Preprint version of
Lindsay Grey Cowell and Barry Smith, "Infectious Disease Ontology", in Vitali Sintchenko, *Infectious Disease Informatics*, New York: Springer, 2010, 373-395.

19 Infectious Disease Ontology

Lindsay Grey Cowell[1] and Barry Smith[2]

[1]Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, North Carolina, U.S.A.

[2]Department of Philosophy, Center of Excellence in Bioinformatics and Life Sciences, University at Buffalo, Buffalo, New York, U.S.A.

Abstract In the last decade, technological developments have resulted in tremendous increases in the volume and diversity of the data and information that must be processed in the course of biomedical and clinical research and practice. Researchers are at the same time under ever greater pressure to share data and to take steps to ensure that data resources are interoperable. The use of ontologies to annotate data has proven successful in supporting these goals and in providing new possibilities for the automated processing of data and information. More recently, ontologies have been shown to have significant benefits both for the analysis of data resulting from high-throughput technologies and for automated reasoning applications, and this has led to organized attempts to improve the structure and formal rigor of ontologies in ways that will better support computational analysis and reasoning. In this chapter, we describe different types of vocabulary resources and emphasize those features of formal ontologies that make them most useful for computational applications. We describe current uses of ontologies and discuss future goals for ontology-based computing, focusing on its use in the field of infectious diseases. We review the largest and most widely used vocabulary resources relevant to the study of infectious diseases and conclude with a description of the Infectious Disease Ontology suite of interoperable ontology modules that together cover the entire infectious disease domain.

Acknowledgments: LGC's contributions were supported by a Career Award from the Burroughs-Wellcome Fund and NIAID grants R01 AI077706 and R01 AI068804. BS's contributions were funded in part through the NIH Roadmap for Medical Research grant to the National Center for Biomedical Ontology (1 U 54 HG004028). Initial development of the Infectious Disease Ontology as well as the Infectious Disease Ontology meetings were generously supported by the Burroughs-Wellcome Fund.

19.1 Vocabulary Resources for Biomedicine

Vocabulary resources have been used in biology and medicine at least since the time of Linnaeus, whose work on classification extended not only to organisms but also, in his *Genera morborum* (1763), to the classification of diseases. Linnaeus' work (and through it Aristotle's ideas on classification) continues to play an influential role in terminology and taxonomy work today.

Initially, vocabularies and terminologies existed in the form of printed dictionaries compiled for human use, and such resources continue to play an important role, for example in education. The primary use of vocabulary resources of interest to us, however, is in fostering the presentation of biomedical and clinical data and information in ways that can support the use of computation in research. In this context, vocabulary resources have been developed for purposes of bibliographic search, coding of clinical and public health data, and database interoperability. For example:

- The Medical Subject Headings (MeSH) vocabulary (http://www.nlm.nih.gov/mesh/meshhome.html), first published in 1954, is used to support literature indexing and document retrieval for the MEDLINE database of biomedical literature;
- The International Classification of Diseases (ICD) (http://www.who.int/classifications/icd/en/), first published as the International List of Causes of Death in 1893, is the international standard for coding diagnostic information for health and vital records and is also commonly used for hospital billing purposes;
- SNOMED (http://www.snomed.org), first released in 1965, was initially developed to support documentation of pathology data and is projected to become a worldwide reference vocabulary for structured clinical documentation;
- The Gene Ontology (GO) (http://www.geneontology.org/), created in 1998, is a vocabulary resource for the annotation of gene and gene-product data facilitating interoperability between a large number of diverse databases, especially in the domain of model organism research.

In the last decade, there has been an increasing need for biology and medical terminologies to support more sophisticated computational algorithms requiring high precision. This is a consequence of i) tremendous increases in the volumes and types of data and information coming out of biomedical and clinical research, resulting in the need for computational assistance for the analysis and interpretation of these data, ii) pressure to implement electronic health records, and iii) in-

creased interest in the possibilities of automated reasoning for biomedical research, clinical decision support, and biosurveillance.

In addition to the increased need for machine interpretable vocabulary resources, there is a growing need also for vocabularies to be interoperable across institutional and disciplinary boundaries. In both the biological and clinical domains, interoperability across sub-disciplines is critical to advancing scientific understanding. The emergence of translational medicine as a new field and the push to use clinical data for research have increased the need for interoperability between the research and clinical care domains. The formation of public data repositories and the movement of patients between health care systems both put additional requirements on vocabularies to be interoperable across institutions. Analogous requirements are also increasingly being felt in the domains of public health and disease and pathogen surveillance.

