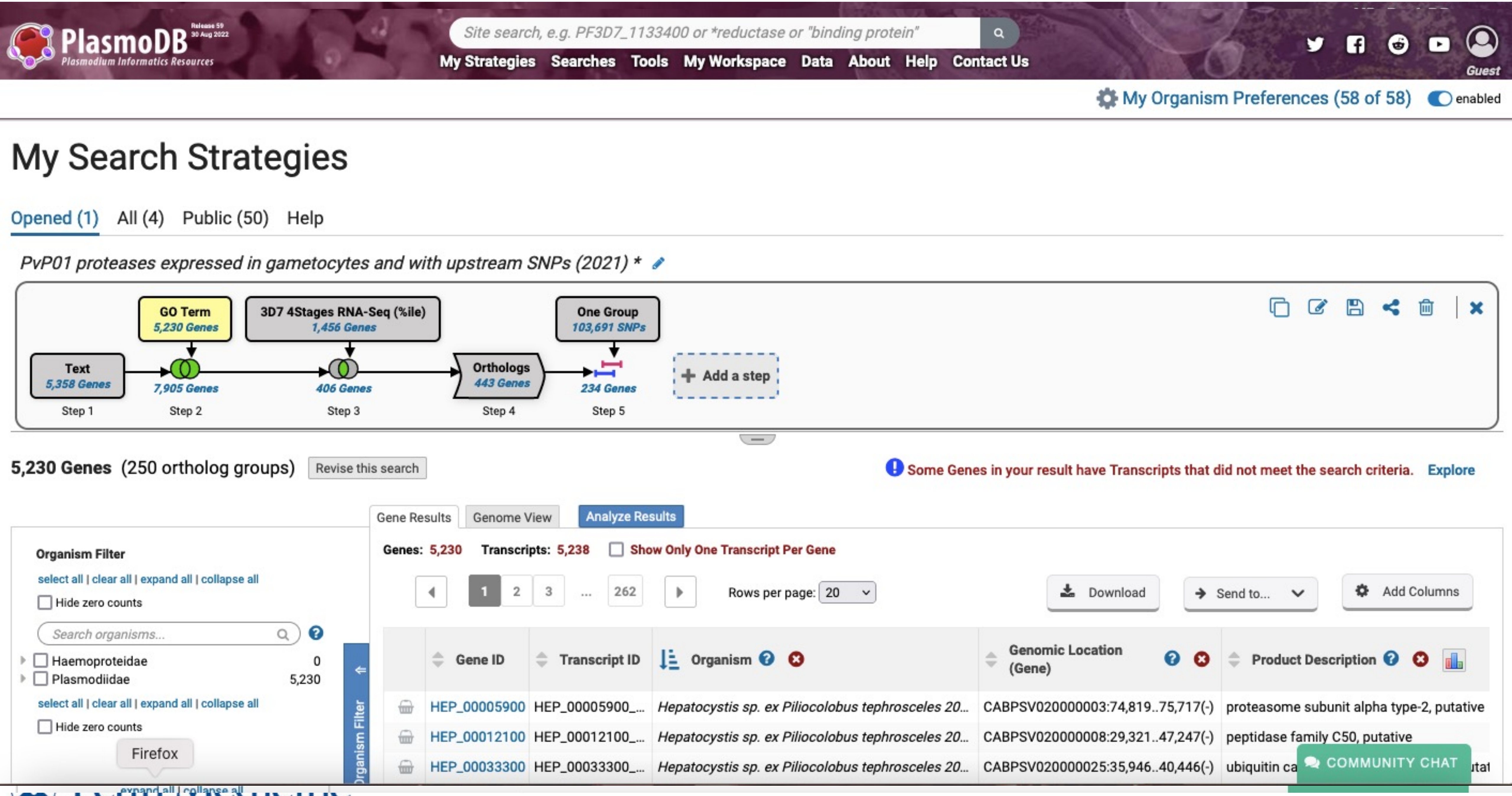




Cristian Cocos

Penn Medicine

hDB
ctor & Host Informatics Resources



PlasmoDB
Release 59
30 Aug 2022
Plasmodium Informatics Resources
Site search, e.g. PF3D7_1133400 or *reductase or "binding protein"
My Strategies
Searches
Tools
My Workspace
Data
About
Help
Contact Us
Guest
My Organism Preferences (58 of 58)
enabled

My Search Strategies

*PvP01 proteases expressed in gametocytes and with upstream SNPs (2021) **

5,230 Genes (250 ortholog groups) Revise this search

Some Genes in your result have Transcripts that did not meet the search criteria. Explore

Genes: 5,230 Transcripts: 5,238 □ Show Only One Transcript Per Gene

Organism Filter

select all | clear all | expand all | collapse all

□ Hide zero counts

□ Haemoproteidae	0
□ Plasmodiidae	5,230

select all | clear all | expand all | collapse all

□ Hide zero counts

Gene ID	Transcript ID	Organism	Genomic Location (Gene)	Product Description
HEP_00005900	HEP_00005900_...	*Hepatocystis sp. ex Piliocolobus tephrosceles 20...*	CABPSV020000003:74,819..75,717(-)	proteasome subunit alpha type-2, putative
HEP_00012100	HEP_00012100_...	*Hepatocystis sp. ex Piliocolobus tephrosceles 20...*	CABPSV020000008:29,321..47,247(-)	peptidase family C50, putative
HEP_00033300	HEP_00033300_...	*Hepatocystis sp. ex Piliocolobus tephrosceles 20...*	CABPSV020000025:35,946..40,446(-)	ubiquitin ca... utat

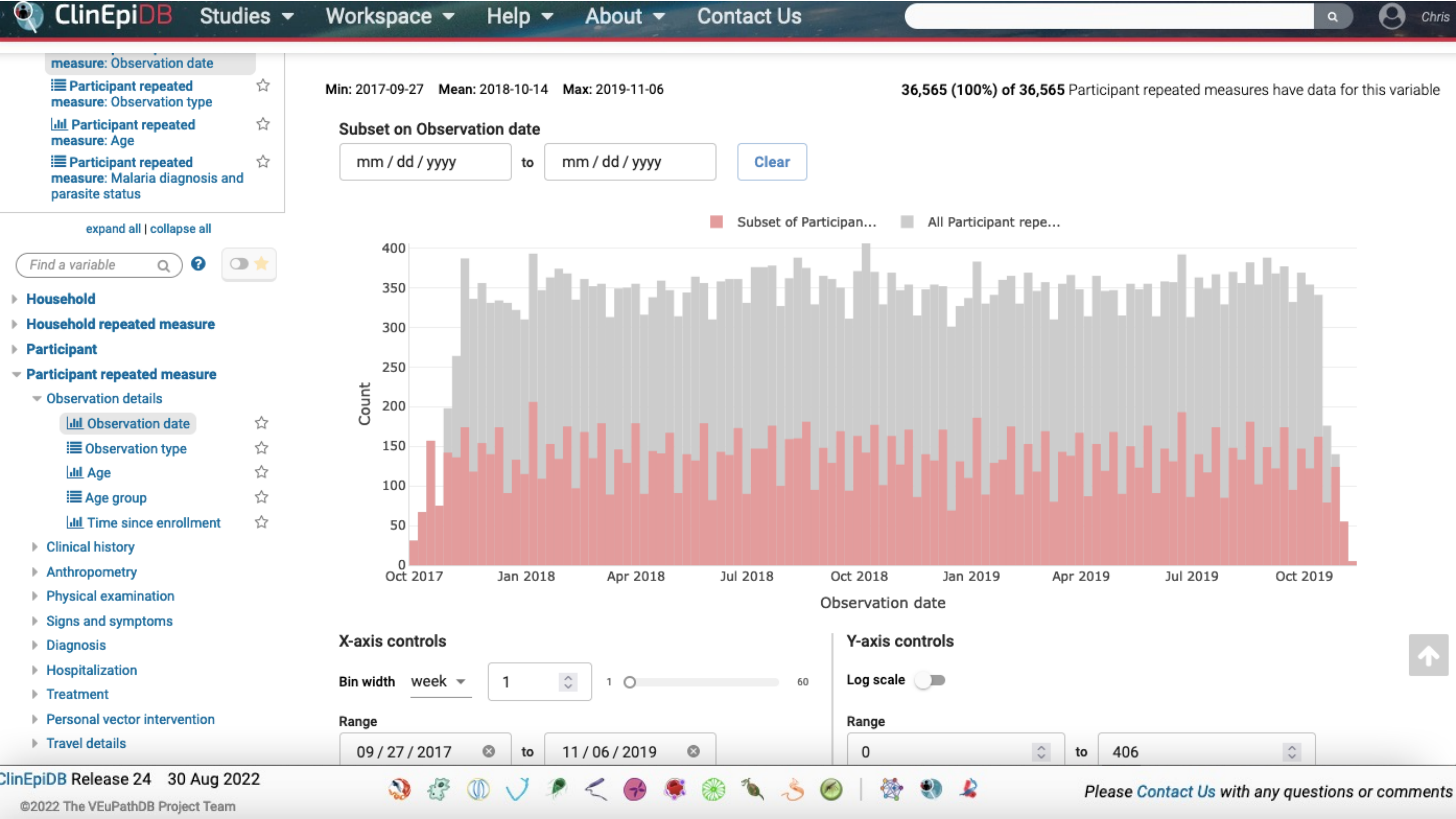
VEuPathDB has expanded to include ClinEpiDB and MicrobiomeDB

- Now working with datasets from clinical epidemiology (ClinEpiDB) and microbiome (MicrobiomeDB) s al) genomic datasets typically
h few "metadata" – **sa** **imental details**. Expanded system
graphi **en social** **g**
are gen **iDB and M** **a**

- ess clini database ou
mplex s ankaka E, e
OS, **Helb** Kissinger JC, **Lindsay** B, Roos DS, S rt
Zheng J, **Tomko** SS. Gates Open Res. 2020 Apr 6;3:1661. doi: 10.12688/gatesopenres.13087.2. eCollection 2019.

- Microbi tems biology platform for integrating, mining and analyzing microbi ents. Oliveira FS, Brestelli J, Cade S, Zheng J, Iodice J, Fischer S, Aurre issinger JC, Brunk BP, Stoeckert CJ Jr, Fernandes GR, Roos DS, **Beiting** cids Res. 2018 Jan 4;46(D1):D684-D691.

Penn Medi



measure: Observation date

Participant repeated measure: Observation type

Participant repeated measure: Age

Participant repeated measure: Malaria diagnosis and parasite status

expand all | collapse all

Find a variable

- Household
- Household repeated measure
- Participant
- Participant repeated measure
  - Observation details
    - Observation date
    - Observation type
    - Age
    - Age group
    - Time since enrollment
  - Clinical history
  - Anthropometry
  - Physical examination
  - Signs and symptoms
  - Diagnosis
  - Hospitalization
  - Treatment
  - Personal vector intervention
  - Travel details

Min: 2017-09-27 **Mean:** 2018-10-14 **Max:** 2019-11-06

36,565 (100%) of 36,565 Participant repeated measures have data for this variable

