Ontologies and controlled vocabularies

Why use ontologies and CVs?

- Very important in all data collection and analysis to manage and share large data sets
- To use the same data labels universally
- To enable quick retrieval of data
- To enable easy comparison of data
- To remove ambiguities

Ambiguities in names

- What is a cell?

Ambiguities in names

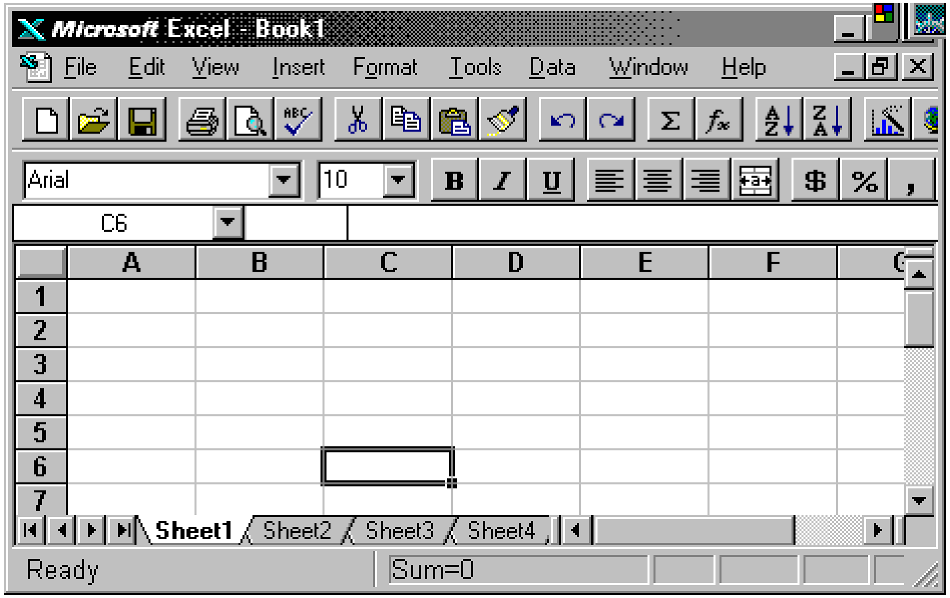
- What is a cell?

OR

Ambiguities in names

- What is a cell?

OR

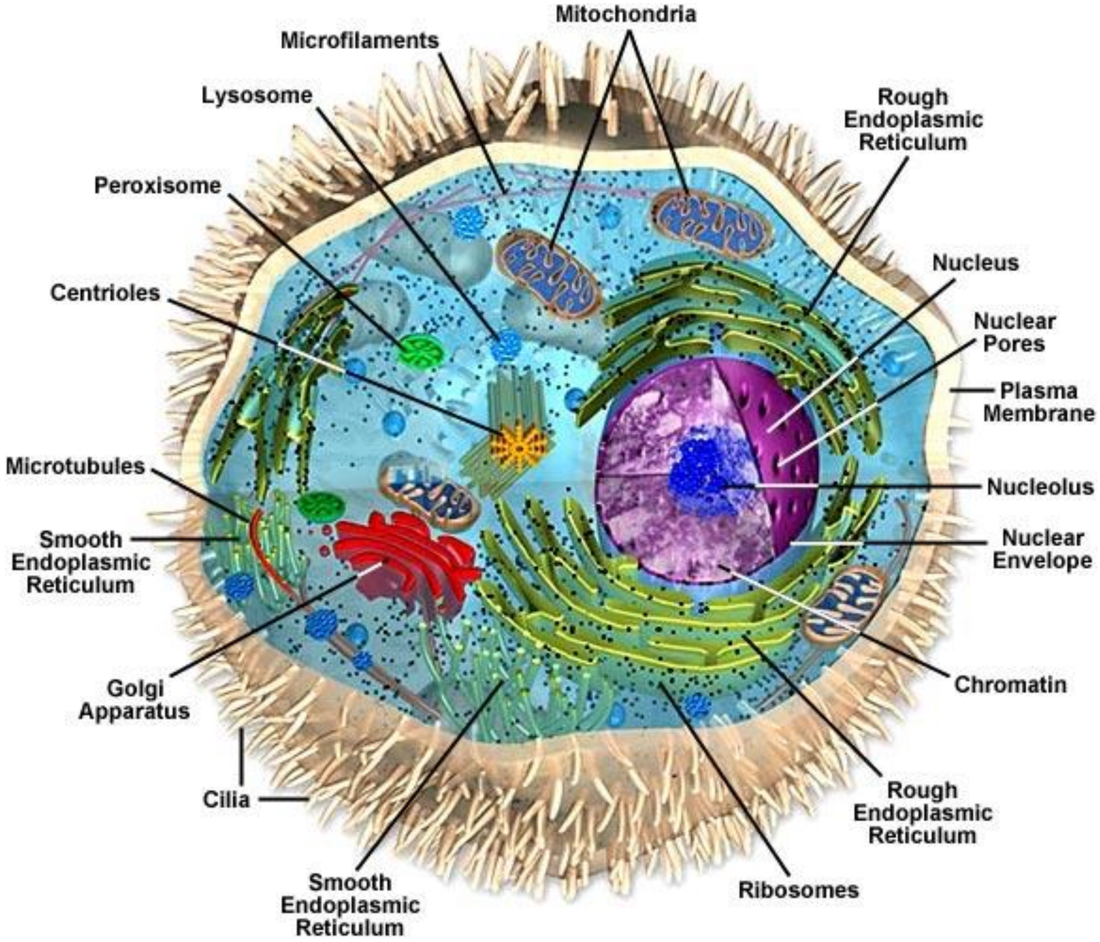


Computational Biology @ UCT

UNIVERSITY OF CAPE TOWN

Ambiguities in names

- What is a cell?