Unfortunately, existing biomedical and clinical vocabularies are in many ways incompatible because they were developed for a variety of different purposes and by multiple separate communities. They have different underlying semantics, employ different linguistic and logical structures, and manifest varying degrees of formal rigor (described in detail below). As a consequence, they are not interoperable and most do not support sophisticated computing of the sort that is becoming central to informatics-driven biomedical research. Increasing reliance on the computer processing of data and information and the need for cross-domain interoperation have highlighted the need for more structure and formal rigor in vocabulary resources. Due to their enhanced formal capabilities to support computing, interoperation, and reasoning, ontologies are being advanced as a new kind of terminology resource that can provide a necessary foundation for biomedical and clinical research in the future.

In what follows, we describe, the different types of vocabulary resources available in the infectious disease domain, covering the spectrum of terminology-based representational artifacts from simple taxonomies, wordlists, glossaries, and loosely structured thesauri through data dictionaries to the more highly formalized 'ontologies' now increasingly being applied in biomedical research. We will emphasize those features of formal ontologies that make them most useful for computational applications. We will then describe the various uses of ontologies in biomedical and clinical research, describe existing vocabulary resources relevant to infectious diseases, and conclude with some speculations concerning the potential uses of ontologies in the future.

19.2 Types of Vocabulary Resources

All vocabulary resources consist of terms; they differ in how these terms are presented and organized. Most importantly for our present purposes, vocabulary resources differ in whether terms are provided with definitions, in the types of relationships asserted between terms or the entities to which the terms refer, and in the degree of logical rigor underlying definitions and relations.

The simplest vocabulary resources are term lists (with or without definitions), containing no information about how the terms or the entities to which the terms refer are related to each other beyond what can be inferred from the terms themselves when considered linguistically. Examples include nomenclatures such as the Human Genome Organization (HUGO) Gene Nomenclature (http://www.hugo-international.org/committee_nomen.htm) and the Nomenclature for Factors of the Human Leukocyte Antigen (HLA) system (http://www.anthonynolan.org.uk/HIG/nomen/reports/homen/reports.html).

The majority of vocabulary resources, however, assert a simple term hierarchy or taxonomy in which the relationships between terms indicate that one term has a narrower meaning than another, or that one type of thing (e.g. dog) is classified as a subtype of another type of thing (e.g. animal). ICD and MeSH are examples of this type of resource. Vocabulary resources that assert a richer set of relations are less common. The best example is the Foundational Model of Anatomy (FMA) (http://sig.biostr.washington.edu/projects/fm/), which includes backbone hierarchies structured by means of taxonomic (*is_a*) and partonomic (*part_of*) relations and various formally defined spatial relations representing adjacency, connectedness and relative position.

Many vocabulary resources, including many medical glossaries, have poor structural organization and provide at best definitions written in natural language for interpretation by human users. This means that they are poorly suited for computational purposes. Providing definitions based on a formal theory (such as [1]) enhances the potential utility of a vocabulary resource for computation, but requires a non-trivial investment of resources, especially for the large vocabulary resources often found in the biomedical domain.

Similarly, there is great variability in the degree to which the relations used in the structure of vocabulary resources are formalized in a way that supports automatic reasoning. In MeSH for example, relations are presented primarily in an implicit fashion through the relative position of terms in the MeSH hierarchy. Among vocabulary resources with explicitly asserted relations, the vast majority provides either no definition of the relations, or provides only natural language descriptions of the intended meaning of relational expressions. At the other, more formally rigorous, end of the spectrum are a growing number of vocabulary resources employing relations defined according to a formal theory, for example within the context of the Semantic Web [2] and of the Open Biomedical Ontologies (OBO) Foundry Initiative [3].

Following what is increasingly becoming standard usage, we shall here employ the term ‘ontology’ to refer to a vocabulary resource that is structured by means of relations between its terms and is logically formalized in the sense that the developers adhere to a logical theory in the definition of terms and relations, for example as outlined in [1,4]. Vocabulary resources of this sort are standardly represented as graph-theoretical structures built up out of terms as the nodes of the graph and relations as edges [5]. While there are a variety of other meanings associated with the term ‘ontology’, the usage here is consistent with that of large in-

fluential ontology developer and user groups, including the Gene Ontology Consortium (http://www.geneontology.org/), the W3C community (http://www.w3.org/), and the OWL Web Ontology Language community (http://www.w3.org/2004/OWL).