Subset on Observation date

mm / dd / yyyy to mm / dd / yyyy Clear

X-axis controls

Bin width week 1 1 60

Range

09 / 27 / 2017 to 11 / 06 / 2019

Y-axis controls

Log scale

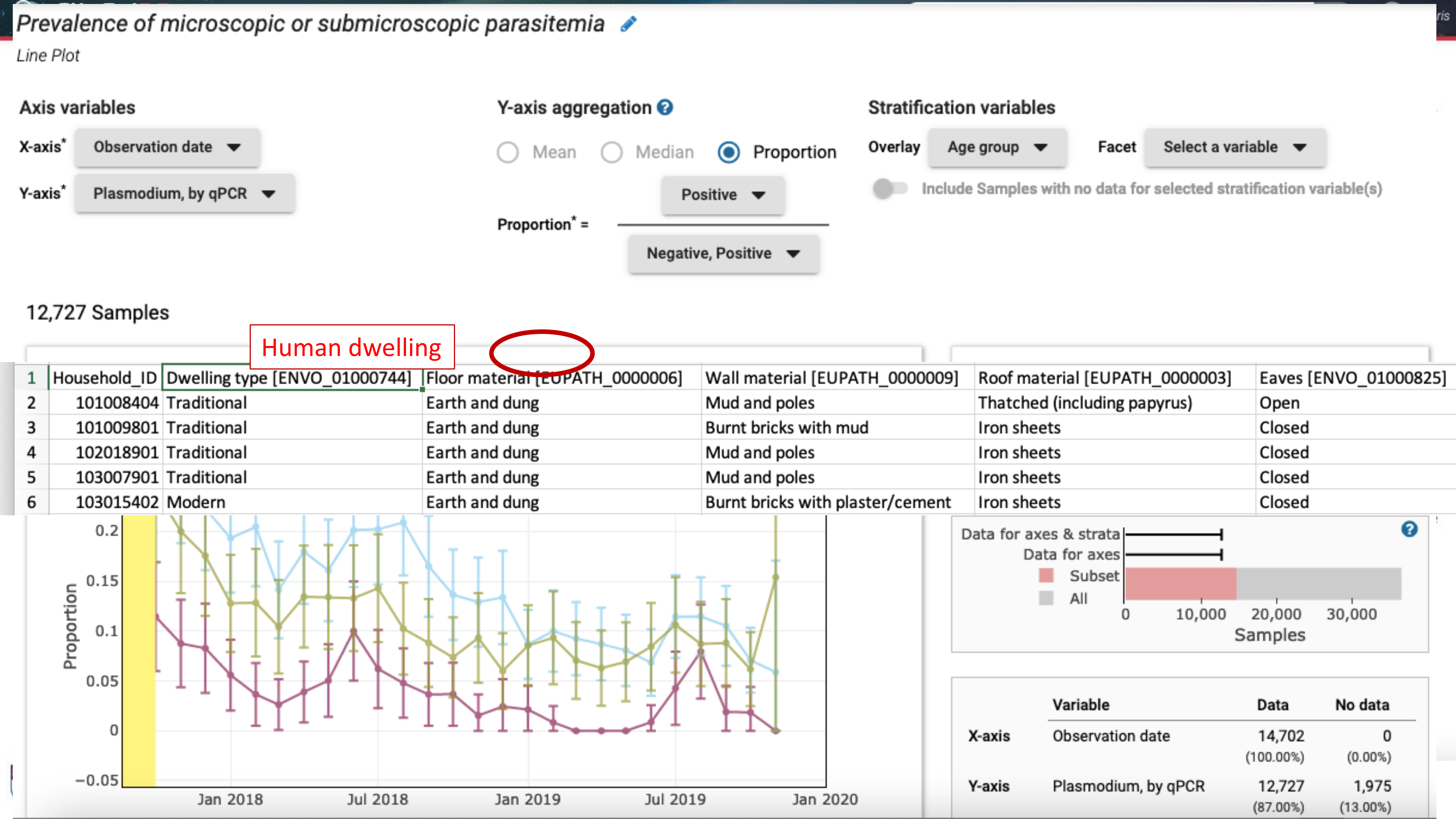
Range

0 to 406

ClinEpiDB Release 24 30 Aug 2022

©2022 The VEuPathDB Project Team

Please Contact Us with any questions or comments



Prevalence of microscopic or submicroscopic parasitemia

Line Plot

Axis variables

X-axis* Observation date

Y-axis* Plasmodium, by qPCR

Y-axis aggregation

○ Mean ○ Median ● Proportion

Proportion* = Positive / Negative, Positive

Stratification variables

Overlay Age group

Facet Select a variable

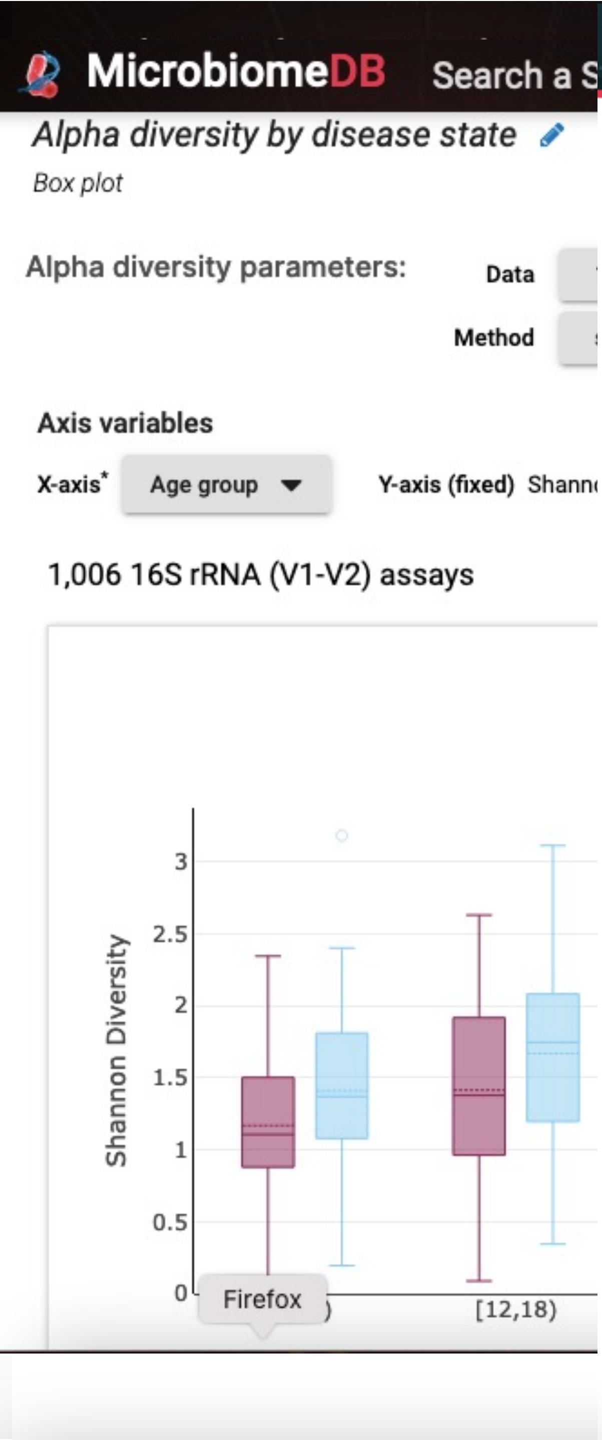
Include Samples with no data for selected stratification variable(s)

12,727 Samples

Human dwelling

1	Household_ID	Dwelling type [ENVO_01000744]	Floor material [EUPATH_0000006]	Wall material [EUPATH_0000009]	Roof material [EUPATH_0000003]	Eaves [ENVO_01000825]
2	101008404	Traditional	Earth and dung	Mud and poles	Thatched (including papyrus)	Open
3	101009801	Traditional	Earth and dung	Burnt bricks with mud	Iron sheets	Closed
4	102018901	Traditional	Earth and dung	Mud and poles	Iron sheets	Closed
5	103007901	Traditional	Earth and dung	Mud and poles	Iron sheets	Closed
6	103015402	Modern	Earth and dung	Burnt bricks with plaster/cement	Iron sheets	Closed

	Variable	Data	No data
X-axis	Observation date	14,702 (100.00%)	0 (0.00%)
Y-axis	Plasmodium, by qPCR	12,727 (87.00%)	1,975 (13.00%)



Alpha diversity by disease state

Box plot

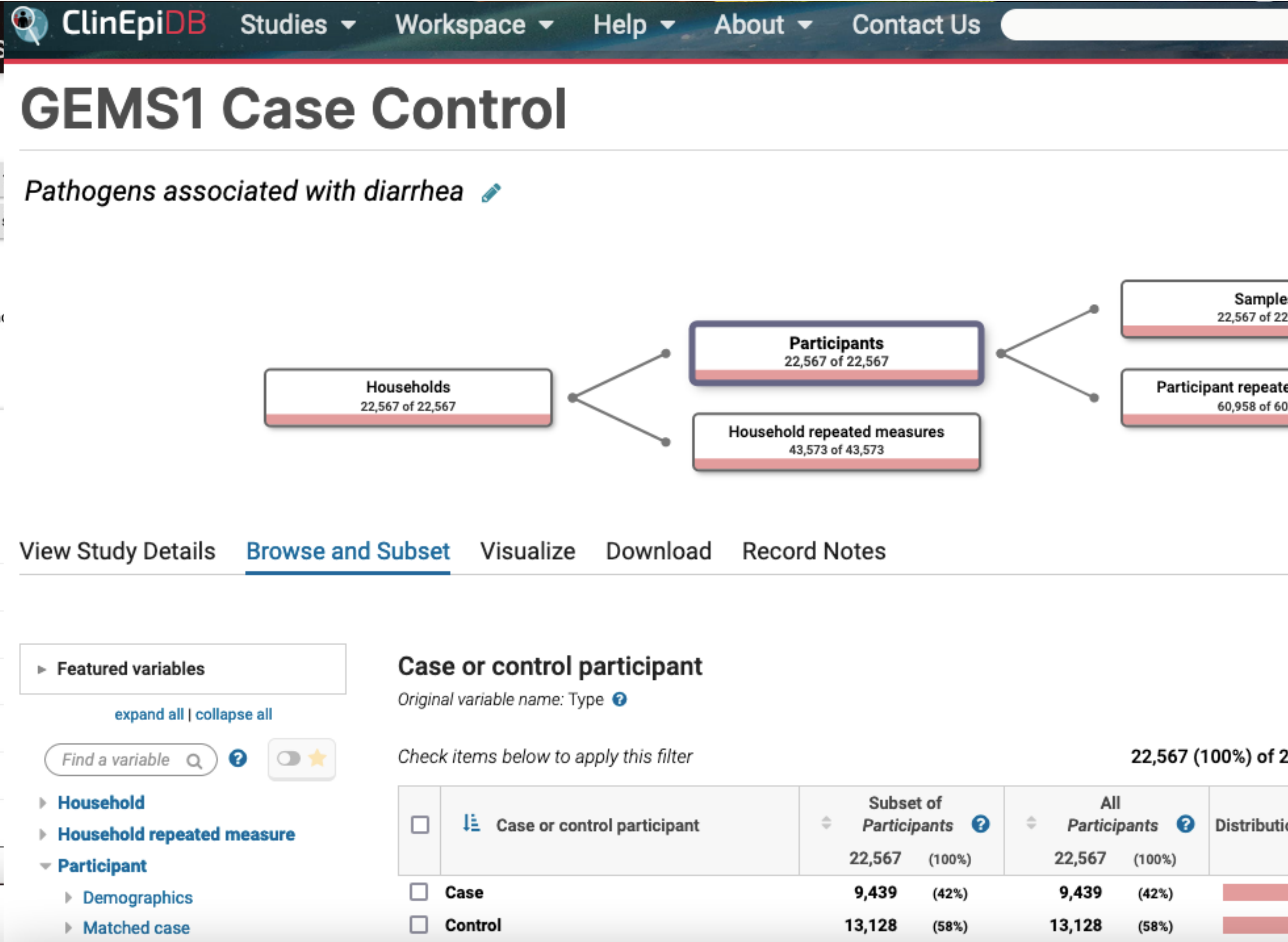
Alpha diversity parameters: Data

Method

Axis variables

X-axis* Age group Y-axis (fixed) Shann

1,006 16S rRNA (V1-V2) assays



GEMS1 Case Control

Pathogens associated with diarrhea

Case or control participant

Original variable name: Type

Check items below to apply this filter

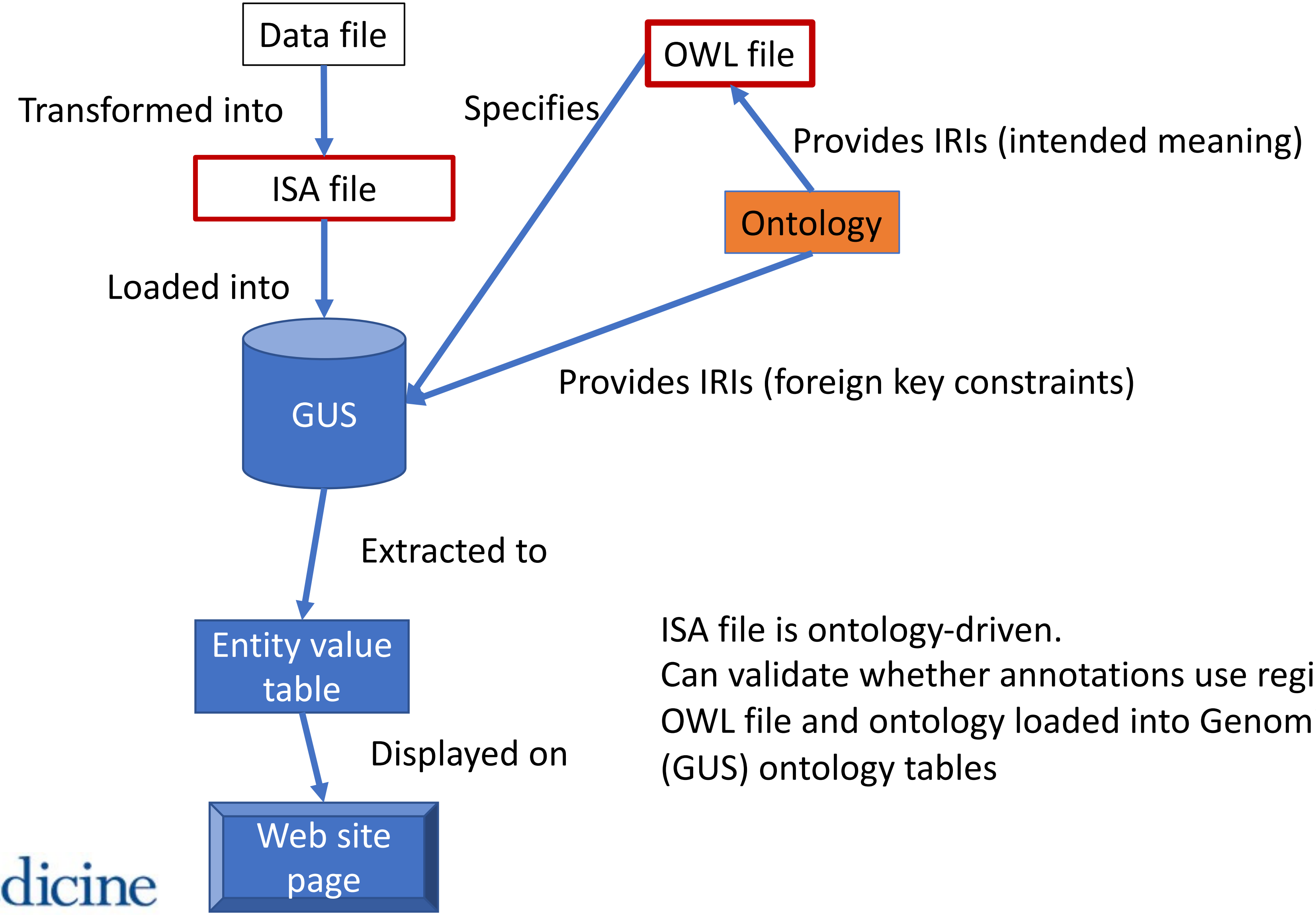
22,567 (100%) of 2

☐	Case or control participant	*Subset of Participants*		*All Participants*		Distributi
		22,567	(100%)	22,567	(100%)	
☐	Case	9,439	(42%)	9,439	(42%)	
☐	Control	13,128	(58%)	13,128	(58%)	

VEuPathDB Ontology supports the harmonization of annotations/data variables for genomic, epidemiological, and microbiome datasets

- All annotations/data variables with the same label are equivalent and have the same IRI
- Data variables are organized following the same web display hierarchy across sites (not the same as the ontology hierarchy).
- Goal is to provide consistency (i.e., semantic harmony) – terms have the same meaning everywhere and can be found in the same way.