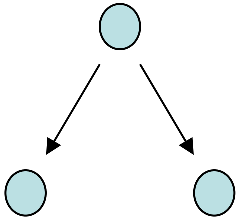


Ambiguities in naming continued

- The same **name** can be used to describe different **concepts**, e.g:
  - Glucose synthesis
  - Glucose biosynthesis
  - Glucose formation
  - Glucose anabolism
  - Gluconeogenesis
- All refer to the process of making glucose
- Makes it difficult to compare the information
- Solution: use **Ontologies** and **Data Standards**

Ontologies

- An ontology is a formal specification of terms and relationships between them – widely used in biology and boinformatics (e.g. taxonomy)
- The relationships are important and represented as graphs
- Ontology terms should have definitions
- Ontologies are machine-readable
- They are needed for ordering and comparing large data sets



Open Biomedical Ontologies

http://www.obofoundry.org/

- Central location for accessing well-structured controlled vocabularies and ontologies for use in the biological and medical sciences.
- Provides simple format for ontologies that can encode terms, relationships between terms and definitions of terms including those taken from external ontologies.

Scope of Open Biomedical Ontologies

- Anatomy
- Animal natural history and life history
- Chemical
- Development
- Ethology
- Evidence codes
- Experimental conditions
- Genomic and proteomic
- Metabolomics
- OBO relationship types
- Phenotype
- Taxonomic classification

Ontologies of use to H3Africa

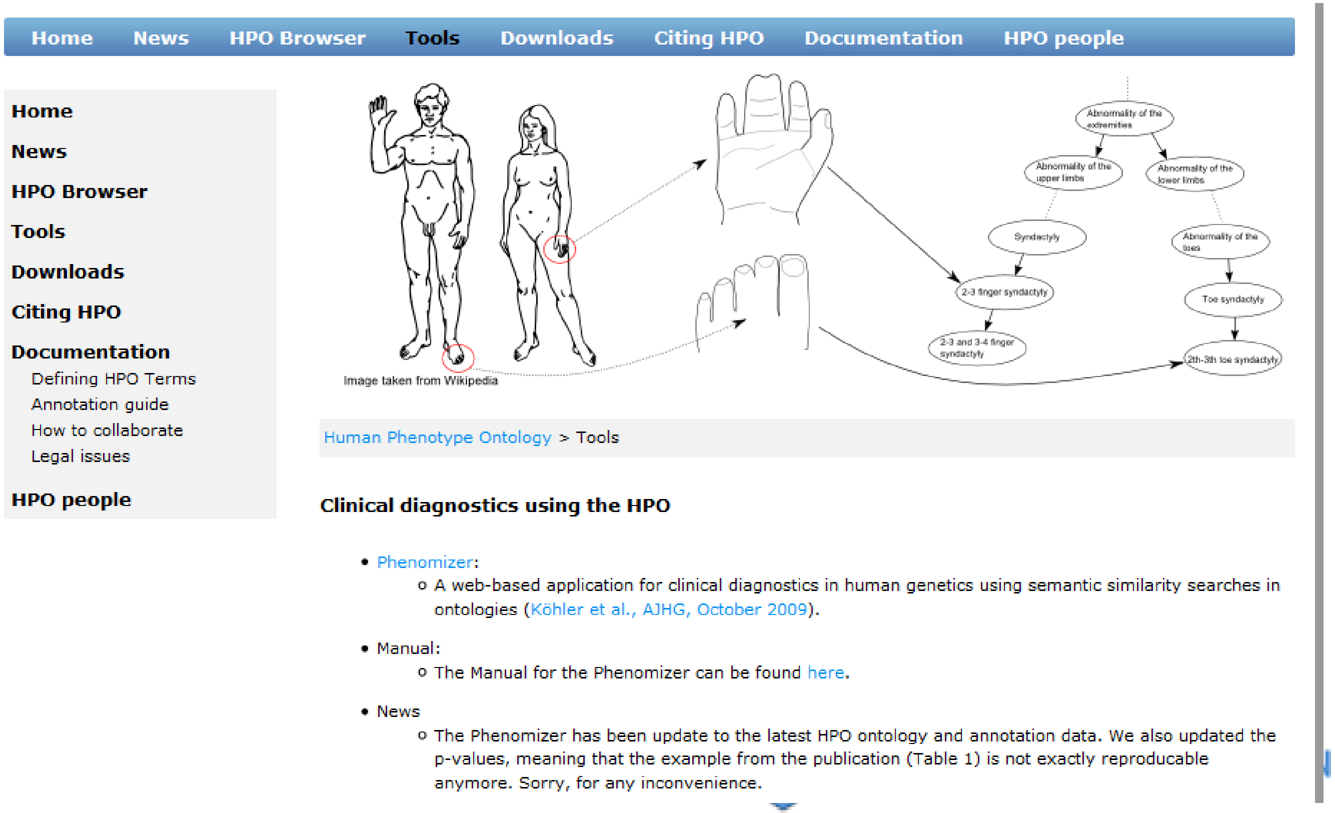
- Phenotypes
  - Mammalian phenotype ontology
  - Human phenotype ontology
  - PhenoDB, PhenoTips and PhenomeCentral
  - PhenX Toolkit
  - Symptom ontology
  - Disease Ontology
  - OMIM, SNOMED
- Experiment
  - Experimental factor ontology

Phenotypes

- …"observable morphological, physiological and behavioural characteristics of an individual in the context of the environment" (MP Ontology)
- Physicians, researchers and clinical laboratories need systems to enable standardized and structured phenotypic data to be collected from patients.