The different uses for which the different vocabulary resources have been built have determined to a large extent the degree and type of structure, level of detail, and logical formalism used in their construction. We argue, however, that even when the intended application does not require a highly structured and formalized vocabulary resource, there are benefits to be gained from developing the resource with a structured and formalized approach in ways that adhere to best practice guidelines. First, such an approach results in vocabulary resources that have fewer developer-introduced errors. Second, the resulting vocabulary resources can be subjected to automated error checking [6,7,8]. Third, structured and formalized resources are likely to be free of idiosyncratic features and are therefore more broadly applicable. Thus the development of a structured and formalized vocabulary resources can facilitate their reusability and utility as biomedical research becomes increasingly reliant on computation [9].

A simple illustration of the advantages already resulting from a greater formal organization of a vocabulary resource is how this organization makes possible a more complete and more focused retrieval of data. Without formal organization, searches against data catalogued on the basis of mere word lists are restricted to the use of string matches, which is highly ineffective especially in a domain like infectious diseases, where data is derived from many heterogeneous sources and nomenclature is poorly standardized. Formal organization means that, when collecting information about a given disease or pathogen, we can automatically extend our search to include corresponding subtypes or variants independently of how the latter are named. Another simple benefit of formal organization is the ability to ensure that the effects of changes to a classification are automatically propagated to all relevant parts of the classification.

The use of classifications which rest on a well-defined and reliably executed use of the subclass or subtype relation (called in what follows '*is_a*') is crucial to the realization of these benefits. Here the test of reliability is conformity to the rule: if type *A* is classified as a subtype of *B*, then all instances of *A* (for example, all cases of a given infectious disease) are also instances of *B*.

One consequence of conformity to this rule is that the *is_a* relation will be transitive (if we know that *A is_a B* and *B is_a C*, then we can infer also that *A is_a C*). For example, if we know that *Staphylococcus aureus is_a Staphylococcus* and *Staphylococcus is_a bacterium*, then we can infer that *Staphylococcus aureus is_a bacterium*.

Another consequence is that all instances of *A* will inherit the properties shared by all instances of *B*. For example, bacteria of the genus Staphylococcus are facultative anaerobes. If this is asserted in the ontology, along with *Staphylococcus aureus is_a Staphylococcus*, *Staphylococcus aureus* will inherit the property of being a facultative anaerobe. Inheritance is an important source of potential bene-

fits from the use of vocabulary resources in automatic reasoning. Definition and use of the *is_a* relation are discussed in more detail below.

Terminological note: Where type *A* stands in an *is_a* relation to type *B* in a classificatory hierarchy, we shall also describe '*A*' as the child term and '*B*' as parent. Any given child can have sibling terms in the sense of terms that share a common parent. Further discussion of the different types of vocabulary resources can be found in [9,10,11,12].

19.3 Features of Ontologies Needed to Support Informatics

For ontologies to support sophisticated computational algorithms with high precision, it is necessary that they be developed in accordance with certain principles of ontology development best practice. In particular, adherence to the following has been shown to enhance support for computation: i) the use of Aristotelian definitions with a single mode of classification; ii) the use of single inheritance hierarchies; iii) the use of relations with formal, logical definitions based on a distinction between types and instances; and iv) writing definitions and ontology assertions as compositions of ontology terms and relations rather than as natural language.

The definition of types in an ontology serves an important purpose beyond describing the meaning of the term that refers to the type, and that is to specify the placement of the types in the ontology's inheritance hierarchy. This is accomplished through the use of Aristotelian definitions, the form of which is *A is_a B* which *C*, where *A* is the type being defined, *B* is its *genus* (parent or supertype), and *C* is the *differentia* [1,13]. It is the first part of the definition, *A is_a B*, which results in inheritance, as *A* will inherit all of the properties of *B*, including those properties *B* inherits from its parent. *B* may have many subtypes, and it is the differentia, *C*, that distinguishes *A* from the other subtypes of *B*. For example, in a hierarchy of disease types, one could define *infectious disease* as a *disease* which is caused by an infection.