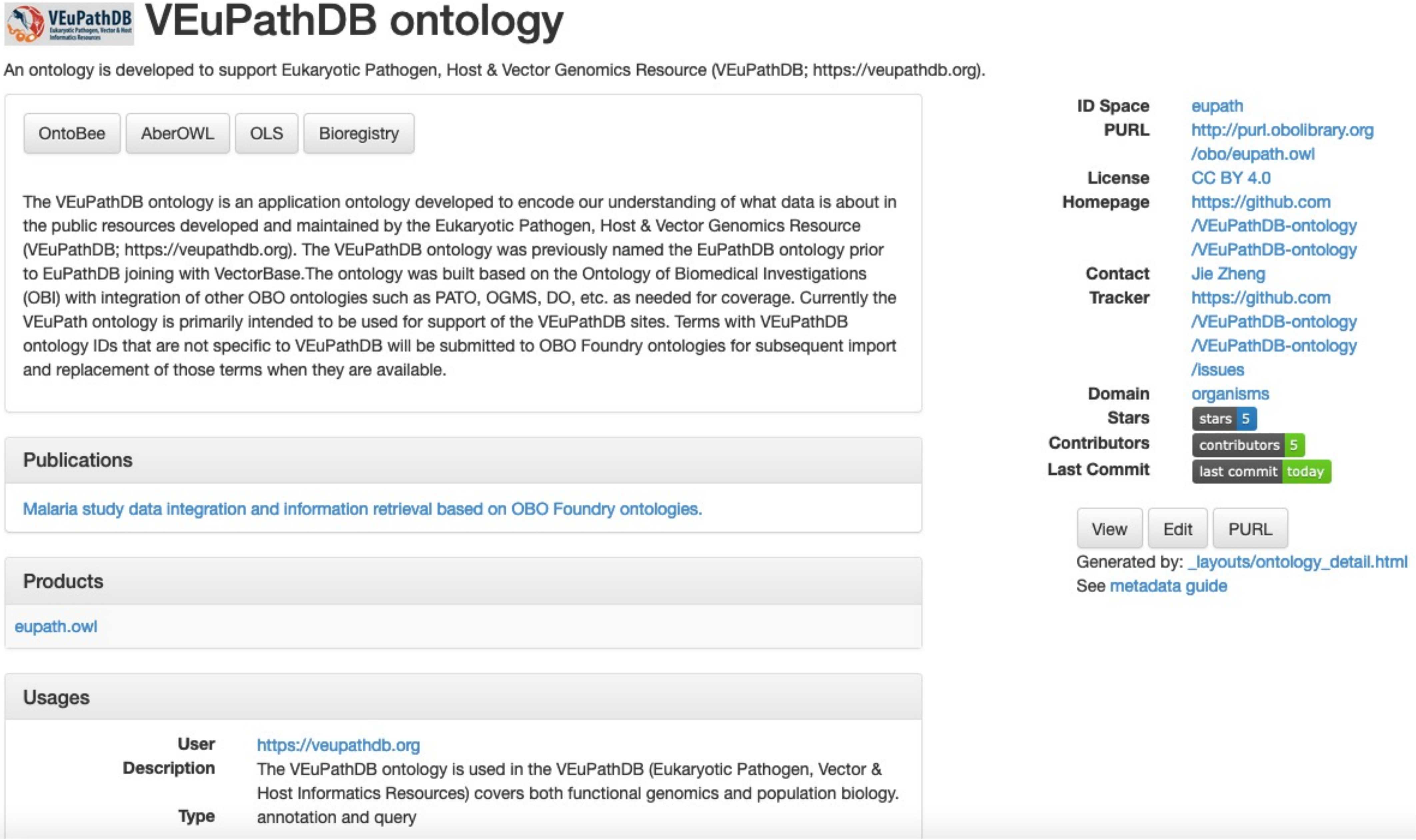
Semantic consistency of VEuPathDB dataset annotations is ontology-driven through "harmony" of IRIs during data loading



ISA file is ontology-driven.
Can validate whether annotations use registered ontology terms.
OWL file and ontology loaded into Genomics Unified Schema (GUS) ontology tables

Penn Medicine

https://obofoundry.org/ontology/eupath.html



VEuPathDB ontology

An ontology is developed to support Eukaryotic Pathogen, Host & Vector Genomics Resource (VEuPathDB; https://veupathdb.org).

The VEuPathDB ontology is an application ontology developed to encode our understanding of what data is about in the public resources developed and maintained by the Eukaryotic Pathogen, Host & Vector Genomics Resource (VEuPathDB; https://veupathdb.org). The VEuPathDB ontology was previously named the EuPathDB ontology prior to EuPathDB joining with VectorBase.The ontology was built based on the Ontology of Biomedical Investigations (OBI) with integration of other OBO ontologies such as PATO, OGMS, DO, etc. as needed for coverage. Currently the VEuPath ontology is primarily intended to be used for support of the VEuPathDB sites. Terms with VEuPathDB ontology IDs that are not specific to VEuPathDB will be submitted to OBO Foundry ontologies for subsequent import and replacement of those terms when they are available.

Publications

Malaria study data integration and information retrieval based on OBO Foundry ontologies.

Products

eupath.owl

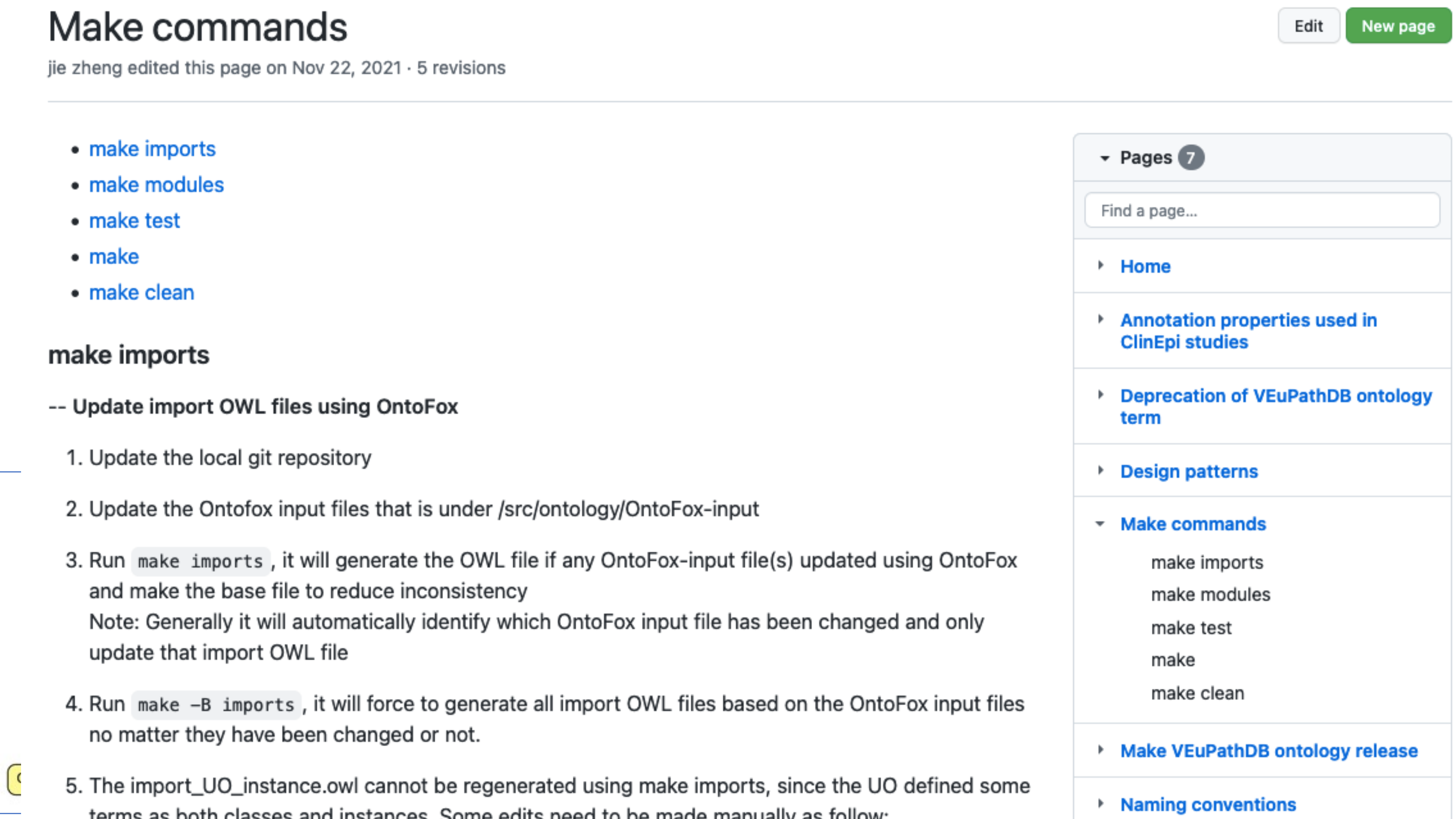
Usages

User	https://veupathdb.org
Description	The VEuPathDB ontology is used in the VEuPathDB (Eukaryotic Pathogen, Vector & Host Informatics Resources) covers both functional genomics and population biology.
Type	annotation and query

ID Space	eupath
PURL	http://purl.obolibrary.org/obo/eupath.owl
License	CC BY 4.0
Homepage	https://github.com/VEuPathDB-ontology/VEuPathDB-ontology
Contact	Jie Zheng
Tracker	https://github.com/VEuPathDB-ontology/VEuPathDB-ontology/issues
Domain	organisms
Stars	stars 5
Contributors	contributors 5
Last Commit	last commit today

Generated by: _layouts/ontology_detail.html

See metadata guide



Make commands

jie zheng edited this page on Nov 22, 2021 · 5 revisions

- make imports
- make modules
- make test
- make
- make clean

make imports

-- Update import OWL files using OntoFox

1. Update the local git repository
2. Update the Ontofox input files that is under /src/ontology/OntoFox-input
3. Run `make imports`, it will generate the OWL file if any OntoFox-input file(s) updated using OntoFox and make the base file to reduce inconsistency
   Note: Generally it will automatically identify which OntoFox input file has been changed and only update that import OWL file
4. Run `make -B imports`, it will force to generate all import OWL files based on the OntoFox input files no matter they have been changed or not.
5. The import_UO_instance.owl cannot be regenerated using make imports, since the UO defined some terms as both classes and instances. Some edits need to be made manually as follow:

Pages 7

Find a page…

Home

Annotation properties used in ClinEpi studies

Deprecation of VEuPathDB ontology term

Design patterns

Make commands

make imports

make modules

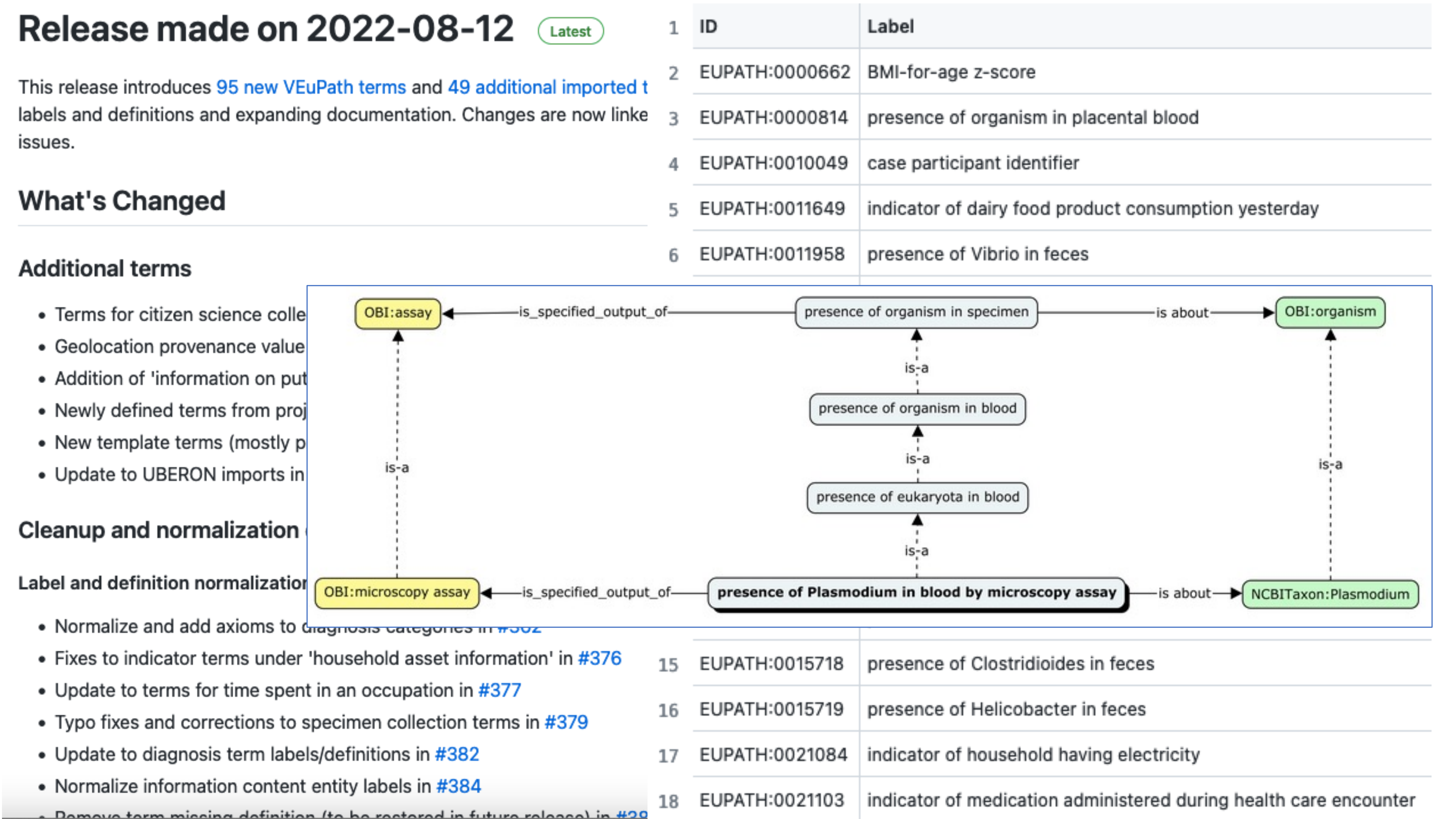
make test

make

make clean

Make VEuPathDB ontology release

Naming conventions



Release made on 2022-08-12

Latest

This release introduces 95 new VEuPath terms and 49 additional imported t
labels and definitions and expanding documentation. Changes are now linke
issues.

What's Changed

Additional terms

- Terms for citizen science colle
- Geolocation provenance value
- Addition of 'information on put
- Newly defined terms from proj
- New template terms (mostly p
- Update to UBERON imports in

Cleanup and normalization

Label and definition normalization

- Normalize and add axioms to diagnosis categories in #362
- Fixes to indicator terms under 'household asset information' in #376
- Update to terms for time spent in an occupation in #377
- Typo fixes and corrections to specimen collection terms in #379
- Update to diagnosis term labels/definitions in #382
- Normalize information content entity labels in #384
- Remove term missing definition (to be restored in future release) in #38

ID	Label
EUPATH:0000662	BMI-for-age z-score
EUPATH:0000814	presence of organism in placental blood
EUPATH:0010049	case participant identifier
EUPATH:0011649	indicator of dairy food product consumption yesterday
EUPATH:0011958	presence of Vibrio in feces
EUPATH:0015718	presence of Clostridioides in feces
EUPATH:0015719	presence of Helicobacter in feces
EUPATH:0021084	indicator of household having electricity
EUPATH:0021103	indicator of medication administered during health care encounter

OBI:assay
is_specified_output_of
presence of organism in specimen
is about
OBI:organism
is-a
presence of organism in blood
is-a
presence of eukaryota in blood
is-a
is-a
is-a
OBI:microscopy assay
is_specified_output_of
presence of Plasmodium in blood by microscopy assay
is about
NCBITaxon:Plasmodium

The VEuPathDB Ontology is an application ontology supporting VEuPathDB resources

- Not trying to capture parasite biology. Employs other ontologies like the the Ontology of Parasite Lifecycles (OPL) to do that.
- Primarily reuses terms from OBO Foundry. Imports terms from over 50 sources.
- However, ~1900 terms out of ~5000 have EUPATH prefix (created by VEuPathDB) so there are a lot of terms only in the VEuPathDB ontology!
- Many of these are placeholders. Active effort to get these in domain / reference ontologies.
- Many are precomposed terms reflecting researcher-based distinctions.
  - E.g. ’presence of Plasmodium in blood by microscopy assay’, ‘specimen used for DNA PCR’
  - Universal or arbitrary grouping?

What is the value of having this VEuPathDB "application" ontology in the OBO Foundry?

- DISCLAIMER: These are my views and don't reflect the OBO Foundry!
- Provides a basis for shared terms typically to be put in a domain/ reference ontology.
    - Get to share how we are ontologizing data from surveys and case report forms.
    - Show the design patterns we've come up with and naming conventions.
- Is it ever OK to reuse our classes outside our projects? Why do we have them?
    - Need terms for database releases on different schedule from external ontology releases.
    - Contact us if you have need for the same term and we'll work together to get it in the most relevant ontology. E.g., OBI, ENVO, OMRSE, PRO
- Note this raises the issue of ontology class expressions (TBox) for 'universals' as opposed to arbitrary groupings.
    - Do classes for arbitrary groupings belong in reference ontologies?
    - Probably not, but do need a home because still be of general use and therefore has value for reuse.

Aspiring to common terms and patterns across diverse domains and diverse ontology backgrounds with OBI

OBI
Ontology for Biomedical Investigations

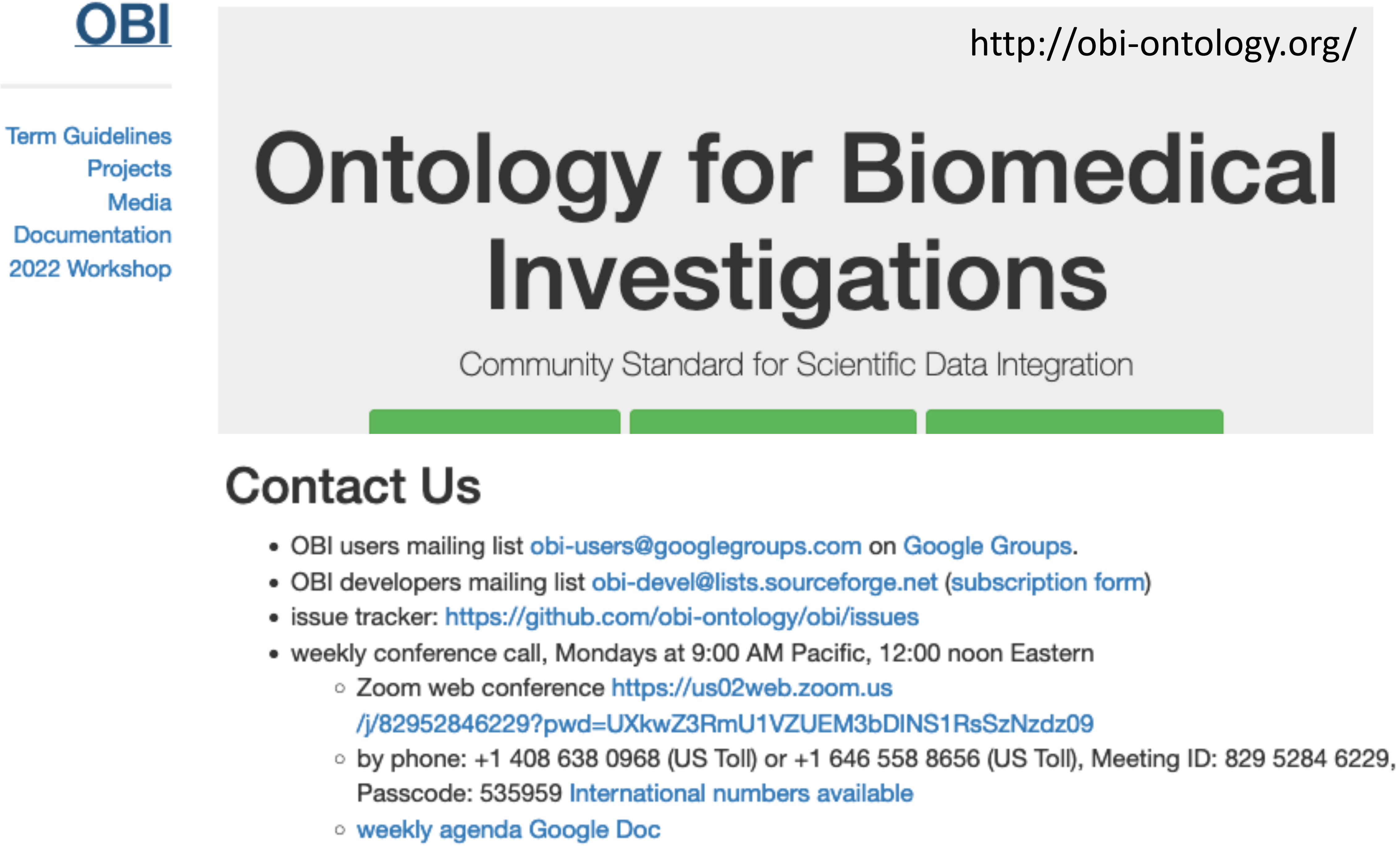
- We could not have made the progress we have in VEuPathDB using ontologies without OBI and the OBO Foundry.
- OBI is the Ontology for Biomedical Investigations and arose initially from the MGED ontology to cover microarray assays and then FuGO, functional genomics ontology.
  - Recognition that different technologies generating large datasets had a common need to describe experimental conditions, protocols, and designs.
  - The same specimens could be used for transcriptomics, proteomics, and metabolomics experiments.

Think OBI first when developing ontology terms related to peforming research!

Ontology for Biomedical Investigations

- Community benefits of working with OBI (extends to other OBOF)
    - Get broader input, more likely to be interoperable, and bigger impact
    - Weekly meetings where you can champion your terms
    - Not just an issue tracker but also GitHub pull requests (can track your terms!)
- Vita R, Zheng J, Jackson R, Dooley D, Overton JA, Miller MA, Berrios DC, Scheuermann RH, He Y, McGinty HK, Brochhausen M, Lin AY, Jain SB, Chibucos MC, Judkins J, Giglio MG, Feng IY, Burns G, Brush MH, **Peters B**, Stoeckert CJ Jr. Standardization of assay representation in the Ontology for Biomedical Investigations. Database (Oxford). 2021 Jul 9;2021:baab040.
    - Design patterns and ROBOT templates
    - Working together to share and apply best practices
    - "Both the ontology terms and the OBI community were improved through this **collaborative community effort,** which made developers more aware of terms outside their area of expertise and gave them a better understanding of assay terms as a whole."

http://obi-ontology.org/



OBI

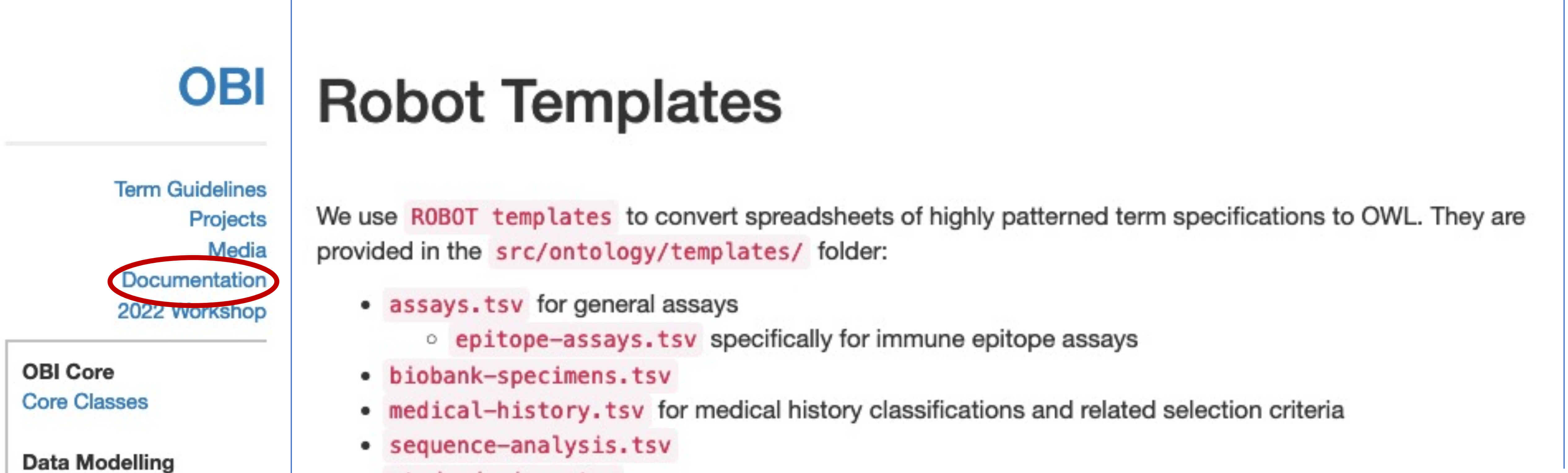
Term Guidelines
Projects
Media
Documentation
2022 Workshop