Phenotype ontology

- http://www.human-phenotype-ontology.org/
- Developed starting from OMIM + literature
- Has 10,000 terms and 50,000 annotations to hereditary di



- Home
- News
- HPO Browser
- Tools
- Downloads
- Citing HPO
- Documentation
  - Defining HPO Terms
  - Annotation guide
  - How to collaborate
  - Legal issues
- HPO people

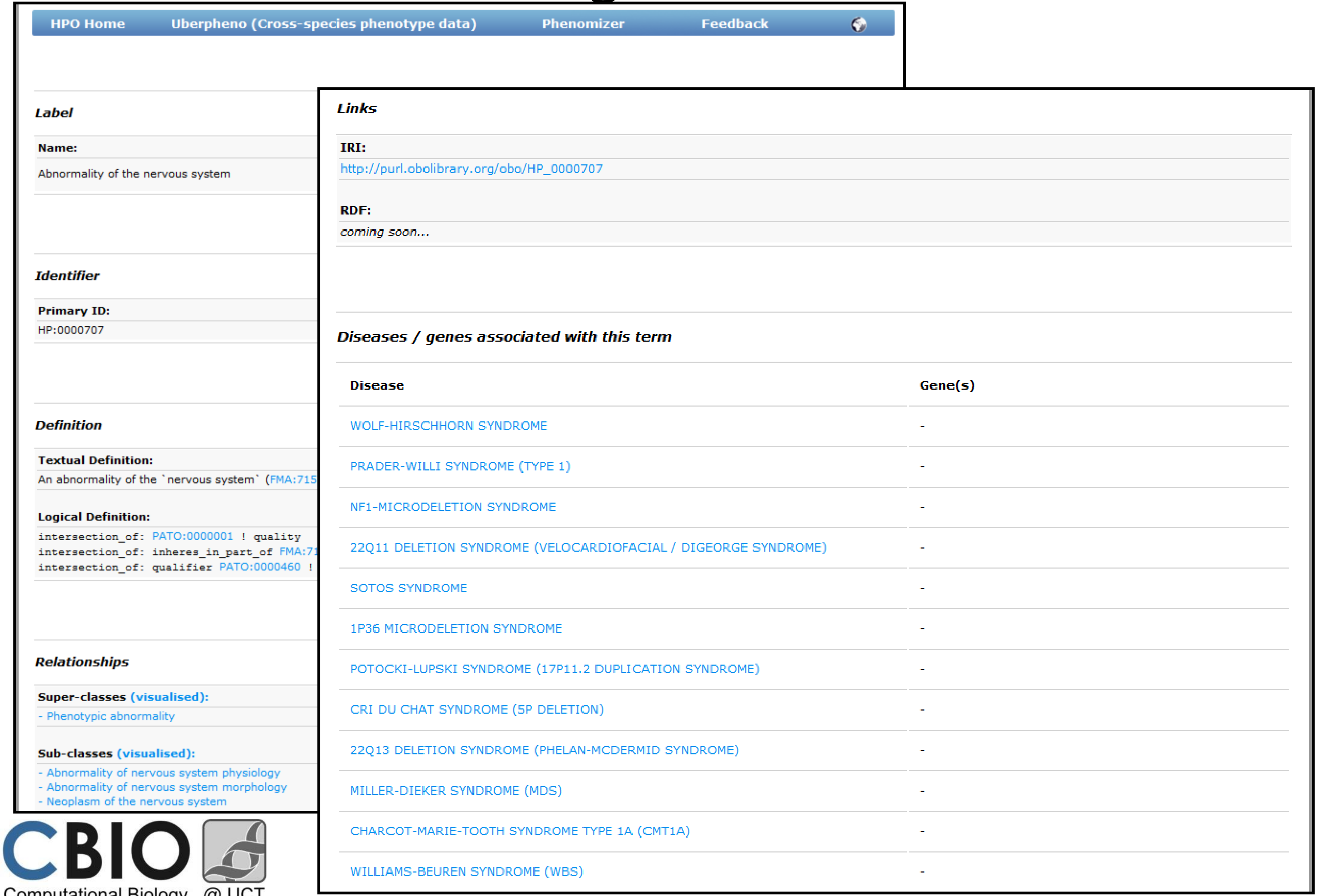
Human Phenotype Ontology > Tools

Clinical diagnostics using the HPO

- Phenomizer:
  - A web-based application for clinical diagnostics in human genetics using semantic similarity searches in ontologies (Köhler et al., AJHG, October 2009).
- Manual:
  - The Manual for the Phenomizer can be found here.
- News
  - The Phenomizer has been update to the latest HPO ontology and annotation data. We also updated the p-values, meaning that the example from the publication (Table 1) is not exactly reproducable anymore. Sorry, for any inconvenience.

CBIO Computational Biology @ UCT

Browsing the HPO



Label

Name:

Abnormality of the nervous system

Identifier

Primary ID:

HP:0000707

Definition

Textual Definition:

An abnormality of the `nervous system` (FMA:715

Logical Definition:

```
intersection_of: PATO:0000001 ! quality
intersection_of: inheres_in_part_of FMA:71
intersection_of: qualifier PATO:0000460 !
```

Relationships

Super-classes (visualised):

- Phenotypic abnormality

Sub-classes (visualised):

- Abnormality of nervous system physiology
- Abnormality of nervous system morphology
- Neoplasm of the nervous system

Links

IRI:

http://purl.obolibrary.org/obo/HP_0000707

RDF:

coming soon...

