A Review of Relevant Ontologies and Application of Reasoners

Melissa Haendel, PhD
@ontowonka
Outline

- Using ontologies and reasoners for classification
- Anatomy and Stage Ontologies
- Example of ontologies and reasoning at work: diagnosing diseases
- Environmental ontologies
- How to exchange data better
What is an Ontology?

Definition: A formal conceptualization of a specified domain

Key Features:
- Terms are defined
- Relationships between terms are defined, allowing logical inference and sophisticated data queries
- Terms are arranged in a hierarchy
- Expressed in a knowledge representation language such as RDFS, OBO, or OWL

Examples:
SNOMED, Foundational Model of Anatomy, Gene Ontology, Linnean Taxonomy of species
Example taxonomy

- cell (CL_0000000)
  - neural cell (CL_0002319)
    - neuron (CL_0005040)
      - enteric neuron (CL_0007011)
    - neuron-associated cell (CL_0000095)
      - efferent neuron (CL_000527)
  - glial cell (CL_0000125)
    - microglial cell (CL_0000129)
    - macroglial cell (CL_0000126)
      - oligodendrocyte (CL_0000128)
      - astrocyte (CL_0000127)
        - cerebellar astrocyte (CL_0002603)
        - hippocampal astrocyte (CL_0002604)
Ontologies enable queries to “just work” as you would hope.

Without ontological “subsumption” reasoning, synonym formalism, the user would either need to do 17 different queries, or get an incomplete set of results.
Ontologies support automated consistency checking, inferred classification along different axes, and powerful graph-based applications.
Ontologies are formal classifications

Appendage
  Tail
  Median fin
  Paired fin
  Pectoral fin
  Pelvic fin

appendage
  tail
  Paired fin
  Pectoral fin
  Pelvic fin

A
Relationships also support classification

‘pectoral fin radial’ SubClassOf part_of some ‘fin’
Necessary and sufficient conditions

Any sense organ that functions in the detection of smell is an olfactory sense organ.
These are necessary and sufficient conditions, also called an equivalent class axiom.
Using reasoners to detect errors

**UBERON: bone**

**Vertebrata**

**UBERON: tibia**

**Homo sapiens**

*Drosophila melanogaster*

part_of

is_a

Fruit fly FBbt ‘tibia’

Human FMA ‘tibia’
Using reasoners to detect errors

- **UBERON: bone** only_in_taxon **Vertebrata**
  - is_a
  - **Drosophila melanogaster** part_of
  - **UBERON: tibia**
  - is_a
  - **Homo sapiens** part_of

Fruit fly FBbt ‘tibia’

Human FMA ‘tibia’
Using reasoners to detect errors

Disjoint_with

**UBERON: bone**

only_in_taxon

**Vertebrata**

**Drosophila melanogaster**

part_of

**UBERON: tibia**

**Homo sapiens**

part_of

✗

**Fruit fly FBbt ‘tibia’**

**Human FMA ‘tibia’**
A compendium of interoperable ontologies

- Functional Genomics: Gene *function*
- Transcriptomics, proteomics: Gene *expression*
- Phenomics and assays: Effects of gene *mutations and environment and their measurement*
- Environments: drugs, exposures, life history
- Disease: Effects of gene *mutations + phenotypes environment + staging*
Anatomy and stage ontologies
The Zebrafish Anatomy and stage ontologies

Diagram: The diagram illustrates the development stages and ontologies of the zebrafish heart. It shows the progression from heart primordium to heart, highlighting stages such as 5-9 somite, 20-25 somite, 26+ somite, prim-5, prim-25, high-pec, and adult. The development stages are connected through arrows indicating the sequence of events.

Key terms: part_of, is_a, start_stage, end_stage, develops_from.

Cardiovascular system: This is a central component of the diagram, with connections to various stages and organs.

Cavitated compound organ: Another key component, with connections to the cardiovascular system and germ layers.

Developmental Time: The timeline at the bottom of the diagram indicates the progression through stages.
The Zebrafish Anatomy and stage ontologies

A. 
- Neural Plate
- Neural Keel
- Neural Rod
- Neural Tube

B. 
- Neural rod
- Neural keel
- Neural plate

11 hpf

Time (hours post fertilization)
Uberon: bridging semantics for anatomy


A merger of disease ontologies
The challenge of multiple perspectives: how can we bridge these?

