Proceedings of the First International Conference on Biomedical Ontology (ICBO 2009) 2009:39-42.

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

Olivier Bodenreider[1], Anita Burgun[2]
[1] LHNCBC, National Library of Medicine, Bethesda, MD, USA
[2] U936, IFR 140, Faculté de Médecine, University of Rennes 1, France

Abstract

There is a plethora of disease ontologies available, all potentially useful for the annotation of biological datasets. We define seven desirable features for such ontologies and examine whether or not these features are supported by eleven disease ontologies. The four ontologies most closely aligned with our desiderata are Disease Ontology, SNOMED CT, NCI thesaurus and UMLS.

Introduction

Ontologies have been developed for the annotation of biological datasets from multiple perspectives including functional annotation of gene products (Gene Ontology), molecular sequences (Sequence ontology) and phenotypes (Mammalian Phenotype Ontology, Phenotypic Quality Ontology). Entries in biological datasets also need to be linked to diseases, either human diseases or experimental models of diseases in model organisms. Ontologies of diseases include the Disease Ontology (DO), from the Open Biomedical Ontology (OBO) family. The NCI Thesaurus was developed for the annotation of cancer research and includes many diseases, but its focus on cancer can be a limitation for use in other domains.

On the other hand, terminologies have been long been developed for the purpose of annotating clinical records, including the International Classification of Diseases (ICD) and SNOMED CT. However, these terminologies have not been widely adopted by biomedical researchers for annotating disease entities in biological datasets. Moreover, neither terminology is free of intellectual property restrictions and a license or fee may be required for their use, which represents a limiting factor.

Finally, terminology integration resources such the Unified Medical Language System (UMLS) Metathesaurus and NCBO's BioPortal both integrate more than one hundred biomedical terminologies, including all those mentioned above. Moreover, both resources provide mappings across terminologies, facilitating the integration of biological and clinical data required for translational medicine. However, their use may necessitate significant training.

The objective of this study is to propose a list of desirable features for an ontology of diseases suitable for the annotation of biological datasets, and to analyze a list of candidate terminologies through the framework provided by these features.

Desiderata for selecting ontologies have been established by the OBO Foundry[1]. While interesting and potentially relevant to the domain of diseases, we find some of these criteria unnecessarily restrictive for the purpose of annotating biological datasets, while key criteria (from our specific perspective) are missing. A brief analysis of the OBO Foundry criteria in the context of our study is proposed in the discussion.

This work also differs from Cimino's desiderata for controlled medical vocabularies[2] in that we focus on content and usability for a particular purpose in addition to representation issues and development process.

Methods

We first select a list of biomedical terminologies and ontologies (hereafter referred to simply as ontologies) potentially suitable for the annotation of diseases in biological datasets. We establish a list of characteristics from these ontologies, focusing on those characteristics which represent potential barriers to adoption of these terminologies by biomedical researchers. We apply the list of features to each candidate ontology and summarize our findings in a feature x ontology matrix.

Candidate ontologies

In order to identify candidate ontologies for diseases, we explored the two major repositories of biomedical ontologies: The Unified Medical Language System (UMLS) and NCBO's BioPortal. We investigated ontologies whose focus is on human diseases and phenotypes, as well as ontologies which contain a significant number of disease entities. In practice, we exploited the metadata provided with OBO ontologies and selected those ontologies for which the domain is contains "phenotype" or "health". No similar mechanism is available for the UMLS and we simply used our knowledge of the source vocabularies to make our selection. References to the ontologies discussed below are listed in Table 1. This selection process led to the identification of eleven ontologies potentially suitable for the annotation of diseases in biological datasets.