In addition to the use of Aristotelian definitions, it is recommended that a single mode of classification be adopted for any given hierarchy, that is, all types within a single hierarchy should be differentiated based on the same type of criterion. It is further recommended that each type have only a single parent type. Hierarchies in which all types have only a single parent are referred to as single inheritance hierarchies, whereas hierarchies in which types can have more than one parent are referred to as multiple inheritance hierarchies. The problem with using multiple modes of classification and with allowing multiple inheritance is that the meaning of the *is_a* relation becomes uncertain, resulting in errors on the part of both maintainers and users of an ontology [14] and the inability to use the hierarchy for automated reasoning. For example, in SNOMED, *is_a* has in some contexts the meaning "has cause" (e.g. *Tuberculosis of meninges is_a Mycobacteriosis*), while in others it means "has location" *(e.g. Tuberculosis of meninges is_a Disorder of meninges*). The use of *is_a* with multiple meanings is often referred to

as '*is_a* overloading' [15]. While in practice it can be difficult to avoid multiple inheritance, even within a single mode of classification, multiple modes of classification (and therefore multiple meanings for *is_a*) should be avoided by using the corresponding specific relations (e.g. *has_location*). The benefits are not only an ontology that has fewer errors, is easier to maintain, and can be used for automated reasoning, but also a reduced loss of information by using the more specific representation. Other considerations in the classification of biological entities are outlined in detail in [13,14].

Successful inferencing over the relations asserted between ontology types relies on a single, logical definition for each relation with clearly specified implications. This is best accomplished by distinguishing between types (e.g. influenza infection) and instances (e.g. each of the individual cases of influenza infection), and defining the relations between types in terms of the relations between the corresponding instances [4]. Thus, a type-level relation *R* will be defined in terms of the instance-level relation **R** by: $X\ R\ Y =_{\text{def}}$ for every instance *x* of *X*, there exists at least one instance *y* of *Y* such that *x* **R** *y*, where uppercase indicates types (*X*, *Y*) and lowercase indicates instances (*x*, *y*). For example, *human has_part brain* means that every instance of *human* has as part of it some instance of *brain*. Defining the relations between types in terms of the relations between instances, and specifying that the type-level relation *X R Y* holds when the instance-level relation *x* **R** *y* holds for *all* instances of *X* ensures that *X R Y* holds universally. This, in turn, ensures transitivity, which can be used for automated reasoning: if *X R Y* and $Y\ R_1\ Z$, then there is some relation R_2 such that $X\ R_2\ Z$. The distinction between types and instances corresponds to the distinction between A-boxes and T-boxes used in the Owl/Semantic Web community [16].

In almost all natural-language-based vocabulary resources thus far, terms and definitions have been treated in effect as black boxes, so that their logical content is not accessible to computational tools. The GO, along with its sister ontologies in the OBO Foundry, has initiated an ambitious strategy to expose the compositional character of compound terms and definitions by conceiving them as cross-products of simpler terms, some of which are derived from other ontologies [3,17]. For example, rather than defining *Tuberculosis of the meninges* with the natural language phrase "Tuberculosis of the meninges is a *Mycobacterium tuberculosis* infection in which the site of infection is the meninges", one can instead use formally defined relations between ontology terms to create structured phrases such as:

> *Tuberculosis of the meninges is_a Mycobacterium tuberculosis infection* THAT *has_location meninges*

where *meninges* is a term in an anatomy ontology, such as the FMA, and *Mycobacterium tuberculosis infection* is a term in an ontology of infectious diseases, such as the Infectious Disease Ontology described below, and is itself defined as a cross-product. By this means, the potential for the ontology to support automatic

reasoning and error checking is enhanced, and so also is its capacity to integrate data in the direction of enhanced semantic interoperability.

That the enhanced formalism and logical rigor of ontologies relative to other vocabulary resources brings significant benefits to applications is perhaps best evidenced by the relative numbers of citations for the GO, SNOMED, and the Unified Medical Language System (UMLs) in the PubMed database. As the name implies, the UMLS, initiated in 1986, is an attempt to provide a unified terminology system for the medical domain. The goal is two-fold: to make the many medically relevant vocabulary resources interoperable, and to create a single, broad coverage resource. The strategy used by the UMLS developers is to integrate the many existing medical terminologies by providing joint access to them through mappings between their terms. The UMLS includes the GO and SNOMED, as well as MeSH and ICD, among its source terminologies. Despite its short history and small domain relative to SNOMED and the UMLS, the GO has become the most cited vocabulary resource in PubMed, with over 450 citations per year [18]. In contrast, the number of UMLS citations has remained constant over the last 10 years [18]. From 2001 to 2007, among papers that cite the GO, SNOMED, the UMLS, the FMA, MeSH, the National Cancer Institute Thesaurus (NCIT), and the Logical Observation Identifiers, Names, and Codes (LOINC) vocabulary, the proportion of GO citations increased from about 5% to about 85%, while the proportion citing SNOMED decreased from about 20% to about 5% and the proportion citing the UMLS decreased from about 55% to about 5% [18].