Disease classifications and lists...there are a lot of them.
4 disease resources plus mappings: Hemolytic anemia

ORDO/Orphanet (yellow)

OMIM (brown)

SubClassOf (solid line)

MESH (grey)

Xref (dashed grey line)

Mungall

Harmonizing disease vocabularies: http://bit.ly/Monarch-Disease
Hemolytic Uremic Syndrome, Atypical, Susceptibility To, 1
Hemolytic Uremic Syndrome, Atypical, Susceptibility To, 2
Hemolytic Uremic Syndrome, Atypical, Susceptibility To, 3
Hemolytic Uremic Syndrome, Atypical, Susceptibility To, 4
Hemolytic Uremic Syndrome, Atypical, Susceptibility To, 5
Hemolytic Uremic Syndrome, Atypical, Susceptibility To, 6

Typical hemolytic-uremic syndrome

Kidney disease

Nephrotic Syndrome, Type 7

Complement Factor H Deficiency

Atypical hemolytic-uremic syndrome with H factor anomaly
Atypical hemolytic-uremic syndrome with anti-factor H antibodies
Atypical hemolytic-uremic syndrome with I factor anomaly
Atypical hemolytic-uremic syndrome with B factor anomaly
Atypical hemolytic-uremic syndrome with MCP/CD46 anomaly
Atypical hemolytic-uremic syndrome with C3 complement deficiency
Atypical hemolytic-uremic syndrome with DGKE deficiency
BOOM Bayes OWL Ontology Merging: Finds the set of hypothetical axioms that maximises $P(O^p)$
## MonDO: Merged Ontology of Disease Entities

<table>
<thead>
<tr>
<th>“Ontology”</th>
<th>Classes (before, after merge)</th>
<th>SubClass axioms</th>
<th>Mappings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inputs:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOID</td>
<td>6878 → 6012</td>
<td>7082</td>
<td>36656</td>
</tr>
<tr>
<td>MESH (D)</td>
<td>11314 → 4152</td>
<td>19036</td>
<td></td>
</tr>
<tr>
<td>OMIM (D)</td>
<td>7783 → 7783</td>
<td>0</td>
<td>31242</td>
</tr>
<tr>
<td>Orphanet (D)</td>
<td>8740 → 4683</td>
<td>15182</td>
<td>20326</td>
</tr>
<tr>
<td>OMIA</td>
<td>4833 → 4833</td>
<td>3120</td>
<td>355</td>
</tr>
<tr>
<td>DC</td>
<td>209 → 208</td>
<td>310</td>
<td>316</td>
</tr>
<tr>
<td>Medic</td>
<td>0</td>
<td>8630</td>
<td>3435</td>
</tr>
<tr>
<td><strong>Output:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MonDO</td>
<td>39757 → 27617</td>
<td>44837</td>
<td></td>
</tr>
</tbody>
</table>

Phenotype ontologies
Different communities use different languages

- Palmoplantar hyperkeratosis
- Thick hand skin
- Ulcerated paws
- Degenerate fin epithelium
Challenge: Each data source uses their own vocabulary/ontology
Challenge: Each data source uses their own phenotype vocabulary/ontology
Decomposition of complex concepts allows interoperability

Human phenotype

“Palmoplantar hyperkeratosis”

= 

PATO

increased

Uberon

Stratum corneum layer of skin

GO

keratinization

Species neutral ontologies, homologous concepts
Semantic similarity of phenotypes for disease discovery

FMA+PATO  MP  ZFA+PATO  FBbt+PATO

The Human Phenotype Ontology for deep phenotyping

- Hyposmia
- Abnormal eye morphology
- Abnormality of globe location
- Deeply set eyes
- Motor neuron atrophy

- sensory perception of smell
- eyeball of camera-type eye
- motor neuron

34571 annotations in 22 species
157534 phenotype annotations
2150 phenotype annotations
Ontologies at work: Data integration and disease diagnosis
### A: Data types covered by Monarch data sources

<table>
<thead>
<tr>
<th>Data source in Monarch</th>
<th>G</th>
<th>2</th>
<th>P/D</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClinVar</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>CTD</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Gene Revs</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>OMIM db</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>HPOA</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Orphanet</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>GWAS</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Coriell</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>KEGG</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>OMIA</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>AnimalQTLDB</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>MPD</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>MMRRC</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>WmBase</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>FlyBase</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>IMPC</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>MGI</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>ZFIN</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Ensembl</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>NCBI</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>UCSC</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>HGNC</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>BioGrid</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Panther</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
</tbody>
</table>