Diseases / genes associated with this term

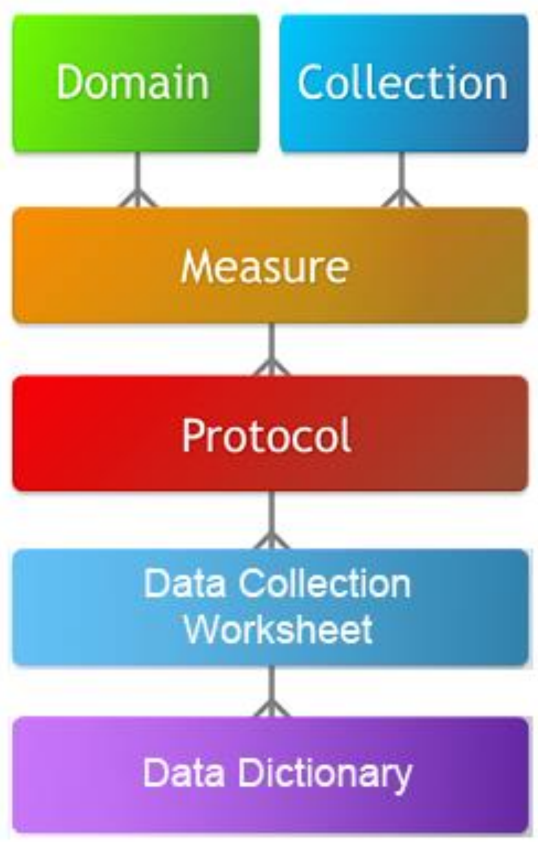
Disease	Gene(s)
WOLF-HIRSCHHORN SYNDROME	-
PRADER-WILLI SYNDROME (TYPE 1)	-
NF1-MICRODELETION SYNDROME	-
22Q11 DELETION SYNDROME (VELOCARDIOFACIAL / DIGEORGE SYNDROME)	-
SOTOS SYNDROME	-
1P36 MICRODELETION SYNDROME	-
POTOCKI-LUPSKI SYNDROME (17P11.2 DUPLICATION SYNDROME)	-
CRI DU CHAT SYNDROME (5P DELETION)	-
22Q13 DELETION SYNDROME (PHELAN-MCDERMID SYNDROME)	-
MILLER-DIEKER SYNDROME (MDS)	-
CHARCOT-MARIE-TOOTH SYNDROME TYPE 1A (CMT1A)	-
WILLIAMS-BEUREN SYNDROME (WBS)	-

PhenX toolkit

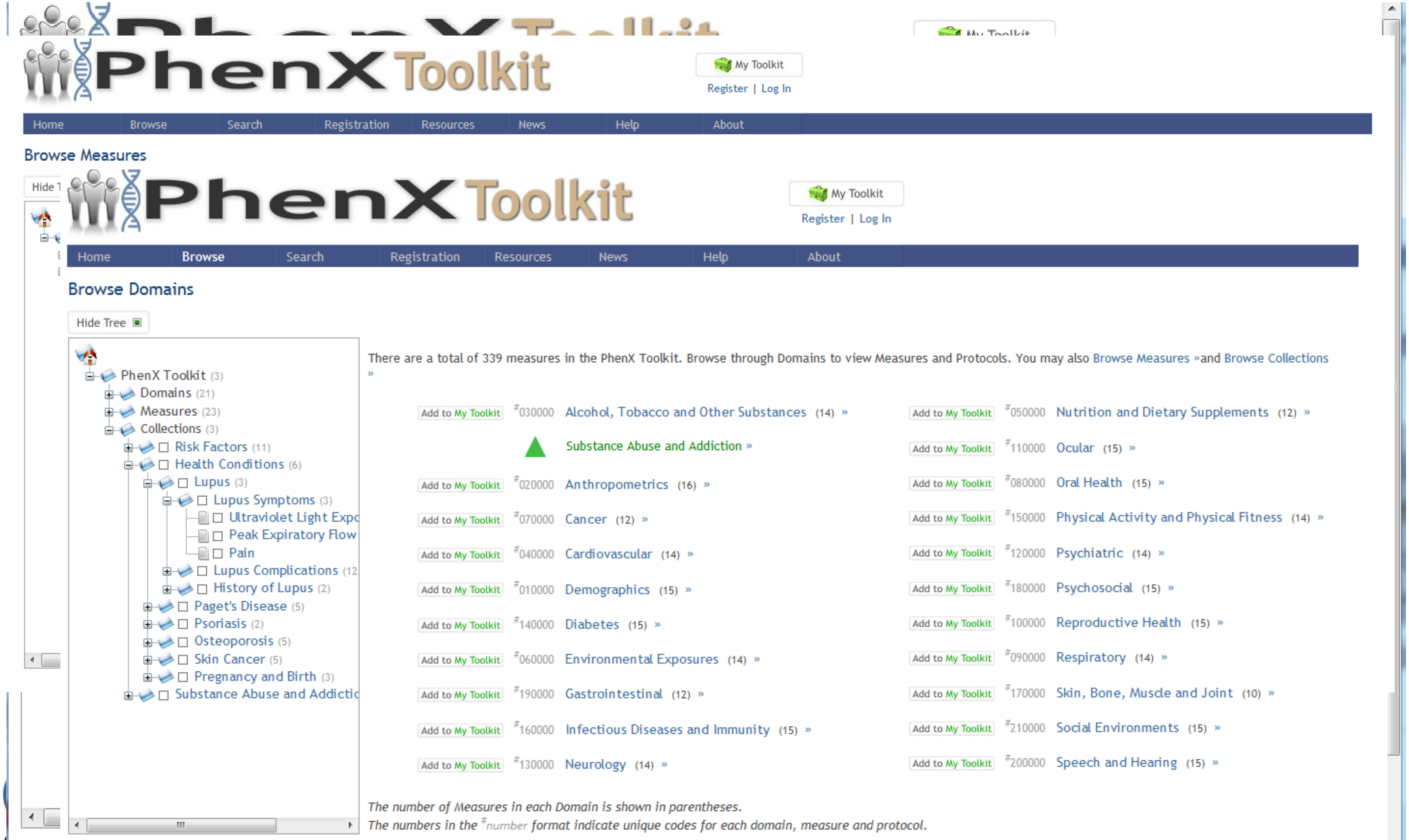
- Well-established, broadly validated measures of phenotypes and exposures relevant to investigators in human genomics, epidemiology, and biomedical research
- Provides detailed protocols, information about the measures
- Provides a toolkit for developing CRF