As can be seen from the description of ontology uses below, the utility of ontologies in computational applications depends not just on adherence to development principles like those outlined above, but also on the breadth of the developer and user communities. When each community develops and uses its own ontology, many of the benefits of ontology are not realized. To address both of these issues, the Open Biomedical Ontologies (OBO) Foundry (http://obofoundry.org) [3] was initiated in 2006. The goals of the Foundry are to foster the pursuit of best practice in ontology development on the basis of an evolving set of design principles and to provide a foundation for the coordinated development of ontologies by large developer and user communities. Its ontologies are designed to represent in an interoperable fashion the biomedical reality from which data are sampled. Their development within the framework of a common top-level ontology, the Basic Formal Ontology and the consistent employment of a constrained set of logically defined relations allows Foundry ontologies to be used together as modules of a larger system for computational applications.

There are currently some 35 member ontologies at varying stages of development in the OBO Foundry. There are OBO Foundry ontologies covering many of the domains relevant to infectious diseases, including proteins (the Protein Ontology [19]; cells (the Cell Ontology [20]); human anatomy (the FMA [1]); anatomy for important vector species (the Tick Gross Anatomy Ontology and the Mosquito Gross Anatomy ontology (http://www.anobase.org/)); and biological processes,

molecular functions, and cellular components (the Gene Ontology (http://www.geneontology.org).

19.4 Uses of Ontologies in Informatics-driven Research and Care

Vocabulary resources have a long history of use in clinical settings, primarily to support the coding of clinical data for health records, laboratory reports, and hospital billing, the coding of public health data for monitoring disease incidence and prevalence, and the coding of knowledge for clinical decision support systems. In basic biomedical research, the primary use of vocabulary resources has, until recently, been to support bibliographic searches and database integration. However, the logical rigor and formalism underlying biomedical ontologies has increased significantly in recent years, allowing biomedical ontologies to be applied for a larger variety of purposes.

For ontologies and the data annotated in their terms, we find a variety of different types of uses in biomedicine, outlined in [9,18,21], including terminology management; text-mining; integration, interoperability, and sharing of data; data interpretation and analysis; and knowledge reuse, reasoning, and decision support. Ontologies support terminology management in aligning independently developed terminologies with overlapping content [22,23]. They also bring benefits in managing changes to terminologies by allowing flexible response to new scientific discoveries, as contrasted with the relative inflexibility of more traditional database approaches, where a database schema may need to be revised in its entirety when one aspect of classification changes.

Ontologies are increasingly used to add value to more traditional vocabulary resources, whose informal structure and lack of systematic definitions "is generally deemed to be inadequate with respect to the requirements of health care information systems that depend on clear communication of complex medical and biological information in a form that is usable by computers" [9]. Applying a formal structure to vocabulary resources allows enhanced opportunities for both manual and automatic error checking. Ontological methods are used to detect errors in definitions and to analyze the meanings of terms and represent those meanings formally [7,8,24]. Additionally, ontological methods are used to detect errors in classification, such as the improper assignment of *is_a* relations arising through inadequate treatment of negation, or the improper assignment of part-whole relations resulting from an inconsistent use of terms in different parts of terminology [6,7].

In the area of text-mining, vocabulary resources are used to facilitate the retrieval of information from biomedical literature (reviewed in [18,25]). The greatest success has come from the assignment of terms from vocabulary resources to individual documents within large collections, a process referred to as indexing. MeSH has long been used to index documents within the PubMed database [26], and, more recently, ontologies have been used for this purpose, allowing text-

mining algorithms to take advantage of the richer set of relations and their formal definitions [27,28,29]. The identification of documents that are relevant to a query within a collection (document retrieval) is greatly facilitated by utilizing the ontologies' structure. For example, the hierarchy of *is_a* relations can be used to expand a query to include parents or children of the original query term, significantly improving recall. *part_of* relations can be similarly used, retrieving documents that refer to fingers or palms in response to a query for documents that refer to hands.

After the identification of relevant documents, text-mining often progresses to information extraction, the identification within documents of statements about pre-specified entities. Named entity recognition is the simplest approach in which a list of entities of interest is provided as input to the information extraction algorithm. The terms from ontologies can serve as an important source of term lists for named entity recognition, and the ontologies' structure can serve to improve information extraction just as document retrieval is improved.

Within the area of infectious disease research, ontology-supported text-mining is used to monitor news reports from all over the world so as to detect disease outbreaks, monitor the geographic distribution of diseases (BioCaster, http://biocaster.nii.ac.jp/; EpiSpider, http://www.epispider.org/) and predict candidate vaccine epitopes [30]. Ontologies have also been developed to support text-mining about Dengue fever, specific Dengue virus serotypes [31], and vaccine development and efficacy. The Vaccine Investigation and Online Information Network (VIOLIN, http://www.violinet.org) was established as a central repository for literature related to vaccine research and the data resulting from vaccine research. In addition to a variety of data analysis tools, VIOLIN provides several text-mining tools supported by its Vaccine Ontology (VO), as well as MeSH and the Textpresso Ontology [28].