### B: Monarch data sources and ontology annotations

- **Annotations**:
  - *Annotated to many sources*
  - *Annotated to all sources*

- **Bridging Ontologies**:
  - UMLS
  - MONDO
  - Elements of Morphology
  - UPheno
  - UBERON
  - GO
  - ECO
  - RO

### C: Mappings to bridging ontologies

- **ONTologies**:
  - MedGen
  - MeSH
  - DOID
  - OMIM
  - ORDO
  - HP
  - EFO
  - VT
  - WT
  - FBcv
  - WA
  - MA
  - FBbt
  - EMAPA
  - ZP
  - ZFA
  - SO
  - FALDO

- **Created**: Maintained

- **Contributed**: Contributed

- **Annotations**: Annotated to many sources

- **Annotations**: Annotated to all sources
Harmonizing diseases, phenotypes, anatomy, and genotypes

91% of our 2.2 Million G2P associations require integrating 2 or more data sources
Phenotypic matchmaking for disease diagnostics

Gene Profile

Gene D
One or more mutations known to cause "Disease D"

Disease D
Hypoplasia of the frontal lobes
Distal lower limb atrophy
Optic nerve hypoplasia
Lyssencephaly
Contractures of the large joints
Focal seizures
Decreased body weight

Patient A Phenotype Profile

Patient A
Aplasia/ Hypoplasia of the cerebrum
Muscular atrophy
Abnormal optic nerve morphology
Abnormal cortical gyration
Flexion contracture
Seizures
Decreased body weight

Closest term in common
Hyperthermia

Gene M or Genotype M
Mouse M

Homology

Orthology

HP Terms
Bridging ontology term
MP Terms
Combining genotype and phenotype data for variant prioritization

Whole exome

Remove off-target and common variants

Mendelian filters

Variant score from allele freq and pathogenicity

Phenotype score from phenotypic similarity

PHIVE score to give final candidates

Variant Score based on allele frequency and pathological impact

Phenotypic Similarity Score based on similarity of observed phenotypes

Putting all that data to use to diagnose a rare platelet syndrome


Ranked STIM-1 variant maximally pathogenic based on cross-species G2P data, in the absence of traditional data sources


Phenotypic profile

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>STIM-1</td>
<td>Heterozygous, missense mutation</td>
</tr>
</tbody>
</table>

Stim1Sax/Sax

N/A

NIH Undiagnosed Diseases Program

MG1

QMIM

Gene

<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>STIM-1</td>
<td>Heterozygous, missense mutation</td>
</tr>
</tbody>
</table>

N/A
What about environment and exposure ontologies?
“the environment is everything that isn’t me”

—Albert Einstein
Can we sensibly make an ontology of everything that isn’t me?
1. Occupational diseases caused by exposure to agents arising from work activities

1.1. Diseases caused by chemical agents

1.1.1. Diseases caused by beryllium or its compounds
1.1.2. Diseases caused by cadmium or its compounds
1.1.3. Diseases caused by phosphorus or its compounds
1.1.4. Diseases caused by chromium or its compounds
1.1.5. Diseases caused by manganese or its compounds
1.1.6. Diseases caused by arsenic or its compounds
1.1.7. Diseases caused by mercury or its compounds
1.1.8. Diseases caused by lead or its compounds
1.1.9. Diseases caused by fluorine or its compounds
1.1.10. Diseases caused by carbon disulfide
1.1.11. Diseases caused by halogen derivatives of aliphatic or aromatic hydrocarbons
1.1.12. Diseases caused by benzene or its homologues
1.1.13. Diseases caused by nitro- and amino-derivatives of benzene or its homologues
1.1.14. Diseases caused by nitroglycerine or other nitric acid esters
1.1.15. Diseases caused by alcohols, glycols or ketones
1.1.16. Diseases caused by asphyxiants like carbon monoxide, hydrogen sulfide, hydrogen cyanide or its derivatives
1.1.17. Diseases caused by acrylonitrile
1.1.18. Diseases caused by oxides of nitrogen
1.1.19. Diseases caused by vanadium or its compounds
1.1.20. Diseases caused by antimony or its compounds
Can we make these lists computable?