Ontology for Biomedical Investigations

Community Standard for Scientific Data Integration

Contact Us

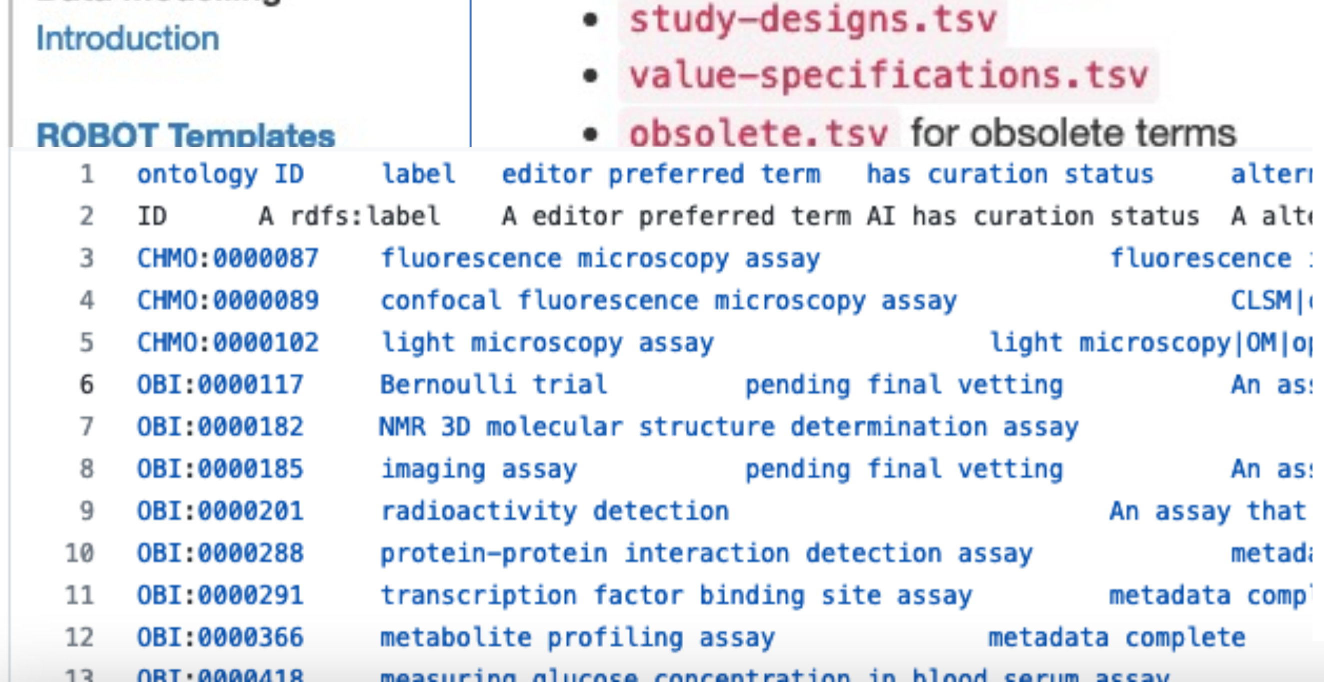
- OBI users mailing list obi-users@googlegroups.com on Google Groups.
- OBI developers mailing list obi-devel@lists.sourceforge.net (subscription form)
- issue tracker: https://github.com/obi-ontology/obi/issues
- weekly conference call, Mondays at 9:00 AM Pacific, 12:00 noon Eastern
    - Zoom web conference https://us02web.zoom.us/j/82952846229?pwd=UXkwZ3RmU1VZUEM3bDlNS1RsSzNzdz09
    - by phone: +1 408 638 0968 (US Toll) or +1 646 558 8656 (US Toll), Meeting ID: 829 5284 6229, Passcode: 535959 International numbers available
    - weekly agenda Google Doc



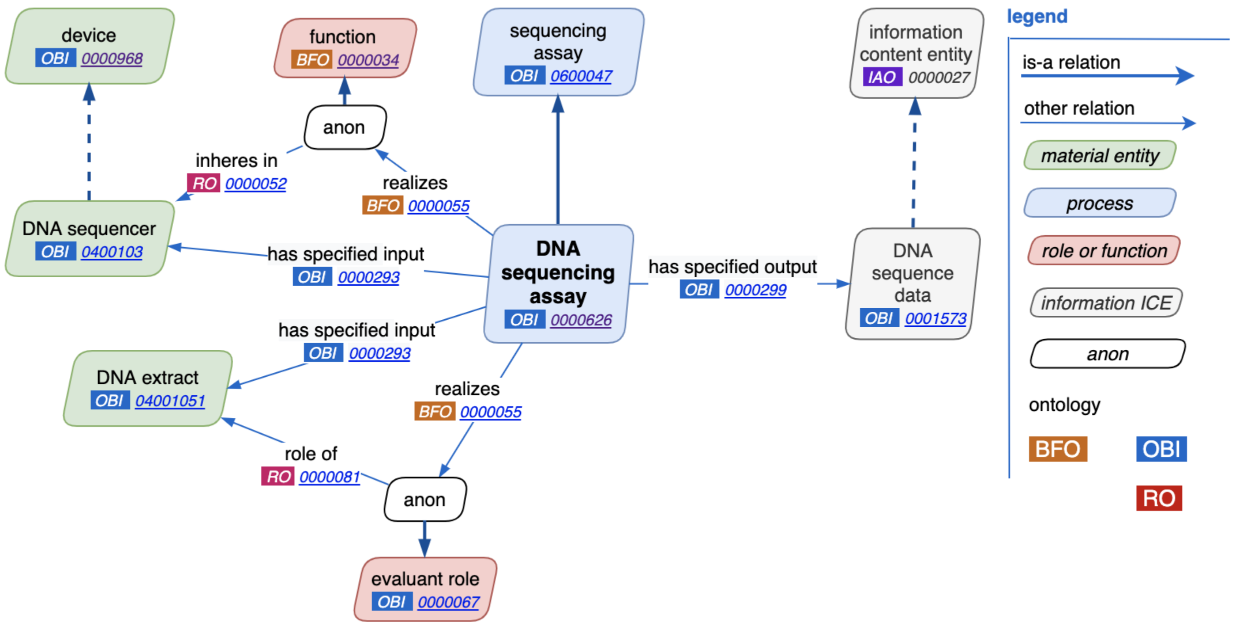
Robot Templates

We use `ROBOT templates` to convert spreadsheets of highly patterned term specifications to OWL. They are provided in the `src/ontology/templates/` folder:

- `assays.tsv` for general assays
  - `epitope-assays.tsv` specifically for immune epitope assays
- `biobank-specimens.tsv`
- `medical-history.tsv` for medical history classifications and related selection criteria
- `sequence-analysis.tsv`
- `study-designs.tsv`
- `value-specifications.tsv`
- `obsolete.tsv` for obsolete terms



```
ontology ID    label    editor preferred term    has curation status    altern
ID    A rdfs:label    A editor preferred term AI has curation status    A alte
CHMO:0000087    fluorescence microscopy assay    fluorescence 
CHMO:0000089    confocal fluorescence microscopy assay    CLSM|
CHMO:0000102    light microscopy assay    light microscopy|OM|op
OBI:0000117    Bernoulli trial    pending final vetting    An ass
OBI:0000182    NMR 3D molecular structure determination assay
OBI:0000185    imaging assay    pending final vetting    An ass
OBI:0000201    radioactivity detection    An assay that
OBI:0000288    protein-protein interaction detection assay    metada
OBI:0000291    transcription factor binding site assay    metadata comp
OBI:0000366    metabolite profiling assay    metadata complete    metabolite assay|metabolomic assay    An assay that detects and identifies ch
OBI:0000418    measuring glucose concentration in blood serum assay    metadata incomplete    An assay that determines the concentration of g
```



Aspiring to common terms and patterns across diverse domains and ontology backgrounds with the OBO Foundry

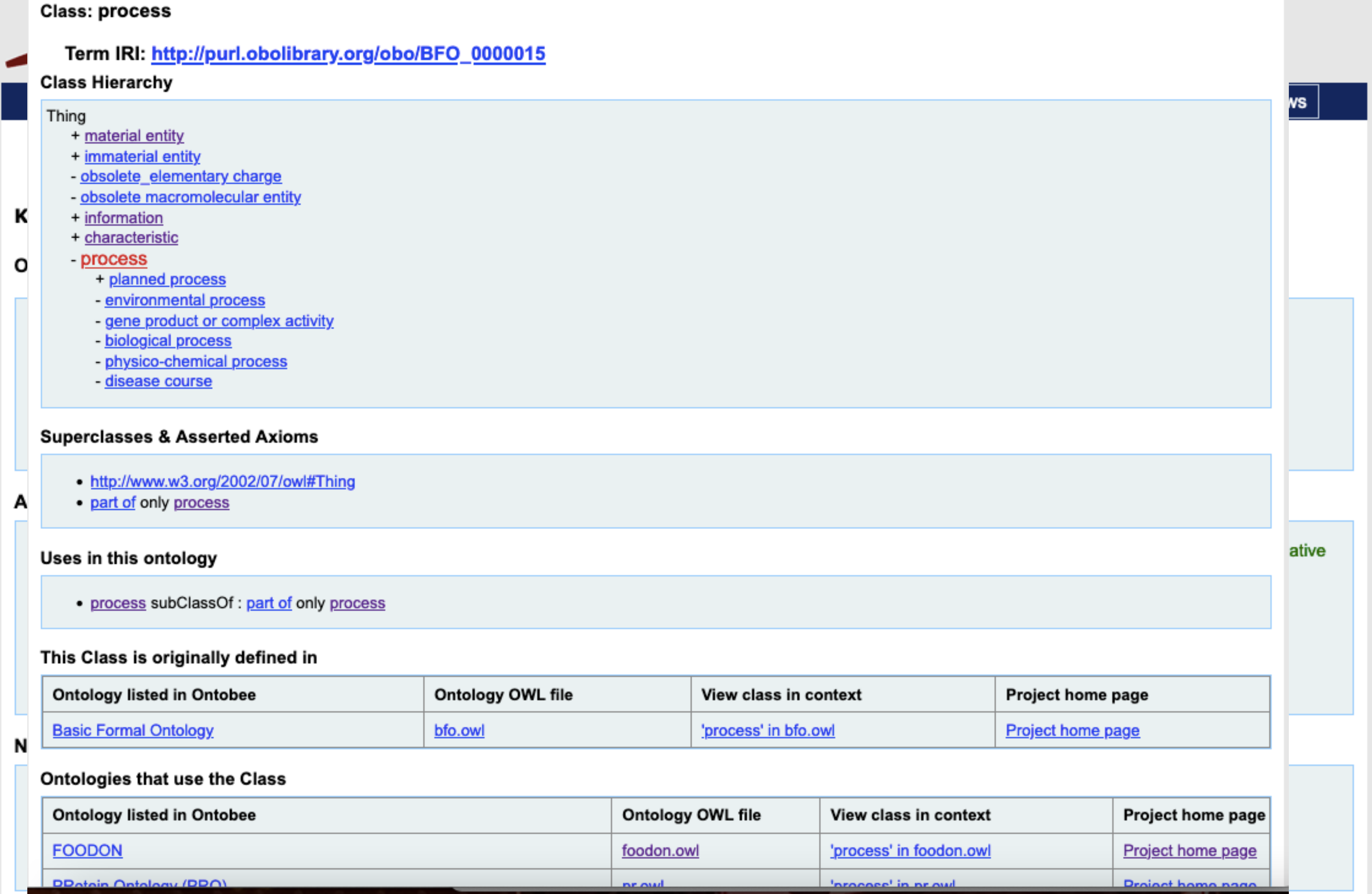
- OBO Foundry (OBOF) arose to provide interoperability for ontologies covering different domains
  - Requires principles for building and documenting the ontologies
  - Benefits from a common upper level and relations
  - Mainly for reference ontologies but many project-based ontologies have become involved
- Ontologies for complex data can cross many domains and require a meta-community like the OBOF to get coverage at least at a mid to upper level.
- OBO Foundry arose out of a recognition that communities should work together on ontologies.

OBOF is about representing what happened in reality not how things are stored in a database (or "cognitive representations on the part of domain experts”).*

- BFO (Basic Formal Ontology) provides an upper level ontology to distinguish between material entities, processes, and information.
  - Now an ISO standard (in First Order Logic - not just OWL, thank you Alan Ruttenberg!)
  - Not the only upper level ontology and not everyone is happy about exposing non-ontologists to ‘continuants’ and ‘occurrents’ so still part of the struggle for semantic harmony across communities.
- But we all use the Relations Ontology (RO)!
  - A requirement for being part of the OBO Foundry.
  - Connections (object properties) between ontology terms are often between different domains and RO provides the primary ones.
  - RO used in TBox definitions but are for ABox assertions on Instances of classes

Penn Medicine

* Ceusters, Smith. A realism-based approach to the evolution of biomedical ontologies. AMIA Annu Symp Proc. 2006;2006:121-5.