Currently the most successful use of ontologies is to support integration, interoperability, and the sharing of data through data annotation. The best example is use of the GO for the creation of annotations by the curators of model organism databases [32,33,34] and genome annotation centers [35]. GO curators are striving to capture, in a form accessible to computational algorithms, information about the contributions of gene products to biological systems, as reported in the scientific literature. The annotation process unfolds in a series of steps [36]. First, specific experiments, documented in the biomedical literature, are identified as relevant to the responsibilities of a given ontology curator. Second, the curator applies expert knowledge to the documentation of the results of each selected experiment. This process entails determining which entities (for example which proteins) are being studied in the experiment, the nature of the experiment itself, and (in the case of the Gene Ontology) the molecular functions, biological processes and cellular components that the experiment identifies as being associated with that gene product. The curator then creates an annotation, which captures the appropriate relationships between the corresponding ontology types and the database entry for the gene product type. The annotated data then become accessible through the use

of the associated Gene Ontology term as a search vehicle and becomes automatically combined with many other types of relevant and useful information as a result of the fact that the curators of many other types of data are using the same controlled vocabulary resource to annotate their data. Developing the ontology in tandem with the process of curation of data also provides a means of ensuring that the ontology is maintained in a way that keeps pace with the advance of science as recorded in the published literature and ensures that the vocabulary provides the resources needed to express the most recent scientific results.

The GO and other ontologies are used for annotation of genes and gene products in a variety of databases relevant to infectious disease research. In addition to the annotation of data for humans and for model organisms, such as mice, which are used to study the host immune response, ontologies are used to annotate data in:

- the ApiDB databases (http://eupathdb.org/eupathdb/), which include genomic and other data for *Cryptosporidium*, *Giardia*, *Plasmodium*, *Theileria*, *Toxoplasma*, and *Trichomonas* strains;
- VectorBase (http://www.vectorbase.org), which includes genomic and other data for invertebrate vectors of human pathogens, including *Anopheles gambiae*, *Aedes aegypti*, *Ixodes scapularis*, *Pediculus humanus*, and *Culex quinquefasciatus*;
- the Integrated Microbial Genomes System (http://img.jgi.doe.gov/), Microbes Online database (http://www.microbesonline.org/), the Pathogen-Host Interaction Data Integration and Analysis System (http://phidias.us), BioHealthBase (http://www.biohealthbase.org/), and the National Microbial Pathogen Data Resource (http://www.nmpdr.org/), among others [37], together include annotations for the genomes of hundreds of bacterial and viral species, as well as a significant number of eukaryotic pathogen species;
- many of the databases listed at http://databases.biomedcentral.com/ under the "infectious diseases" subject area.

In addition to the annotation of genomic data, the use of ontologies and other vocabulary resources to annotate other types of data is also becoming common. For example, data in ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/) has been annotated with GO terms as well as terms from the Microarray Gene Expression Data (MGED) Ontology [38]. Of particular interest in the study of infectious diseases is the use of MeSH to annotate human disease names to microarray data in the Gene Expression Omnibus [39] and the use of GO and SNOMED to annotate pathways and integrate disease and pathway information [40].

Ontology annotations not only provide a basis for database interoperability, but also significantly enhance the interpretation of data from genome-wide and high-throughput experiments, as for example in [41,42,43,44]. A variety of software tools has been developed to use ontologies and other vocabulary resources for the

analysis and interpretation of microarray data, including Onto-Tools [40], GoMiner [45], GOTree Machine [46], MeSHer [47], and more recently [48,49].

Ontology annotations have formed the basis for new bioinformatics approaches for the analysis of such data [49,50]. One such method for the analysis of microarray data is the CLASSIFI algorithm [51] which determines, for sets of genes clustered based on their expression levels, whether particular gene ontology terms are overrepresented within any set of genes. Ontologies have also been used to enhance clustering algorithms for microarray data by using the ontology annotations as a second cluster variable [52,53,54]. In another study, proteins were clustered based on the similarity of their GO annotation profiles. The annotations for each protein were represented as a graph, and the graph similarity for pairs of proteins was used as the distance measure for clustering [55]. This method was applied to sets of proteins from two different protein array screens, and in both cases, proteins not identified in the original study were implicated to play a role in the biological process under study [55]. Finally, ontologies have been used to integrate text-mining approaches with microarray data analysis to facilitate disease gene identification [56].