PhenX toolkit

- **Domain** –field of research e.g. demographics, anthropometrics, organ systems, complex diseases
- **Collection** –group of measures e.g. population
- **Measure** –way of collecting data from participant
- **Protocol** –standard way to collect & record measure
- **Data Collection Worksheet** –identifies items in a protocol
- **Data Dictionary -**lists each variable included in a protocol with its attributes, including variable names and unique identifiers



PhenX toolkit example



PhenX Toolkit

My Toolkit

Register | Log In

Home | Browse | Search | Registration | Resources | News | Help | About

Browse Measures

Hide T

PhenX Toolkit

My Toolkit

Register | Log In

Home | Browse | Search | Registration | Resources | News | Help | About

Browse Domains

Hide Tree

- PhenX Toolkit (3)
  - Domains (21)
  - Measures (23)
  - Collections (3)
    - □ Risk Factors (11)
    - □ Health Conditions (6)
      - □ Lupus (3)
        - □ Lupus Symptoms (3)
          - □ Ultraviolet Light Expo
          - □ Peak Expiratory Flow
          - □ Pain
        - □ Lupus Complications (12
        - □ History of Lupus (2)
      - □ Paget's Disease (5)
      - □ Psoriasis (2)
      - □ Osteoporosis (5)
      - □ Skin Cancer (5)
      - □ Pregnancy and Birth (3)
    - □ Substance Abuse and Addictio

There are a total of 339 measures in the PhenX Toolkit. Browse through Domains to view Measures and Protocols. You may also Browse Measures »and Browse Collections »

Add to My Toolkit	#030000	Alcohol, Tobacco and Other Substances	(14) »
		Substance Abuse and Addiction »	
Add to My Toolkit	#020000	Anthropometrics	(16) »
Add to My Toolkit	#070000	Cancer	(12) »
Add to My Toolkit	#040000	Cardiovascular	(14) »
Add to My Toolkit	#010000	Demographics	(15) »
Add to My Toolkit	#140000	Diabetes	(15) »
Add to My Toolkit	#060000	Environmental Exposures	(14) »
Add to My Toolkit	#190000	Gastrointestinal	(12) »
Add to My Toolkit	#160000	Infectious Diseases and Immunity	(15) »
Add to My Toolkit	#130000	Neurology	(14) »
Add to My Toolkit	#050000	Nutrition and Dietary Supplements	(12) »
Add to My Toolkit	#110000	Ocular	(15) »
Add to My Toolkit	#080000	Oral Health	(15) »
Add to My Toolkit	#150000	Physical Activity and Physical Fitness	(14) »
Add to My Toolkit	#120000	Psychiatric	(14) »
Add to My Toolkit	#180000	Psychosocial	(15) »
Add to My Toolkit	#100000	Reproductive Health	(15) »
Add to My Toolkit	#090000	Respiratory	(14) »
Add to My Toolkit	#170000	Skin, Bone, Muscle and Joint	(10) »
Add to My Toolkit	#210000	Social Environments	(15) »
Add to My Toolkit	#200000	Speech and Hearing	(15) »

The number of Measures in each Domain is shown in parentheses.

The numbers in the #number format indicate unique codes for each domain, measure and protocol.