Class: process

Term IRI: http://purl.obolibrary.org/obo/BFO_0000015

Class Hierarchy

Thing
- + material entity
- + immaterial entity
- - obsolete_elementary charge
- - obsolete macromolecular entity
- + information
- + characteristic
- - process
    - + planned process
    - - environmental process
    - - gene product or complex activity
    - - biological process
    - - physico-chemical process
    - - disease course

Superclasses & Asserted Axioms

- http://www.w3.org/2002/07/owl#Thing
- part of only process

Uses in this ontology

- process subClassOf : part of only process

This Class is originally defined in

Ontology listed in Ontobee	Ontology OWL file	View class in context	Project home page
Basic Formal Ontology	bfo.owl	'process' in bfo.owl	Project home page

Ontologies that use the Class

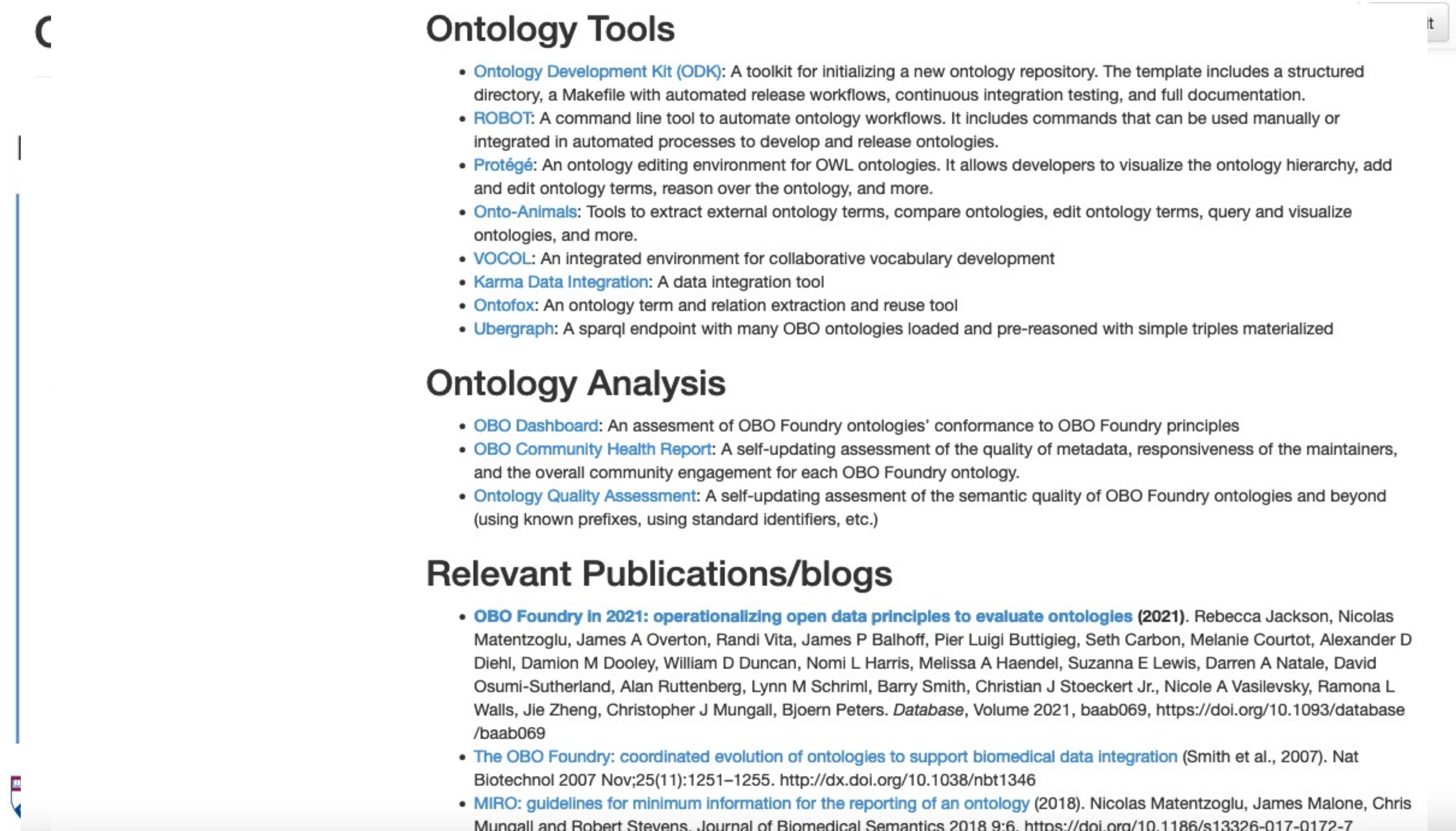
Ontology listed in Ontobee	Ontology OWL file	View class in context	Project home page
FOODON	foodon.owl	'process' in foodon.owl	Project home page
PRotein Ontology (PRO)	pr.owl	'process' in pr.owl	Project home page

Harmonizing communities: OBO Foundry

- Longstanding mechanisms for participation.
  - OBOF web site
  - OBO-discuss mail list
  - Meetings like ICBO

Harmonizing communities: OBO Foundry

- New ways to be part of the community.
  - Slack https://obo-communitygroup.slack.com
    - >50 channels! >240 members in the general channel!
    - Ontologies (e.g., COB), technologies (e.g., SPARQL), tools (e.g., ODK), resources (e.g., Jobs)
    - Governance
  - On-line resources
    - Tool tutorials
    - OBO Semantic Engineering Training https://oboacademy.github.io/obook/ (Open Biological and Biomedical Ontologies Organized Knowledge)
- Please attend the OBO Foundry Town Hall on Wednesday afternoon, Sept. 28.
  - Operations Committee
  - Governance



Ontology Tools

- Ontology Development Kit (ODK): A toolkit for initializing a new ontology repository. The template includes a structured directory, a Makefile with automated release workflows, continuous integration testing, and full documentation.
- ROBOT: A command line tool to automate ontology workflows. It includes commands that can be used manually or integrated in automated processes to develop and release ontologies.
- Protégé: An ontology editing environment for OWL ontologies. It allows developers to visualize the ontology hierarchy, add and edit ontology terms, reason over the ontology, and more.
- Onto-Animals: Tools to extract external ontology terms, compare ontologies, edit ontology terms, query and visualize ontologies, and more.
- VOCOL: An integrated environment for collaborative vocabulary development
- Karma Data Integration: A data integration tool
- Ontofox: An ontology term and relation extraction and reuse tool
- Ubergraph: A sparql endpoint with many OBO ontologies loaded and pre-reasoned with simple triples materialized

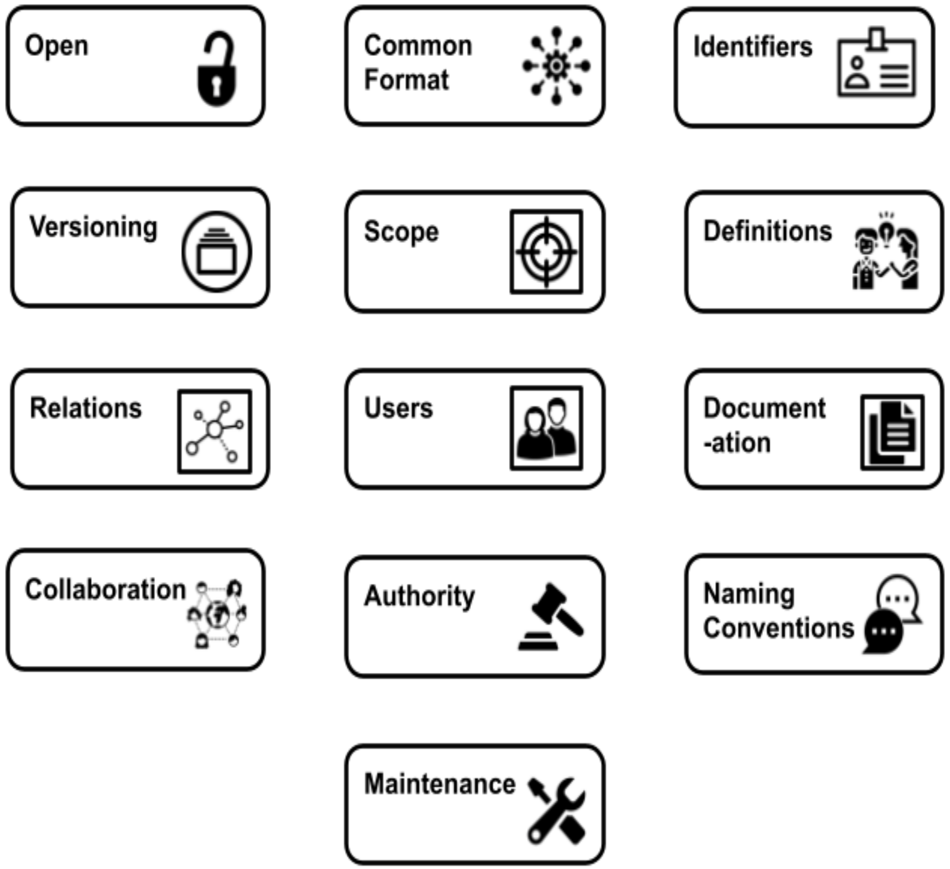
Ontology Analysis

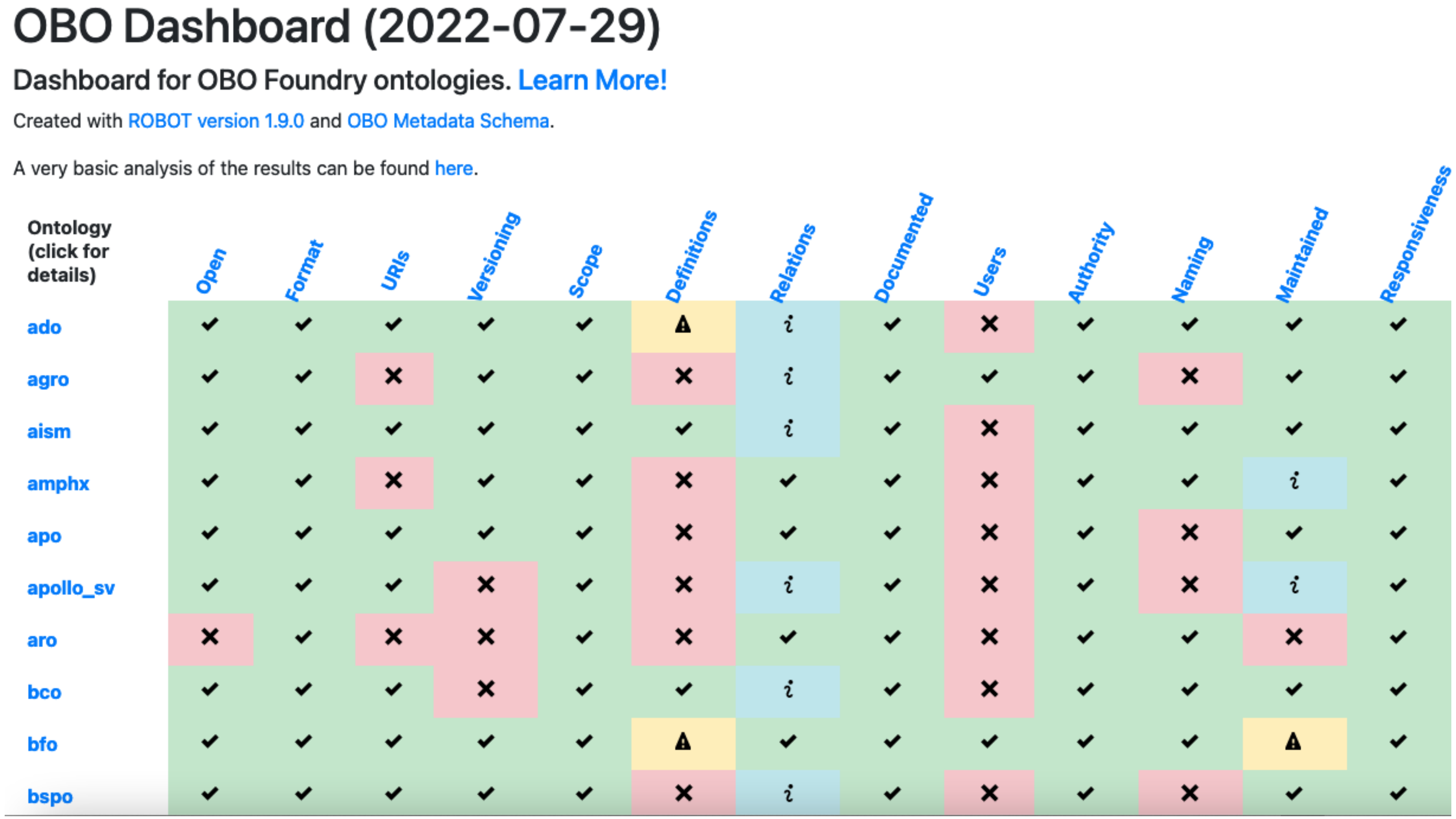
- OBO Dashboard: An assesment of OBO Foundry ontologies' conformance to OBO Foundry principles
- OBO Community Health Report: A self-updating assessment of the quality of metadata, responsiveness of the maintainers, and the overall community engagement for each OBO Foundry ontology.
- Ontology Quality Assessment: A self-updating assesment of the semantic quality of OBO Foundry ontologies and beyond (using known prefixes, using standard identifiers, etc.)

Relevant Publications/blogs

- **OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies (2021)**. Rebecca Jackson, Nicolas Matentzoglu, James A Overton, Randi Vita, James P Balhoff, Pier Luigi Buttigieg, Seth Carbon, Melanie Courtot, Alexander D Diehl, Damion M Dooley, William D Duncan, Nomi L Harris, Melissa A Haendel, Suzanna E Lewis, Darren A Natale, David Osumi-Sutherland, Alan Ruttenberg, Lynn M Schriml, Barry Smith, Christian J Stoeckert Jr., Nicole A Vasilevsky, Ramona L Walls, Jie Zheng, Christopher J Mungall, Bjoern Peters. *Database*, Volume 2021, baab069, https://doi.org/10.1093/database/baab069
- The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration (Smith et al., 2007). Nat Biotechnol 2007 Nov;25(11):1251–1255. http://dx.doi.org/10.1038/nbt1346
- MIRO: guidelines for minimum information for the reporting of an ontology (2018). Nicolas Matentzoglu, James Malone, Chris Mungall and Robert Stevens. Journal of Biomedical Semantics 2018 9:6. https://doi.org/10.1186/s13326-017-0172-7

OBO Foundry recent activities have included working on operationalizing principles





OBO Dashboard (2022-07-29)

Dashboard for OBO Foundry ontologies. Learn More!

