International Conference on Biomedical Ontology
Buffalo, NY - July 24, 2009

**Towards Desiderata for an Ontology of Diseases
for the Annotation of Biological Datasets**

Olivier Bodenreider
National Library of Medicine
Bethesda, Maryland - USA

Anita Burgun
Medical School / Univ. Hospital
University of Rennes, France

Motivation

- ◆ Many terminologies and ontologies available for the domain of phenotypes and diseases, e.g.,
 - Disease Ontology
 - Mouse Phenotype Ontology
 - International Classification of Diseases
 - Disease hierarchy of general terminologies
 - SNOMED CT
 - Medical Subject Headings (MeSH)
 - NCI Thesaurus
 - ...
- ◆ Which one to choose for annotation purposes?

Motivation

- ◆ Some issues with existing disease ontologies
 - Biological ontologies often do not link to human diseases
 - Clinical ontologies are not widely used for the annotation of biological datasets
 - Terminology integration resources are not easy to use
 - Incomplete coverage
 - NCI Thesaurus: focus on cancer

Objectives

- ◆ Propose a list of desirable features for an ontology of diseases suitable for the annotation of biological datasets
- ◆ Analyze a list of candidate terminologies through the framework provided by these features

Methods Overview

- ◆ Select candidate biomedical ontologies covering the subdomain of diseases
- ◆ Establish a list of desirable characteristics from these ontologies
- ◆ Apply the list of characteristics to each ontology
- ◆ Rank candidate ontologies by degree of compliance with the desirable characteristics

Candidate ontologies Methods

- ◆ Explore repositories of ontologies
- ◆ BioPortal (NCBO)
 - Metadata (“phenotype”, “health”)
- ◆ Unified Medical Language System (UMLS)
 - Expert knowledge

Candidate ontologies Results

- ◆ Disease ontology (DO)
- ◆ Online Mendelian Inheritance in Man (OMIM)
- ◆ International Classification of Diseases (ICD)
- ◆ SNOMED CT
- ◆ Medical Subject Headings (MeSH)
- ◆ NCI Thesaurus (NCIt)
- ◆ Unified Medical Language System (UMLS)
- ◆ Human Phenotype Ontology (HPO)
- ◆ Phenotypic Quality Ontology (PATO)
- ◆ Mammalian Phenotype Ontology (MPO)
- ◆ Logical Observation Identifiers Names and Codes (LOINC)

Desirable characteristics Methods

- ◆ The ontology must be *open* and available to be used by all without any constraint
- ◆ The ontology is in, or can be expressed in, a *common shared syntax*. This may be either the OBO syntax, extensions of this syntax, or OWL.
- ◆ The ontologies possesses a *unique identifier space* within the OBO Foundry.
- ◆ The ontology provider has procedures for identifying distinct successive *versions*.
- ◆ The ontology has a clearly specified and clearly *delineated content*.
- ◆ The ontologies include textual *definitions* for all terms.
- ◆ The ontology uses relations which are unambiguously defined following the pattern of definitions laid down in the *OBO Relation Ontology*.
- ◆ The ontology is *well documented*.
- ◆ The ontology has a plurality of independent *users*.
- ◆ The ontology will be developed *collaboratively* with other OBO Foundry members.

<http://www.obofoundry.org/crit.shtml>

Desirable characteristics Results

Weight

- 1 ♦ No intellectual property restrictions
- 1 ♦ Standard, friendly format
- 2 ♦ Existence of a mapping to clinical terminologies
- 2 ♦ Harmonization with other biological ontologies
- 1 ♦ Regular maintenance
- 5 ♦ Exhaustive coverage of diseases
- 2 ♦ Support for automatic reasoning

Example

Legend 1 0.5 0

	Weight	MPO		
freely available	1		1 x 1	1
friendly format (OBO, OWL, RDF)	1		1 x 1	1
mapping to clinical terminologies	2		0 x 2	0
harmonized with OBO ontologies	2		0 x 2	0
regularly maintained	1		1 x 1	1
coverage of diseases	5		0.5 x 5	2.5
support for automatic reasoning	2		0 x 2	0
total	14	39%		5.5

$$5.5 / 14 = 39\%$$

Desirable characteristics matrix

Nature Precedings : doi:10.1038/npre20093536.1 : Posted 1 Aug 2009

		Mendelian Inher. Man	Observation Ident.	Mammalian Phen.	Human Phenotype	Int. Classif. Diseases	Med. Subject headings	Phenotypic Quality	Disease Ontology	SNOMED CT	NCIt	UMLS
Legend		1	0.5	0								
freely available	1
friendly format (OBO, OWL, RDF)	1
mapping to clinical terminologies	2
harmonized with OBO ontologies	2
regularly maintained	1
coverage of diseases	5
support for automatic reasoning	2
Total	14	32%	32%	39%	39%	50%	50%	54%	64%	64%	68%	68%

Findings

Disease Ontology
SNOMED CT
NCI Thes.

Legend	● 1	▲ 0.5	◆ 0	Weight	DO	SN. CT	NCIt	UMLS
freely available	1	0.5	0	1	●	◆	●	▲
friendly format (OBO, OWL, RDF)	1	0.5	0	1	●	◆	●	◆
mapping to clinical terminologies	2	0.5	0	2	●	▲	●	●
harmonized with OBO ontologies	2	0.5	0	2	◆	◆	◆	▲
regularly maintained	1	0.5	0	1	◆	●	●	●
coverage of diseases	5	0.5	0	5	●	●	▲	●
support for automatic reasoning	2	0.5	0	2	◆	●	●	◆
total	14	64%	64%	68%	68%			

- ◆ Wide range in the degree of compliance with our desirable features
- ◆ The top 4 contenders
 - Are close
 - High score for different reasons

Differences with OBO Foundry principles

- ◆ Legacy ontologies
 - Not always freely available
 - Not always available in OWL or OBO format
 - Not always developed collaboratively
 - Not always coordinated with other ontologies
- ◆ But worth considering
 - Coverage
 - Mapped to clinical terminologies
- ◆ OBO Foundry principles might be too rigid for the purpose of selecting disease ontologies for the annotation of biological datasets

Limitations

- ◆ Coarse criteria
- ◆ Arbitrary weighting scheme
 - Needs to be adapted to specific annotation scenarios

Conclusions

- ◆ Framework for evaluating disease ontologies available for annotation purposes
- ◆ Abundance of disease ontologies
 - Not necessarily good news to researchers
 - Should discourage the development of new ontologies in this domain
- ◆ Annotations made to different ontologies need to be reconciled in order to enable semantic interoperability
 - Role of ontology integration resources

Medical Ontology Research

Contact: olivier@nlm.nih.gov

Web: mor.nlm.nih.gov

Olivier Bodenreider

Lister Hill National Center
for Biomedical Communications
Bethesda, Maryland - USA