An important benefit of ontologies is that they facilitate knowledge re-use. While knowledge-based systems that support applications such as decision support in health care are typically dependent on large amounts of current domain knowledge, the capture of such knowledge in computationally accessible information systems through data curation is an expensive and arduous process. In the domain of molecular biology, the widespread adoption of the Gene Ontology as a standard vocabulary has worked well, eliminating the need for developers of different information systems to expend resources capturing the same knowledge. In the clinical domain, however, knowledge capture has standardly been performed with the aid of locally developed database schemas and vocabulary resources, both structured to the specific application at hand. Such database schemas and vocabulary resources do not support the reuse or cumulation of data and often lose their validity within a short space of time. Increasingly, therefore, there is a move, illustrate by the caBIG endeavor, to foster the development of reusable resources for data capture in which, again, ontologies and ontology-related technologies are again playing an important role.

The use of ontologies to support automated reasoning is an active area of research and recent work, described below, has shown that the benefits of even primitive reasoning algorithms can be significant. These results have led to increased interest in developing vocabularies with sufficient formalism to support reasoning as well as in developing reasoning algorithms that make use of the types of information captured in ontologies. An important application area of automated reasoning is clinical decision support.

Query engines have been developed in such a way that the ontology itself is a directly query-able knowledge resource. For example, Emily [57] is a system used to query the FMA for structural relationships between anatomical entities. The FMA also serves as a source of anatomical knowledge in a reasoning application

used to predict the consequences of penetrating injury [58]. The system is used to determine which organs are injured and whether vital structures, such as a coronary artery, are injured given particular projectile trajectories [58]. HyBrow is a system that uses ontologies and ontology annotations as sources of existing knowledge to test whether hypotheses are consistent with existing knowledge and data, to rank hypotheses by the amount of supporting evidence, and to test the implications of hypotheses [59].

Clinical decision support systems (CDSS) are commonly used in the infectious diseases field for diagnostic assistance, guidance in the prescription of anti-infectives, biosurveillance, and vector control (Global Infectious Disease and Epidemiology Network, http://www.gideononline.com, and [60,61,62,63,64,65,66,67]). Vocabulary resources, such as classifications of drug types, serve as a source of knowledge for CDSS. In most cases, however, simple terminology lists or term hierarchies are used, and when vocabulary resources with more complex relations are used, the resources are developed for the purposes of the specific application and do not have sufficient logical formalism to serve the purposes of broad interoperability. For example, the clinical vocabulary resource with the broadest scope, and which also has many ontology-like features, is SNOMED. A recent review of the literature found little evidence that SNOMED is being used for direct care purposes such as CDSS [68]. The use of ontologies, as we have defined them, in CDSS is still a young field of research. One prominent example is the use of ontologies in the Dengue Decision Support System (http://www.rams-aid.org/) developed by the Risk Assessment and Management Solutions for Arthropod-borne and Infectious Diseases group at Colorado State University. There is a growing effort within the OBO Foundry community to develop ontologies with coverage of the clinical domain and to develop ontology-based reasoning algorithms, including those useful within CDSS.

19.5 Vocabulary Resources Relevant to the Field of Infectious Diseases

We provide a brief review of vocabulary resources that have content relevant to the infectious diseases domain, restricting ourselves primarily to those resources that are freely available, widely used and likely to persist. For each resource, we describe its intended use and evaluate its adequacy and prospects for general use in infectious disease research and clinical care, taking account of the considerations outlined below.

The vocabulary resources relevant to this review can be divided into two broad groups: resources produced primarily as terminologies for use in the clinical domain, and resources developed in support of research in the basic biological sciences. In light of the increasing focus on translational medicine, we take it that the trajectory of clinical and biomedical sciences is towards an ever closer alignment of these two groups of resources, which have hitherto evolved almost entirely independently. Therefore, one focus of our evaluation has been to gauge the degree

to which existing clinical and biomedical terminology resources can support this trajectory. The second focus is on evaluating the degree to which such resources support the increasing demand for more sophisticated information processing capabilities.

19.5.1 Medical Subject Headings (MeSH) controlled vocabulary

MeSH is a general purpose vocabulary, initially developed for purposes of indexing and cataloging medical literature, now used to support many text- and literature-mining endeavors. Terms from the MeSH controlled vocabulary are used to annotate biomedical journal article citations and abstracts for the MedLine database. Query interfaces to MedLine, such as PubMed, use MeSH to support the retrieval of MedLine records in ways which supplement the use of simple string searches.