Created with ROBOT version 1.9.0 and OBO Metadata Schema.

A very basic analysis of the results can be found here.

Ontology (click for details)	Open	Format	URIs	Versioning	Scope	Definitions	Relations	Documented	Users	Authority	Naming	Maintained	Responsiveness
ado	✓	✓	✓	✓	✓	⚠	ℹ	✓	✗	✓	✓	✓	✓
agro	✓	✓	✗	✓	✓	✗	ℹ	✓	✓	✓	✗	✓	✓
aism	✓	✓	✓	✓	✓	✓	ℹ	✓	✗	✓	✓	✓	✓
amphx	✓	✓	✗	✓	✓	✗	✓	✓	✗	✓	✓	ℹ	✓
apo	✓	✓	✓	✓	✓	✗	✓	✓	✗	✓	✗	✓	✓
apollo_sv	✓	✓	✓	✗	✓	✗	ℹ	✓	✗	✓	✗	ℹ	✓
aro	✗	✓	✗	✗	✓	✗	✓	✓	✗	✓	✓	✗	✓
bco	✓	✓	✓	✗	✓	✓	ℹ	✓	✗	✓	✓	✓	✓
bfo	✓	✓	✓	✓	✓	⚠	✓	✓	✓	✓	✓	⚠	✓
bspo	✓	✓	✓	✓	✓	✗	ℹ	✓	✗	✓	✗	✓	✓

Challenge of interoperating with other communities

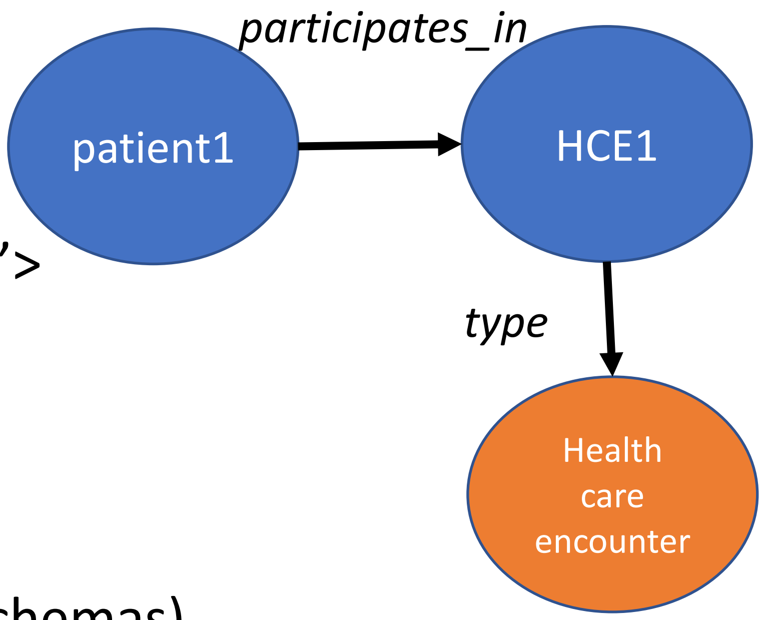
- Criteria for joining the OBO Foundry.
  - Scope: Use case relevant to life sciences.
  - Balance of inclusion and quality.
  - SOP for reviewers on new ontology requests is now available at https://obofoundry.org/docs/SOP.html (open for improvement).
- Ontology / Semantic web communities:
  - Interoperability of OBOF principles and tools with other standards.
  - What are they? What are current connections? ENVO yes but Financial Industry Business Ontology (FIBO)? SNOMED? FHIR?
  - Presence at ICBO / joint conferences (e.g., with US2TS this year) has been a step in the right direction but unclear what progress comes of that.

Applying ontology-driven tools on real-world data: The TURBO story

- Real-World Data (RWD): "data relating to patient health status and/or the delivery of health care routinely collected from a variety of sources.”
  - https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence
  - Also RWD is typically messy! (inconsistent in use of fields and incomplete)
- Billing codes (ICD) and problem lists (SNOMED, LOINC) capture the category for describing the patient diagnoses and findings but don’t provide a representation of the patient and the clinical experience (or a biobank specimen and its history and status).
- Biobank and clinical data needs harmonization and integration with other data and are **focused on individuals** so a good ABox fit.

Real-world data on individuals can best be represented as triples and captured as graphs

- Including relations enables making explicit the implicit connections between data records.
  - Triples provide a mechanism to accomplish this:
    - <subject> <predicate> <object>
  - <patient1> obo:RO_0000056#participates_in <HCE1>
  - <HCE1> rdfs:type <obo:OGMS_0000097#'health care encounter'>



- ***Modeling of real-world data is best done by ontological realism.****
  - Transforms data from different sources (and typically different schemas) to a common representation of what happened in the real world.

Penn Medicine

Ceusters. The place of Referent Tracking in Biomedical Informatics https://osf.io/q8hts/

PennTURBO aims to facilitate clinical research through semantically rich representations of RWD

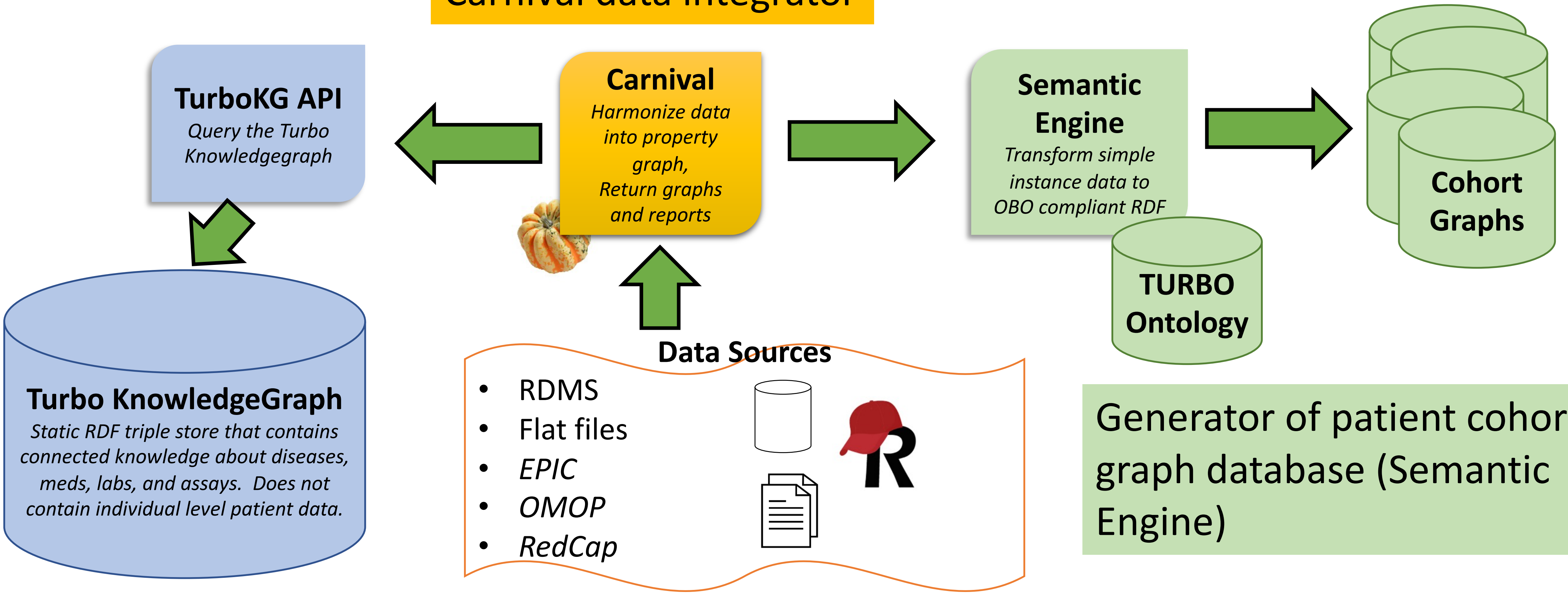
- TURBO stands for Transforming and Unifying Research with Biomedical Ontologies.

Transform *relational data to graphs gaining flexibility of queries*
Unify *data from different sources through alignment to reality (what data is about)*
Research *is supported by tracking provenance of data processing and integration*
Biomedical Ontologies *provide computable hierarchies and connections.*

TURBO can find patients based on demographics (e.g., sex, BMI), diagnosis related to a disease class (e.g., lung cancer), medications prescribed (e.g., opioid orders), lab test results (e.g., blood glucose measurements), and representation in specific resources (e.g., Penn Medicine Biobank, the Penn Cancer Registry).

PennTURBO is comprised of three major components: Carnival, Knowledge graphs, and the Semantic Engine.

Carnival data integrator



TurboKG API

Query the Turbo Knowledgegraph

Turbo KnowledgeGraph

Static RDF triple store that contains connected knowledge about diseases, meds, labs, and assays. Does not contain individual level patient data.

Carnival

Harmonize data into property graph, Return graphs and reports

Semantic Engine

Transform simple instance data to OBO compliant RDF

TURBO Ontology

Cohort Graphs

Data Sources

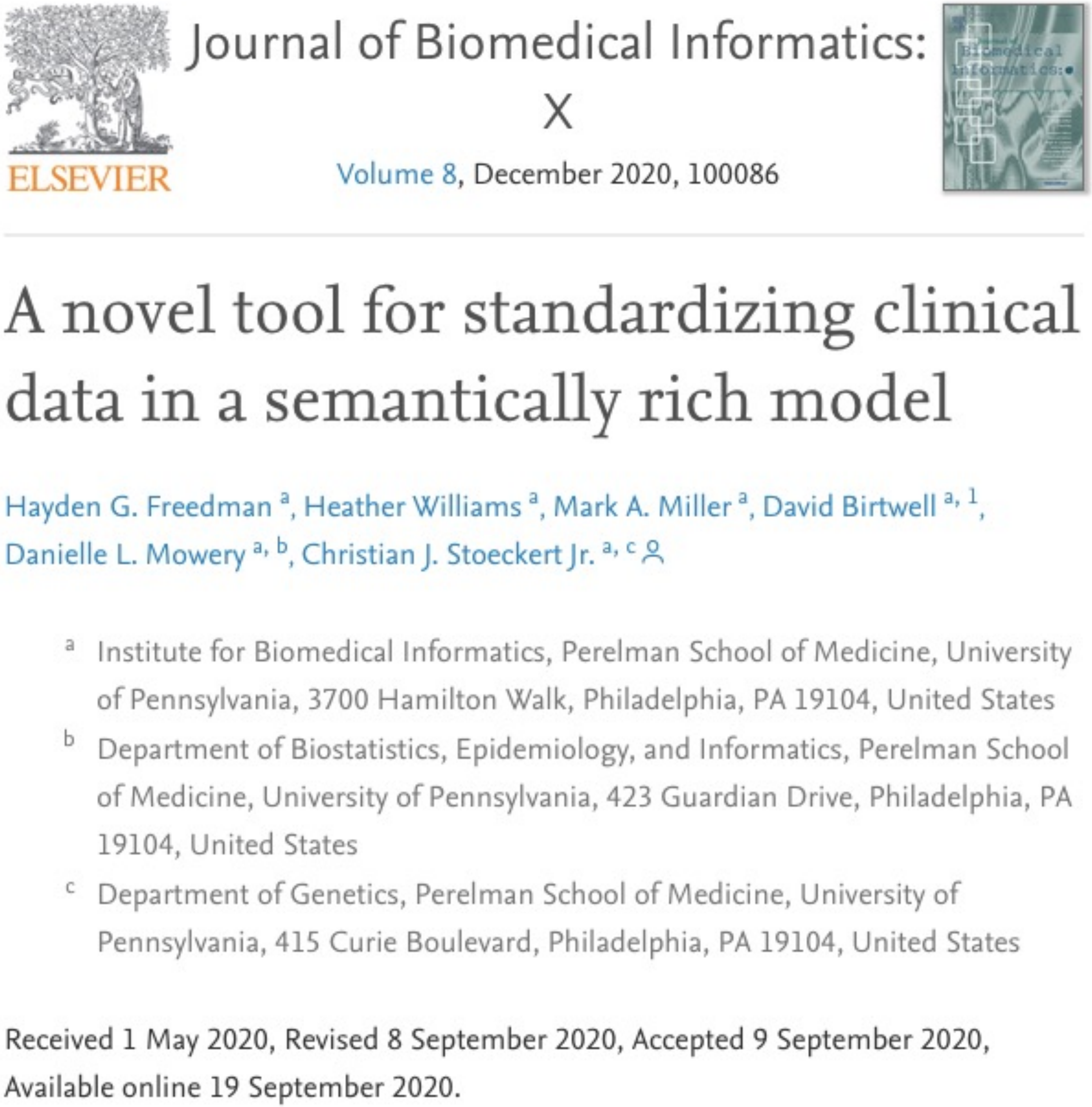
- RDMS
- Flat files
- *EPIC*
- *OMOP*
- *RedCap*

Generator of patient cohort graph database (Semantic Engine)

Knowledgegraphs linking ontologies to searchable fields in EHR

The Semantic Engine uses a graph specification for allowable triples based on the TURBO Ontology