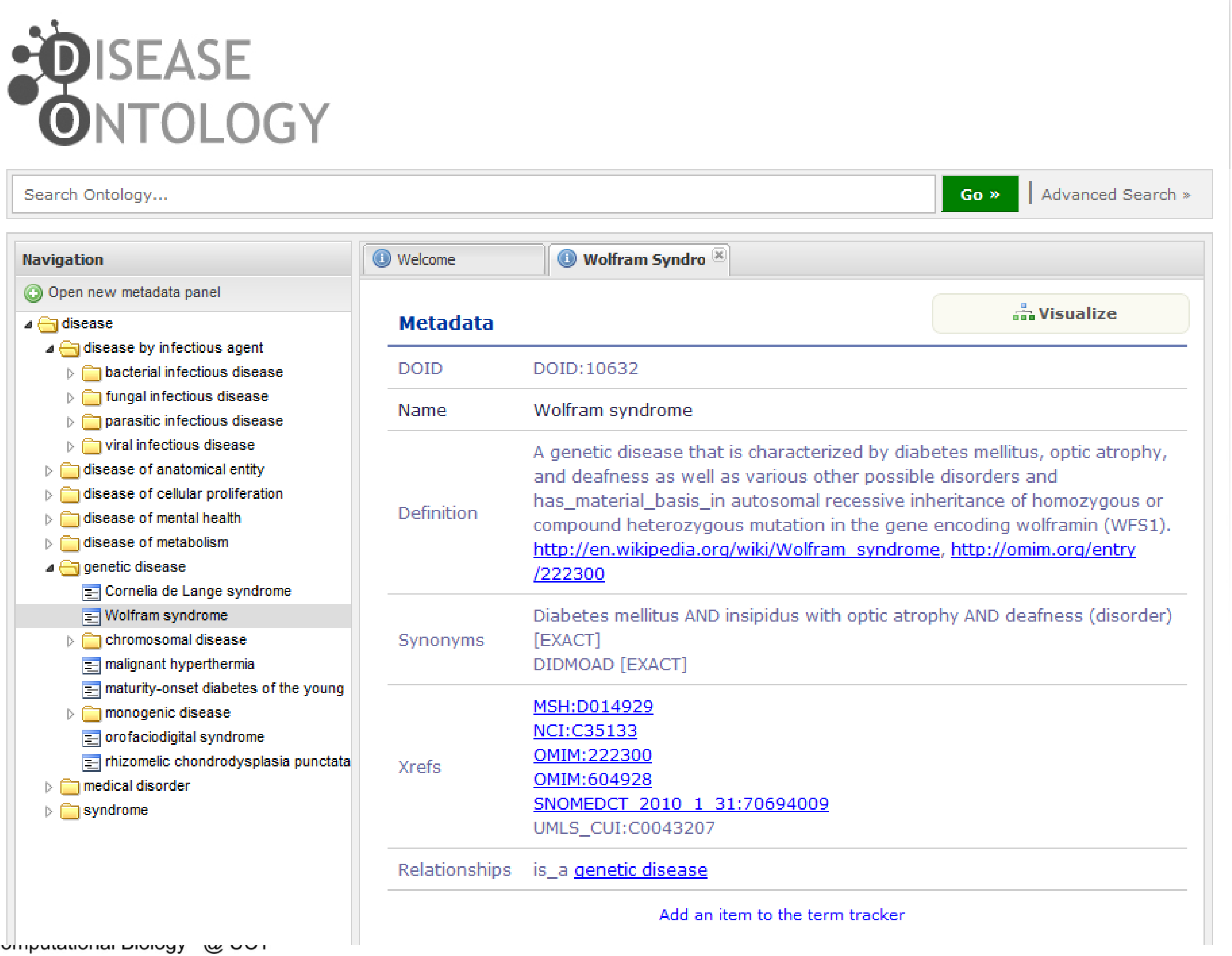
Symptom ontology

- http://bioportal.bioontology.org/ontologies/SYMP and http://symptomontologywiki.igs.umaryland.edu/wiki/index.php/Main_Page
- "A perceived change in function, sensation or appearance reported by a patient indicative of a disease"

Disease Ontology

- http://do-wiki.nubic.northwestern.edu/do-wiki/index.php/Main_Page and http://www.disease-ontology.org/
- “A disease is a disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism."
- Community-driven open source ontology for human diseases
- Link to SNOMED, ICD-9, ICD-10, MeSH, UMLS

Disease ontology example



- disease
  - disease by infectious agent
    - bacterial infectious disease
    - fungal infectious disease
    - parasitic infectious disease
    - viral infectious disease
  - disease of anatomical entity
  - disease of cellular proliferation
  - disease of mental health
  - disease of metabolism
  - genetic disease
    - Cornelia de Lange syndrome
    - Wolfram syndrome
    - chromosomal disease
    - malignant hyperthermia
    - maturity-onset diabetes of the young
    - monogenic disease
    - orofaciodigital syndrome
    - rhizomelic chondrodysplasia punctata
  - medical disorder
  - syndrome