- **Disease Ontology** (DO): Controlled terminology from the OBO family created for annotation purposes as part of the NuGene project at Northwestern University. Coverage restricted to diseases.
- **Online Mandelian Inheritance in Man** (OMIM): Knowledge base on human genetic diseases developed at John Hopkins University and available through the NCBI Entrez system. Its terminological component – including clinical synopses – is available through the UMLS. Coverage restricted to genetic diseases.
- **International Classification of Diseases** (ICD): Classification from the World Health Organization (WHO) family of health classifications, with many local adaptations. ICD9-CM, developed by the Center for Medicare & Medicaid Services (CMS) for use in the US, includes clinical modifications. Coverage restricted to diseases and health problems.
- **SNOMED CT**: The largest clinical terminology developed by the International Health Terminology Standard Development Organization (IHTSDO) for use in electronic health records and adopted by eleven countries to date. Broad coverage including diseases.
- **Medical Subject Headings** (MeSH): Controlled vocabulary developed by the U.S. National Library of Medicine for the indexing and retrieval of the biomedical literature, especially in the MEDLINE bibliographic database. Broad coverage including diseases.
- **NCI Thesaurus** (NCIt): Controlled vocabulary developed by the National Cancer Institute to support the integration of information related to cancer research. Broad coverage including diseases.
- **Unified Medical Language System** (UMLS): Terminology integration system developed by the U.S. National Library of Medicine, establishing a correspondence among terms from different terminologies for a given biomedical entity. Broad coverage including diseases.
- **Human Phenotype Ontology** (HPO): Controlled vocabulary for the phenotypic features encountered in human hereditary and other diseases. Developed by a consortium including Charite Hospital (Berlin) and the University of Cambridge (UK). Coverage restricted to monogenic diseases listed in OMIM.
- **Phenotypic Quality Ontology** (PATO): Ontology of phenotypic qualities, intended for use in a number of applications, primarily defining composite phenotypes and phenotype annotation. Coverage restricted to phenotypes.
- **Mammalian Phenotype Ontology**: Controlled vocabulary for the “robust annotation of mammalian phenotypes” currently used for the annotation of phenotypic data in mouse and rat databases. Developed at the Jackson Laboratory. Coverage restricted to phenotypes.
- **Logical Observation Identifiers Names and Codes** (LOINC): Set of names and codes for laboratory and other clinical observations (elements of clinical phenotypes). Developed at the Regenstrief Institute. Coverage restricted to clinical observations.

Phenotype ontologies for organisms other than *Homo sapiens* were ignored. (e.g., **Yeast phenotypes**). Ontologies of diseases included as part of a broader ontology were ignored when they were unlikely to provide additional coverage or characteristics useful for the discussion in this paper (e.g., **National Drug File Reference Terminology** and **International Classification of Primary Care**). Specialized resources (e.g., **Online Congenital Multiple Anomaly/Mental Retardation Syndromes**, **Infectious Disease Ontology** and **Diagnostic and Statistical Manual of Mental Disorders**), while providing deep coverage of a narrow subdomain of medicine, are unlikely to provide the broad coverage expected from an ontology of diseases and were ignored.

Desirable features

Starting from the ten OBO Foundry principles, we have identified seven desirable features for an ontology of diseases. In each case, the absence of a feature represents a potential barrier to the adoption of a biomedical ontology for the annotation of diseases in biological datasets. Differences with the set of OBO Foundry principles are discussed later in this paper.

- **No intellectual property restrictions**. The use of some vocabularies is limited to certain contexts (e.g., restriction for research purposes vs. production systems for some vocabularies in the UMLS) or to certain countries (e.g., member countries of the IHTSDO for SNOMED CT), or subject to the payment of a fee (e.g., ICD 10). This feature is aligned with Foundry principle #1.

