A Review of Relevant Ontologies and Application of Reasoners

Melissa Haendel, PhD
@ontowanaka
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**
  - **neural cell**
    - **neuron**
    - **neuron-associated cell**
      - **enteric neuron**
      - **afferent neuron**
      - **glial cell**

- **microglial cell**
- **macrogliar cell**

- **oligodendrocyte**
- **astrocyte**

- **cerebellar astrocyte**
- **hippocampal astrocyte**
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

A
Relationships also support classification

‘pectoral fin radial’ \( \text{SubClassOf} \) part_of some ‘fin’
Necessary and sufficient conditions

Any sense organ that functions in the detection of smell is an olfactory sense organ.
Classifying

These are necessary and sufficient conditions, also called an equivalent class axiom
Using reasoners to detect errors

UBERON: bone

Vertebrata

is_a

UBERON: tibia

Homo sapiens

is_a

Drosophila melanogaster

part_of

Fruit fly FBbt ‘tibia’

is_a

part_of

Human FMA ‘tibia’
Using reasoners to detect errors

Drosophila melanogaster

Vertebrata

UBERON: bone

UBERON: tibia

Homo sapiens

Fruit fly FBbt ‘tibia’

Human FMA ‘tibia’
Using reasoners to detect errors

- UBERON: bone
- only_in_taxon
- Vertebrata

- disjoint_with
- is_a

- Drosophila melanogaster
- UBERON: tibia
- Homo sapiens

- part_of
- is_a

- 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

Gene Ontology

Anatomy and Stage Ontologies

Phenotype and Trait Ontology, Ontology of Biomedical Investigations

ENVO, MRE, ZECO, ECTO

Numerous nosologies, MonDO
Anatomy and stage ontologies
The Zebrafish Anatomy and stage ontologies

- part_of
- is_a
- develops_from

- cardiovascular system
- cavitated compound organ

- heart primordium
- heart rudiment
- primitive heart tube
- heart tube
- heart

Developmental Time:
- 5-9 somite
- 20-25 somite
- 26+ somite
- prim-5
- prim-25
- high-pec
- adult
The Zebrafish Anatomy and stage ontologies

A.
- Neural Plate
- Neural Keel
- Neural Rod

B.
- Neural rod
- Neural keel
- Neural plate

11 hpf

Time (hours post fertilization): 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30
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**

**DoID** (blue)

**OMIM** (brown)

**SubClassOf** (solid line)

**MESH** (grey)

**Xref** (dashed grey line)

Mungall
Harmonizing disease vocabularies: http://bit.ly/Monarch-Disease
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>&quot;Ontology&quot;</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>

Different communities use different languages

- Palmoplantar hyperkeratosis
- Degenerate fin epithelium
- Thick hand skin
- Ulcerated paws
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

“Palmoplantar hyperkeratosis”

Species neutral ontologies, homologous concepts

Human phenotype

PATO

increased

GO

keratinization

Uberon

Stratum corneum layer of skin

Autopod
Semantic similarity of phenotypes for disease discovery

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

- Genes
- Variants
- Orthology
- Animals
- Function
- Interactions
- Expression
- Pathways
- Diseases
- Phenotypes

B: Monarch data sources and ontology annotations

C: Mappings to bridging ontologies

- Annotated to many sources
- Annotated to all sources

- UMLs
- MedGen
- MeSH
- DOID
- OMIM
- Mondo
- ORDO
- Elements of Morphology
- HP
- EFO
- VT
- UPheno
- UBERON
- Geno
- GO
- ECO
- RO
- Created
- Maintained
- Contributed
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"

Patient A Phenotype Profile

Closest term in common

HP Terms | Bridging ontology term | MP Terms
--- | --- | ---
Hypoplasia of the frontal lobes | Aplasia/Hypoplasia of the cerebrum |  
Distal lower limb atrophy | Muscular atrophy |  
Optic nerve hypoplasia | Generalized atrophy |  
Lyssencephaly | Generalized atrophy |  
Contractures of the large joints | Optic atrophy |  
Focal seizures | Polymicrogyria |  
Decreased body weight | Abnormal joint mobility |  

Gene M or Genotype M

Mouse M
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

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


Genes

- Heterozygous, missense mutation STIM-1

Phenotypic profile

- Stim1Sax/Sax

N/A

NIH Undiagnosed Diseases Program

MGI Mouse

Stim1Sax/Sax

QuIMIM

CAGGC TGAGG

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?