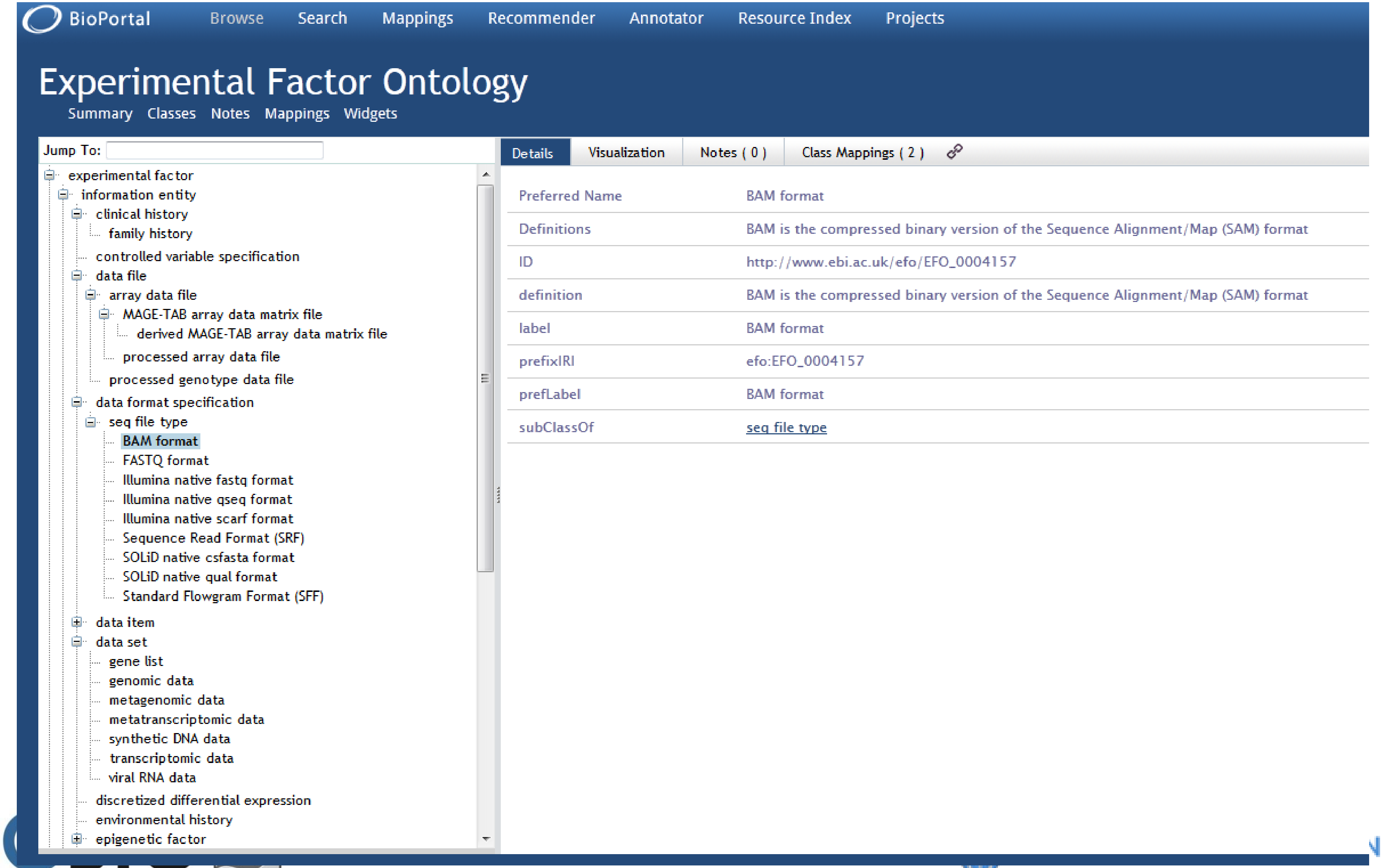
Metadata

DOID	DOID:10632
Name	Wolfram syndrome
Definition	A genetic disease that is characterized by diabetes mellitus, optic atrophy, and deafness as well as various other possible disorders and has_material_basis_in autosomal recessive inheritance of homozygous or compound heterozygous mutation in the gene encoding wolframin (WFS1). http://en.wikipedia.org/wiki/Wolfram_syndrome, http://omim.org/entry/222300
Synonyms	Diabetes mellitus AND insipidus with optic atrophy AND deafness (disorder) [EXACT] DIDMOAD [EXACT]
Xrefs	MSH:D014929 NCI:C35133 OMIM:222300 OMIM:604928 SNOMEDCT_2010_1_31:70694009 UMLS_CUI:C0043207
Relationships	is_a genetic disease

Experimental factor ontology

- http://www.ebi.ac.uk/efo/
- Description of experimental variables
- Combines parts of other ontologies, e.g. anatomy, disease and chemical compounds
- Recommended by EGA