- **Standard, friendly format**. Availability of terminologies in formats that are standard (e.g., RDF, OWL) or friendly to biologists (e.g., OBO) is likely to foster adoption. In contrast, proprietary formats (e.g., RRF for the UMLS Metathesaurus) may represent a barrier to adoption. This feature corresponds roughly to Foundry principle #2.
- **Existence of a mapping to clinical terminologies**. In the era of translational medicine, biological datasets must be linkable to clinical datasets. The existence of mappings between an ontology of diseases used for the annotation of biological datasets and clinical terminologies used in patient records is strong requirement.
- **Harmonization with other biological ontologies**. Similarly to the requirement for integration with clinical terminologies, there is a need for a disease ontology to be integrated – if possible natively – with other biological ontologies. This feature corresponds roughly to Foundry principle #5.
- **Regular maintenance**. The domain of diseases is in constant evolution and an ontology of disease shall reflect emerging diseases and changes in our understanding of the domain of diseases. This feature corresponds roughly to Foundry principle #4.
- **Exhaustive coverage of diseases**. At a given level of granularity, the ontology shall provide an exhaustive coverage of the domain. Terminologies focusing on a specific subdomain may have limited applicability outside this subdomain (e.g., focus on cancer in NCIt).
- **Support for automatic reasoning**. Annotations made to ontologies often form the basis for gaining new knowledge about biomedical entities. In order to process annotations efficiently and automatically, ontologies need to have a robust, formal structure and provide support for automated reasoning (e.g., through subsumption).

A framework for comparing disease ontologies

The desirable features listed above do not all have the same importance from the perspective of an ontology of diseases for annotation purposes. For example, coverage of diseases it of the outmost importance for an ontology of diseases and was given the highest weight (5). Interoperability with other ontologies (clinical and biological) and support for automatic reasoning correspond to major uses of ontologies and are also weighted more (2) than the remaining features (1).

We examined the eleven candidate ontologies through the prism of the seven desirable features. More precisely, for each feature, we rated the ontology semi-quantitatively: 0 (no or minimal support for the feature), 0.5 (partial support of the feature) or 1 (reasonable support for the feature), assessed by the authors. The weights were applied to the ratings. Finally, the score of each ontology was computed by comparing the sum of the scores for each feature to the sum of all weights (14).

Results

The result of assessing the presence of the desirable features in the candidate ontologies is summarized in Figure 1. Support for the desirable features ranges from 32% (OMIM, LOINC) to 68% (NCIt, UMLS). Seven ontologies have a score of 50% or more.

Discussion

Applying the desiderata. The top four contenders identified in our matrix of desirable features x ontologies (Figure 1) are Disease Ontology, SNOMED CT, NCIt and UMLS. Interestingly, these four ontologies made it to the top for slightly different reasons. Depending on what features are most important in a given use case, the ontologies corresponding to this profile of features should be selected.

Phenotypes vs. diseases. Precisely defining phenotype and disease is beyond the scope of this paper. However, we observed that phenotype ontologies containing pre-coordinated concepts (e.g., **Mammalian Phenotype Ontology**, **LOINC**) or supporting post-coordination (e.g., the **Phenotypic Quality Ontology - PATO**), cover low-level phenotypes and clinical observations (e.g., individual anatomical and physiological abnormalities) rather than diseases. Examples of phenotypes form MPO include *enlarged liver*, found in ontologies including MeSH, NCTt and SNOMED CT. In contrast, they mostly contain terms indicating deviation from normal anatomical structures or physiologic states (e.g., *decreased liver weight*), typically absent from the clinically-oriented disease ontologies. Phenotype ontologies seem suitable for the annotation of data with low-level phenotypes, whereas disease ontologies have application in the annotation of higher-order information about diseases, i.e., resulting from some elaborate diagnostic process.

Differences with OBO Foundry criteria. Although some of our desirable features are aligned with principles of the OBO Foundry[1], we found the Foundry principles to be generally too rigid for the purpose of annotating biological datasets and lacking consideration for legacy ontologies. Applying these principles strictly to the selection of ontologies would potentially result in unnecessarily excluding from consideration the datasets annotated to these legacy ontologies.