<table>
<thead>
<tr>
<th>3.</th>
<th>Occupational cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.</td>
<td>Cancer caused by the following agents</td>
</tr>
<tr>
<td>3.1.1.</td>
<td>Asbestos</td>
</tr>
<tr>
<td>3.1.2.</td>
<td>Benzidine</td>
</tr>
<tr>
<td>3.1.3.</td>
<td>Bis-chloride</td>
</tr>
<tr>
<td>3.1.4.</td>
<td>Chromium</td>
</tr>
<tr>
<td>3.1.5.</td>
<td>Coal tar</td>
</tr>
<tr>
<td>3.1.6.</td>
<td>Beta-nafta</td>
</tr>
<tr>
<td>3.1.7.</td>
<td>Vinyl chloride</td>
</tr>
<tr>
<td>3.1.8.</td>
<td>Benzene</td>
</tr>
<tr>
<td>3.1.9.</td>
<td>Toxic naphtha</td>
</tr>
<tr>
<td>3.1.10.</td>
<td>Ionizing</td>
</tr>
<tr>
<td>3.1.11.</td>
<td>Tar, pitch</td>
</tr>
<tr>
<td>3.1.12.</td>
<td>Coke oven</td>
</tr>
<tr>
<td>3.1.13.</td>
<td>Nickel</td>
</tr>
<tr>
<td>3.1.14.</td>
<td>Wood odors</td>
</tr>
<tr>
<td>3.1.15.</td>
<td>Arsenic</td>
</tr>
<tr>
<td>3.1.16.</td>
<td>Beryllium</td>
</tr>
<tr>
<td>3.1.17.</td>
<td>Cadmium</td>
</tr>
<tr>
<td>3.1.18.</td>
<td>Erionite</td>
</tr>
<tr>
<td>3.1.19.</td>
<td>Ethylene oxide</td>
</tr>
<tr>
<td>3.1.20.</td>
<td>Hepatitis B virus (HBV) and hepatitis C virus (HCV)</td>
</tr>
</tbody>
</table>
| 3.1.21. | Cancers caused by other agents at work not mentioned in the preceding items where a classification has been established by administrative edition.
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

**Biome:** Food desert

**Feature:** Store (alcohol, sugar-rich food)

**Material:** Air, high particulate matter

**Process:** decreased investment in infrastructure
CHEBI: chemical classification

- organic aromatic compound
- carbocyclic compound
- organooxygen compound
- role
- biological role
- application
- aetiopathogenetic role
- plasticiser
- diester
- carboxylic ester
- aromatic ester
- benzenoid aromatic compound
- benzenes
- carbonyl compound
- ester
- disobutyl phthalate
- PPAR modulator
- teratogenic agent
- has role
- has role
- has role

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

PECO: Pombe experimental conditions ontology

ZECO: Zebrafish environmental conditions

ExO: Exposure ontology

XCO: Experimental conditions ontology

NCI Thesaurus clinical

SNOMED clinical

MRE: Medically relevant exposures

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

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

HGVs
SNOMED®
PheKB
Standard exchange formats exist for genes … but for phenotypes? Environment?

Genes
Environment
Phenotypes

GFF  VCF  BED

PXF

NEW

Global Alliance for Genomics & Health
If it is alive, it can be PhenoPackaged

Patients & Cohorts

Rare Disease Diagnosis

Model Organisms

Epidemiological Monitoring

Mechanistic Discovery

Drug discovery & Development

Disease vectors

Personalized Medicine

Biodiversity

Crops

Genetic Engineering

Domestic Animals

Epidemiological Monitoring

Environmental Monitoring

Genetic Engineering

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: HHSN2682013000036C, HHSN2682014000093P; NCINCI/Leidos #15X143, BD2K U54HG007990-S2 (Haussler) & BD2K PA-15-144-U01 (Kesselman)