Experimental ontology example 1

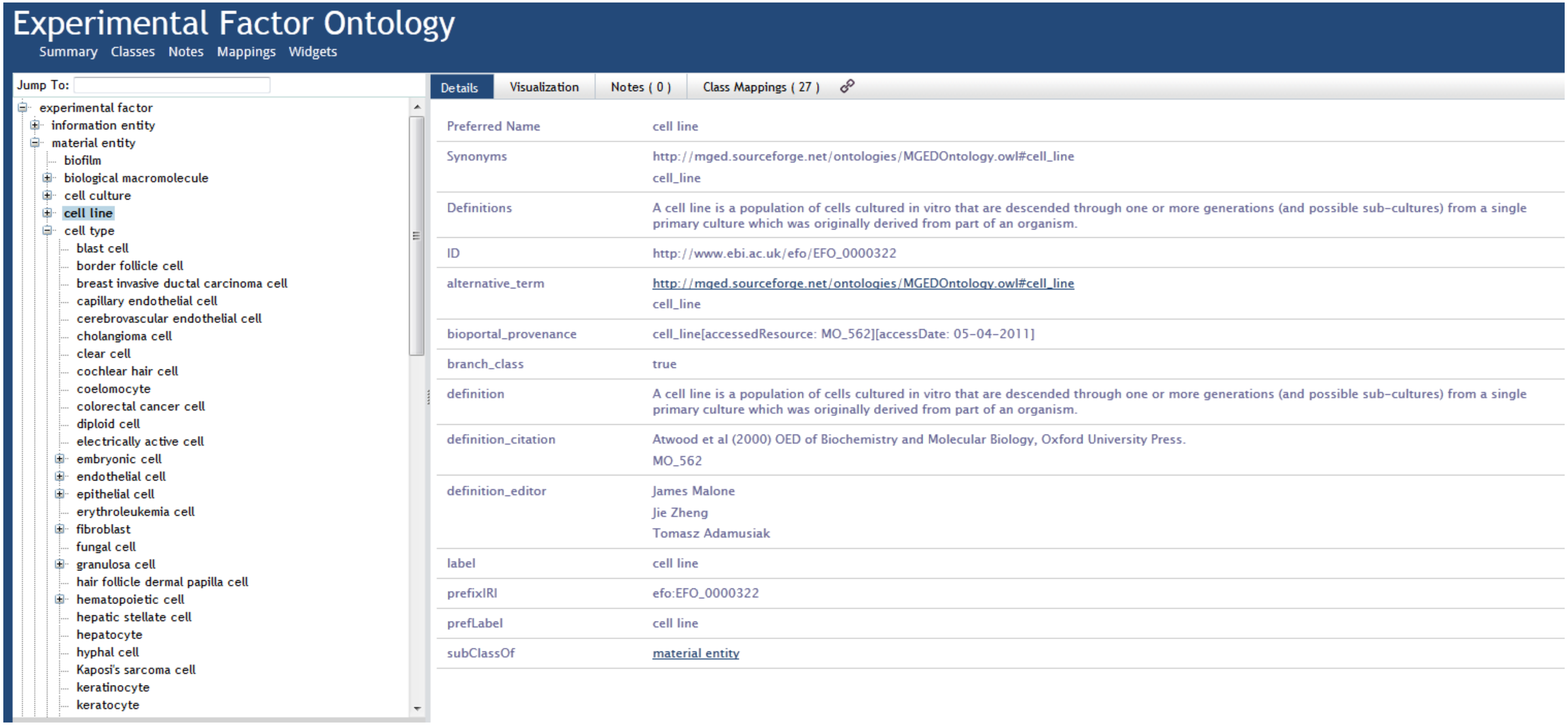


Experimental Factor Ontology

- experimental factor
  - information entity
    - clinical history
      - family history
    - controlled variable specification
    - data file
      - array data file
        - MAGE-TAB array data matrix file
          - derived MAGE-TAB array data matrix file
        - processed array data file
      - processed genotype data file
    - data format specification
      - seq file type
        - BAM format
        - FASTQ format
        - Illumina native fastq format
        - Illumina native qseq format
        - Illumina native scarf format
        - Sequence Read Format (SRF)
        - SOLiD native csfasta format
        - SOLiD native qual format
        - Standard Flowgram Format (SFF)
    - data item
    - data set
      - gene list
      - genomic data
      - metagenomic data
      - metatranscriptomic data
      - synthetic DNA data
      - transcriptomic data
      - viral RNA data
    - discretized differential expression
    - environmental history
    - epigenetic factor

Preferred Name	BAM format
Definitions	BAM is the compressed binary version of the Sequence Alignment/Map (SAM) format
ID	http://www.ebi.ac.uk/efo/EFO_0004157
definition	BAM is the compressed binary version of the Sequence Alignment/Map (SAM) format
label	BAM format
prefixIRI	efo:EFO_0004157
prefLabel	BAM format
subClassOf	seq file type

Experimental ontology example 2



Preferred Name	cell line
Synonyms	http://mged.sourceforge.net/ontologies/MGEDOntology.owl#cell_line cell_line
Definitions	A cell line is a population of cells cultured in vitro that are descended through one or more generations (and possible sub-cultures) from a single primary culture which was originally derived from part of an organism.
ID	http://www.ebi.ac.uk/efo/EFO_0000322
alternative_term	http://mged.sourceforge.net/ontologies/MGEDOntology.owl#cell_line cell_line
bioportal_provenance	cell_line[accessedResource: MO_562][accessDate: 05-04-2011]
branch_class	true
definition	A cell line is a population of cells cultured in vitro that are descended through one or more generations (and possible sub-cultures) from a single primary culture which was originally derived from part of an organism.
definition_citation	Atwood et al (2000) OED of Biochemistry and Molecular Biology, Oxford University Press. MO_562
definition_editor	James Malone Jie Zheng Tomasz Adamusiak
label	cell line
prefixIRI	efo:EFO_0000322
prefLabel	cell line
subClassOf	material entity

H3Africa phenotype harmonization WG

- Leverage the significant investment being made in cohorts and genomic analyses
- Multi-site and cross-consortium analyses – more statistically powerful and informative
- Encourage H3Africa to use PhenX standardized phenotype measures for Case Report Forms
- Encouraged to collect a set of 25 essential and 10 discretionary phenotypes

“Essential” phenotypes

(1) Age & (2) Sex
(3) Country of birth
(4) Current residence
(5) Native language
(6) Ethno-linguistic/tribal affiliation
(7) Country of birth of father and mother
(8) Native language of father and mother
(9) Ethno-linguistic/tribal affiliation of mother and father
(10) Height
(11) Weight
(12) Current medications
(13) Smoking history
(14) Alcohol history

Self-reported personal history (streamlined - yes/no and age at diagnosis):

(15) Hypercholesterolemia
(16) Hypertension
(17) Myocardial infarction
(18) Arrhythmia
(19) Rheumatic fever/rheumatic heart disease
(20) Asthma or reactive airway disease
(21) Stroke
(22) Diabetes
(23) Kidney disease
(24) HIV
(25) Tuberculosis

“Discretionary” phenotypes

- (1) Schizophrenia
- (2) Cancer (especially cervical cancer)
- (3) Malaria
- (4) Trypanosomiasis
- (5) Substances of abuse history
- (6) Blood pressure measurement
- (7) Urinalysis (albumin, creatinine, protein)

more detailed questions on:

- (8) Diabetes - Type 1 and Type 2
- (9) Kidney disease
  (10) Rheumatic heart disease

Additional phenotypes- measurements being collected by at least 2 of the larger Collaborative Center grants

- Waist circumference
- Hip circumference
- Ultrasound fat measurements
- Family history of stroke, heart attack, hypertension

Importance of using ontologies

- Share information in a common structure
- Enable reuse of domain knowledge
- Make naming/entities explicit and unambiguous
- Used for searching of data
- For H3Africa it will increase power to do cross-project studies