Translate them into a form a machine can understand and reason over?
We have a precise machine-readable language for describing some environmental exposures

\[
\text{CHEBI:6651}
\]

CheBI is a chemical ontology
But others are harder to define
The Zebrafish Environmental Conditions Ontology

https://github.com/ybradford/zebrafish-experimental-conditions-ontology
The Environment Ontology

- Originally created for metagenome samples
  - Characterize microbial environments
- Extended for ecological science
  - The “Earth Phenotype Ontology”
- Being adapted for human exposures

Material
- Water
- Soil
- Air

Features
- Natural
- Anthropogenic

Biome
- Terrestrial
- Aquatic
- Polar

Process
- Erosion
- Pollution
- Biological
  - Algal bloom…

Biome: Food desert
Feature: Store (alcohol, sugar-rich food)
Material: Air, high particulate matter
Process: decreased investment in infrastructure

Image: Zol87 CC by/nc
C HeBl: chemical classification

- Organic aromatic compound
- Carboxyclic compound
- Organooxygen compound
- Carbonyl compound
- Esters
- Naphthenic aromatic compound
- Aromatic ester
- Carboxylic ester
- Diester
- Phthalate ester
- Disobutyl phthalate
- PPAR modulator
- Teratogenic agent
- Biological role
- Aetiological role
- Application

monarchinitiative.org
Environmental conditions, treatments and exposures ontology (ECTO)

- PECO: Pombe experimental conditions ontology
- ZECO: Zebrafish environmental conditions
- ExO: Exposure ontology
- NCI Thesaurus clinical
- SNOMED clinical
- XCO: Experimental conditions ontology
- MRE: Medically relevant exposures
- ENVO

https://github.com/cmungall/environmental-conditions
monarchinitiative.org
The Ontology of Biomedical Investigations

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0154556
Recording and exchanging phenotype and environmental data...better
WebPhenote and Noctua

A causal/spatiotemporal network curation environment

Form-based

Graph-based

http://create.monarchinitiative.org/

noctua.berkeleybop.org
Computable encodings are essential

Genes + Environment = Phenotypes

Base pairs
Variant notation (e.g. HGVS)

Medical procedure coding

Human Phenotype Ontology

ICD-10

SNOMED

PheKB
Standard exchange formats exist for genes … but for phenotypes? Environment?
If it is alive, it can be PhenoPackaged

Patients & Cohorts

Rare Disease Diagnosis

Model Organisms

Personalized Medicine

Disease vectors

Epidemiological Monitoring

Mechanistic Discovery

Drug discovery & Development

Biodiversity

Epidemiological Monitoring

Crops

Genetic Engineering

Domestic Animals

Some biodiversity images adapted from http://i.vimeocdn.com/video/417366050_1280x720.jpg
A semantic vision for environmental health research

Laying a Community-Based Foundation for Data-Driven Semantic Standards in Environmental Health Sciences

https://ehp.niehs.nih.gov/15-10438/
For updates on the SEAZIT project and other activities related to *in vitro* alternatives, subscribe to the NICEATM News email list.


- Check the NICEATM News box and click submit.
Acknowledgements

OHSU
Matt Brush
Kent Shefchek
Julie McMurry
Tom Conlin
Nicole Vasilevsky
Dan Keith

Genomics
England/Queen Mary
Damian Smedley
Jules Jacobson

Jackson Laboratory
Peter Robinson

ZFIN
Ceri VanSlyke
Yvonne Bradford

CTD
Carolyn Mattingly

Garvan
Tudor Groza

Max Plank
Pier Buttigieg

Lawrence Berkeley
Chris Mungall
Suzanna Lewis
Jeremy Nguyen
Seth Carbon
Nicole Washington

Charité
Sebastian Kohler

RTI
Jim Balhoff
Becky Boyles

Cyverse
Ramona Walls

FUNDING: NIH Office of Director: 2R24OD011883; NIH-UDP: HHSN268201300036C, HHSN268201400093P, NSF-DEB-0956049, NCINCI/Leidos #15X143, BD2K U54HG007990-S2 (Haussler) & BD2K PA-15-144-U01 (Kesselman)

With special thanks to Julie McMurry for excellent graphic design
The Monarch Initiative

PDs: Melissa Haendel, Chris Mungall, Peter Robinson

Funding:
NIH Office of Director: 1R24OD011883; NIH-UDP: HHSN268201300036C, HHSN268201400093P; NCINCI/Leidos #15X143, BD2K U54HG007990-S2 (Haussler) & BD2K PA-15-144-U01 (Kesselman)