- Application ontology that imports terms from > 20 OBOF ontologies.
- Also includes TURBO project terms
  - "TURBO assertion making process" : A planned process that takes a datum as input and has a rdf:Statement as output. (used to track various recodings)
- Originally built using Ontodog (started based on the Ontology for Biobanking - OBIB) and Ontofox (for imports).
- Made a release this year with the Ontology Development Kit (ODK).
  - https://github.com/PennTURBO/turbo-ontology/releases/tag/2022-05-09



ELSEVIER

Journal of Biomedical Informatics: X

Volume 8, December 2020, 100086

A novel tool for standardizing clinical data in a semantically rich model

Hayden G. Freedman [a], Heather Williams [a], Mark A. Miller [a], David Birtwell [a, 1], Danielle L. Mowery [a, b], Christian J. Stoeckert Jr. [a, c]

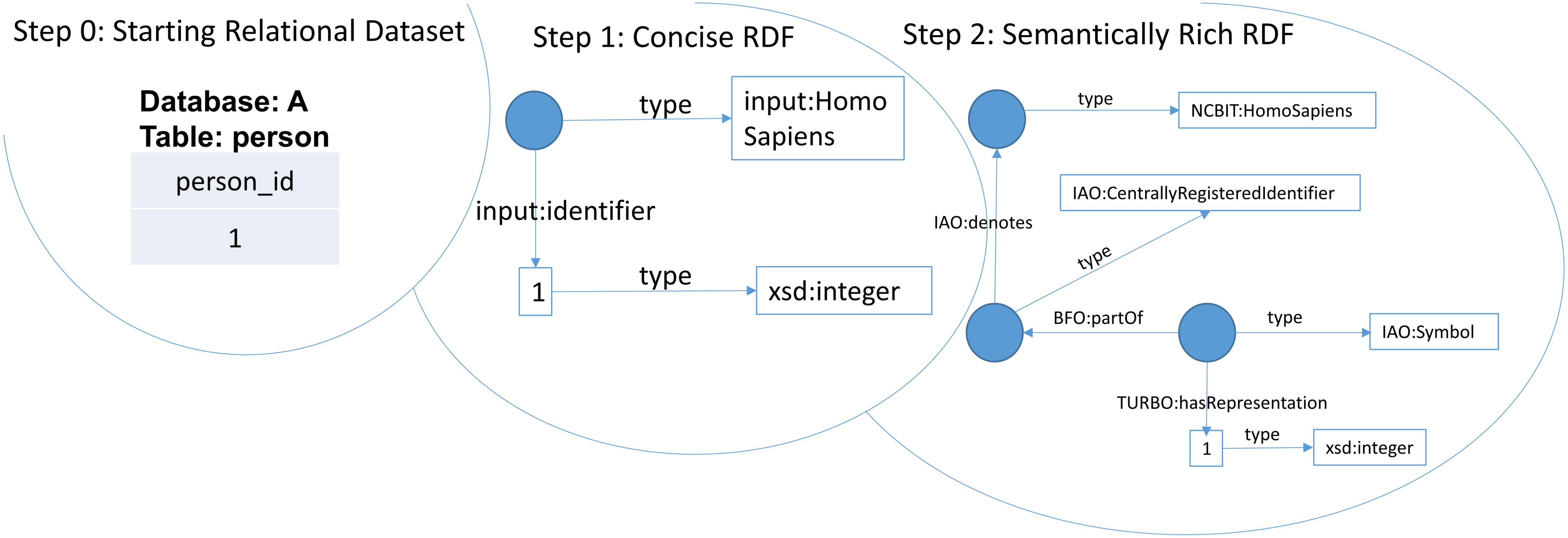
[a] Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, 3700 Hamilton Walk, Philadelphia, PA 19104, United States

[b] Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, 423 Guardian Drive, Philadelphia, PA 19104, United States

[c] Department of Genetics, Perelman School of Medicine, University of Pennsylvania, 415 Curie Boulevard, Philadelphia, PA 19104, United States

Received 1 May 2020, Revised 8 September 2020, Accepted 9 September 2020, Available online 19 September 2020.

Explicitly expressing the semantics that are implicit in the data



Step 0: Starting Relational Dataset

Database: A
Table: person

person_id
1

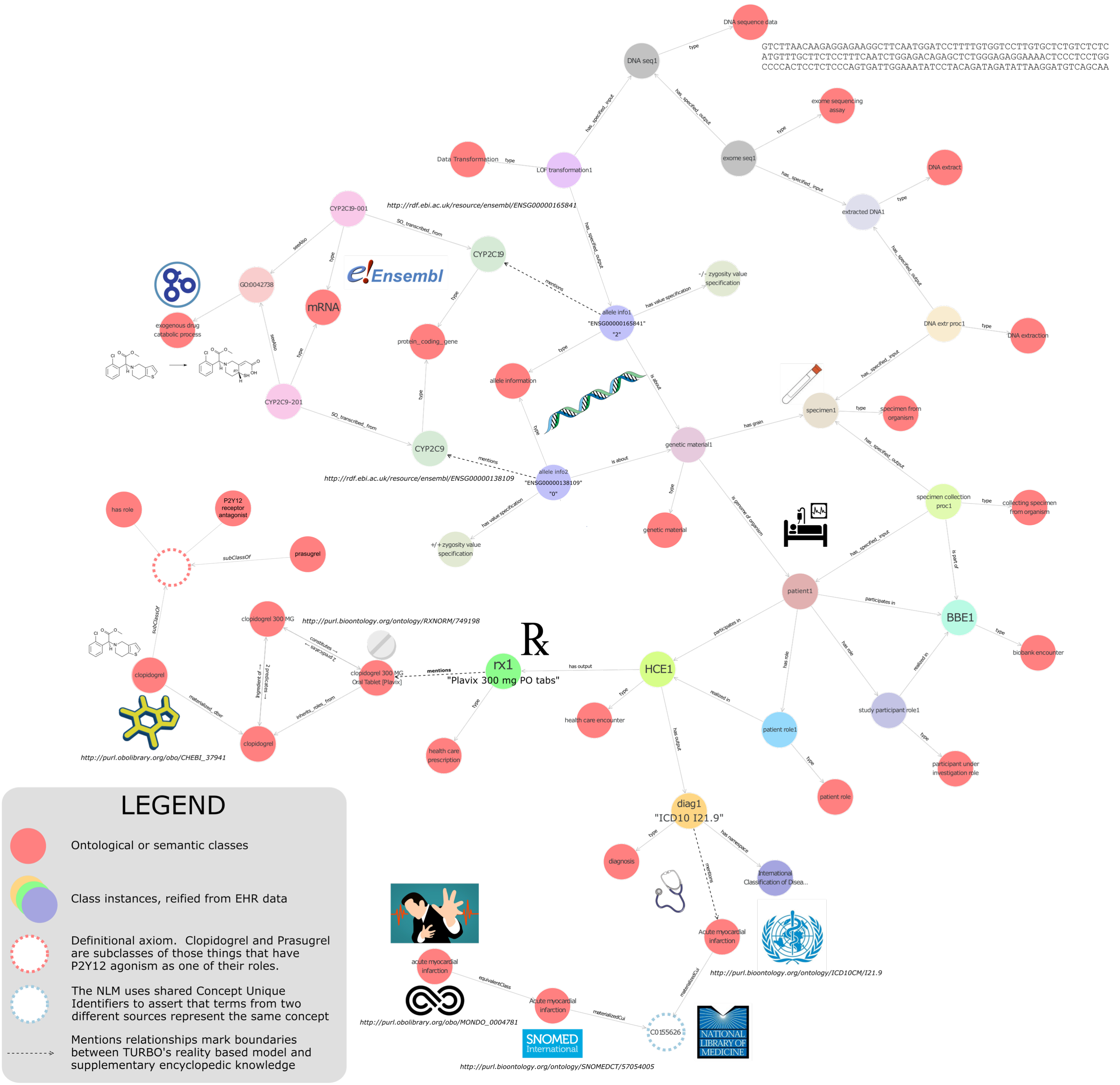
Step 1: Concise RDF

Step 2: Semantically Rich RDF

Differs from ontology in defining only a single way to relate instances

Visualization of a TURBO graph pattern with mock EHR data: CYP2C19*2/*2 MI patient taking Clopidogrel

Clinical picture and therapeutic options: Patient had a recent myocardial infarction. Clopidogrel was prescribed to inhibit platelet aggregation via P2Y$_{12}$ antagonism. It was later determined that the patient has the CYP2C19*2/*2 genotype and lacks the enzyme necessary to convert Clopidogrel into its active form. However, the patient is wildtype for CYP2C9, the gene whose product does the analogous conversion for Prasugrel, another P2Y$_{12}$ inhibitor.



LEGEND

● Ontological or semantic classes

● Class instances, reified from EHR data

○ Definitional axiom. Clopidogrel and Prasugrel are subclasses of those things that have P2Y12 agonism as one of their roles.

○ The NLM uses shared Concept Unique Identifiers to assert that terms from two different sources represent the same concept

- - - -> Mentions relationships mark boundaries between TURBO's reality based model and supplementary encyclopedic knowledge

Penn Medicine

Data fields currently included in the PennTURBO Group’s clinical data model

Category of Data	Fields Modeled
Patient demographics and observations	● Centrally Registered Identifier (CRID) ● Date of Birth ● Gender Identity ● Racial Identity ● Height ● Weight ● BMI ● Systolic Blood Pressure ● Diastolic Blood Pressure
Healthcare and Biobank Encounters	● Encounter Primary Key ● Date of Encounter
Diagnoses	● Diagnosis Primary Key ● Diagnosis Code ● Diagnosis Code Registry (ICD9, ICD10, SNOMED, etc.) ● Diagnosis Description String
Medications	● Medication Primary Key ● Medication Code ● Medication Code Registry (e.g., RxNorm) ● Medication Description String

Penn Medicine

Can it scale? Comparison of the time taken for the Semantic Engine to transform various types of data, by cohort size.

Cohort Size (Patients)	Patient Instantiation Time (seconds)	Encounter Instantiation Time (seconds)	Diagnosis Instantiation Time (seconds)	Medication Expansion Time (seconds)
1,000	3	47	6	21
10,000	29	444	62	238
100,000	255	3,004	542	1,757
1,000,000	2,937	37,724	8,734	24,573

Applying ontology-driven tools on real world data: Challenges to applying TURBO

- Using knowledge graphs to connect RWD to ontologies requires significant mapping.
- Need more developers comfortable with graphs, working with triples.
- Operationalizing in a SQL world, matching ontology expressiveness to the needs of the data.
- Recognition of benefit for data integration where individuals matter.

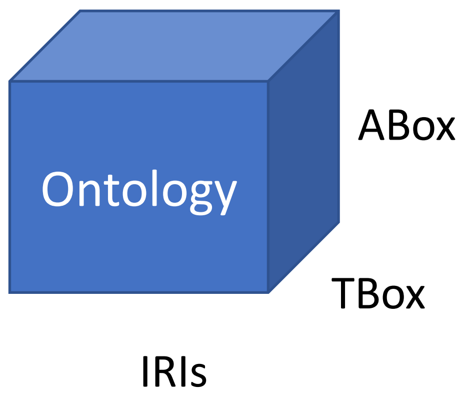
We are developing PennTURBO as generalizable open source projects using GitHub

- TURBO team: Heather Williams, Tom Hutchinson, Danielle Mowery, *Hayden Freedman (UCI) , Mark Miller (LBL)*, Chris Stoeckert, *David Birtwell (USC)*
- Carnival: https://github.com/carnival-data/carnival
  - Heather Williams, Tom Hutchinson, David Birtwell, Louis Lee
- Semantic Engine: https://github.com/PennTURBO/semantic-engine
  - Hayden Freedman and the TURBO team
- Knowledge Graphs
  - Tom Huthchinson, Mark Miller and the TURBO team
  - https://github.com/PennTURBO/medication-knowledgegraph-pipeline
    - text strings to RxNorm to drugs and their roles
  - https://github.com/PennTURBO/disease-diagnosis-knowledgegraph-pipeline
    - ICD codes to disease terms
- Currently working on a lab test knowledge graph pipeline.
- https://pennturbo.github.io/Turbo-Documentation/

Penn Medicine

Insights from a holistic approach in striving for semantic harmony of datasets, communities, and RWD

- Striving to achieve semantic harmony across datasets in VEuPathDB has identified issues to be addressed as a community of ontology developers
  - Patterns for ontology terms
  - Role of application ontologies
- OBI and OBOF are addressing issues of semantic harmony through encouraging community involvement
  - Providing tools (ROBOT, Dashboard), tutorials, and communication platforms but need more outreach.
  - Balancing inclusivity and quality
- Real-world data is about individuals (ABox).
  - Carnival enables the aggregation of data from disparate sources into a unified property graph and provides mechanisms to model and interact with the graph in well-defined ways inspired by OBO Foundry ontologies.
  - TURBO Semantic Engine generates RDF triples with OBOF semantics denoting patients and their health care encounters.



Acknowledgements

- VEuPathDB: National Institute of Allergy and Infectious Diseases, Wellcome Trust (UK)
- ClinEpiDB/ MicrobiomeDB: Bill & Melinda Gates Foundation
- OBI/ OBO Foundry: **Many communities who have contributed!**
  - NIH grant 1R24HG010032: Services to support the OBO Foundry standards (**Bjoern Peters, Chris Mungall**)
  - Volunteers are essential but central funding really accelerates progress!
- TURBO: Support from the Penn Institute for Biomedical Informatics, University of Pennsylvania Health System, and the Penn Institute for Translational Medicine and Therapeutics
- Many many colleagues through out the years!

Thank you!

Penn Medicine