Many legacy disease ontologies are not available in OWL or OBO format, but are widely used. Borrowing from "orthogonal" ontologies is a good principle for the coordinated development of ontologies (i.e., applied in a prospective manner). However, this principle can hardly be held against legacy disease ontologies. The absence of textual definition is a common feature to many legacy disease ontologies. It can be offset in part by the presence of formal definitions (in description logic-based systems) and usage information. Finally, most widely used disease ontologies are developed outside the OBO Foundry and not always in a collaborative manner.

Limitations. The framework provided here for analyzing disease ontologies is relatively coarse and somewhat arbitrary. The list of desirable features and the weights would need to be adapted to specific annotation scenarios. For example, the presence of synonyms is required if annotations are to be discovered automatically in text corpora using text mining techniques.

Conclusions

The plethora of disease ontologies available to biomedical researchers for annotation purposes is not necessarily good news. In this domain in particular, reusing existing ontologies should be carefully considered before starting the development of a new one. Annotations made to different ontologies, including legacy ontologies, will likely need to be reconciled in order to enable interoperability among datasets, which is a strong requirement for translational medicine. Terminology integration systems such as the UMLS are thus expected to play a key role in data integration tasks.

Acknowledgements

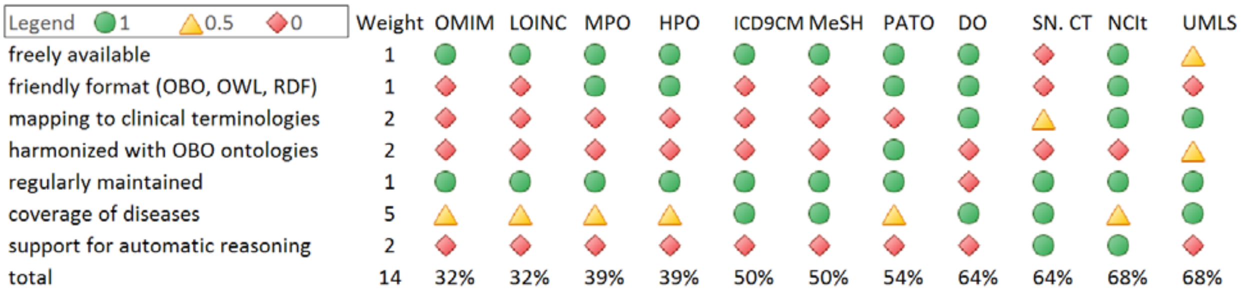
This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

1. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 2007;25(11):1251-5.
2. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4-5):394-403.

Table 1. List of potential disease ontologies discussed in this paper

DO	*Disease Ontology* - http://diseaseontology.sourceforge.net/
OMIM	*Online Mandelian Inheritance in Man* - http://www.ncbi.nlm.nih.gov/omim/
ICD	*International Classification of Diseases* - http://www.who.int/classifications/icd/en/
SNOMED CT	*SNOMED CT* - http://www.ihtsdo.org/
MeSH	*Medical Subject Headings* - http://www.nlm.nih.gov/mesh/
NCI Thes.	*NCI Thesaurus* - http://cancer.gov/cancerinfo/terminologyresources/
UMLS	*Unified Medical Language System* - http://www.nlm.nih.gov/research/umls/
HPO	*Human Phenotype Ontology* - http://www.human-phenotype-ontology.org/index.php/hpo_home.html
PATO	*Phenotypic Quality Ontology* - http://www.bioontology.org/wiki/index.php/PATO:Main_Page
MPO	*Dictionary of Medicines and Devices* - http://www.informatics.jax.org/searches/MP_form.shtml
LOINC	*Logical Observation Identifiers Names and Codes* - http://loinc.org



Legend ● 1 ▲ 0.5 ◆ 0	Weight	OMIM	LOINC	MPO	HPO	ICD9CM	MeSH	PATO	DO	SN. CT	NCIt	UMLS
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%

Figure 1. Desiderata applied to candidate disease ontologies (matrix of desirable features x ontologies)