MeSH is a controlled vocabulary organized as a thesaurus consisting of sets of terms or 'descriptors' in a hierarchical structure that permits searching at various levels of specificity. The relationship between terms in a hierarchy is not *is_a*; rather the terms appear in the MeSH term hierarchies on the basis of relatedness as assessed in terms of fields of study or research (a strategy designed to maximize the utility of MeSH as a literature indexing resource). For example, most of the content relevant to the infectious disease domain is found under one of descriptors *Anatomy, Organisms, Diseases* or *Biological Sciences*. Under *Biological Sciences*, one finds *Public Health*, under which one finds *Disease Outbreaks*, *Disease Reservoirs*, and *Disease Transmission*, along with terms such as *Consumer Product Safety* and *Equipment Reuse*. A natural language note is associated with each term.

MeSH is marked by a broad coverage of topics relevant not only to the domain of infectious diseases but also to microbiology and host immunity. Of all the vocabulary resources we have evaluated, MeSH has the broadest coverage across the entirety of the infectious disease / immunology domain. However, the terms are not linked to any relations, which limits the usefulness of the information contained in MeSH for many purposes. Despite its broad coverage of the subject matter, MeSH cannot be used as a computable vocabulary resource for infectious diseases, though it is highly useful in supporting a variety of string- and statistics-based forms of data and literature mining. Its utility in this respect has been enhanced by its recently completed alignment to the GO [69].

19.5.2 International Classification of Diseases (ICD)

ICD version 10 (ICD-10) is a member of a family of World Health Organization (WHO) international classifications designed to promote international comparability in the collection, processing, classification, and presentation of diagnostics in health epidemiology, health management and mortality statistics. ICD-10 is a classification of diseases and other health problems developed for the purposes of compiling statistics of disease or causes of death. ICD-10 is used to record dis-

ease and other health problems on health and vital records such as death certificates. These records are subsequently used to compile national mortality and morbidity statistics by WHO member states. ICD-10 is also used for general epidemiological and health management purposes, such as monitoring the incidence and prevalence of diseases.

ICD-10 is organized as a term hierarchy in which terms are names of diseases and each term is associated with a code of up to six digits in length, indicating the term's placement in the hierarchy. Terms are defined primarily by their placement in the hierarchy along with statements of inclusion and exclusion. For example, *Tuberculosis* is defined by being a subclass of *Certain infectious and parasitic diseases*, along with the statements "Includes: infections due to *Mycobacterium tuberculosis* and *Mycobacterium bovis*. *Excludes:* congenital tuberculosis, pneumoconiosis associated with tuberculosis, sequelae of tuberculosis, silicotuberculosis."

ICD's coverage of the domain in terms of types of infectious diseases is broad, but information about other aspects of infectious disease is limited and thus the scope of ICD-10 is considered narrow. Because ICD provides a disease classification constructed primarily on the basis of anatomy, it has a relatively robust classification of pathological structures resulting from disease, such as carcinomas and neoplasm, whose classification follows the anatomical partition. For the infectious disease domain, however, a different approach would be needed. The ICD-10 classification of infectious disease is based on many different and inconsistently used classification criteria resulting in a disorganized hierarchy that is counter-intuitive, difficult to navigate, and difficult to construct queries for. Furthermore, there are no formal definitions for terms and no logical basis for the hierarchical structure used. Thus, ICD-10 could not sensibly be used to support either interoperability with other information resources or reasoning within the context of its own hierarchy.

19.5.3 The Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT)

While SNOMED CT is not a fully open source vocabulary resource, its broad scope and the long experience of its use and maintenance, combined with its presumptive status as an international master vocabulary for the coding of clinical information, mean that it is an especially important vocabulary resource for analysis and critical review.

The intended use for SNOMED CT is documentation and reporting of health care information throughout the health care process (medical history, illnesses, treatments, laboratory results, etc.) in software applications used for clinical data collection. The intention is that the processing of health care information recorded in SNOMED CT terms can be used to improve patient outcomes by providing health care providers with more easily accessible and complete information, as

well as to conduct outcomes research, to evaluate the quality and cost of care, and to design effective treatment guidelines.

SNOMED CT is comprised of what are called concepts, concept descriptions, and relationships. A concept is described as a clinical meaning. Concepts are defined by the relationships between them. The primary defining relationship is the *is_a* relation, but there are an additional 50 defining attribute relationships, such as *Finding_site* and *Associated_morphology*.

In general we find that SNOMED CT contains a large number of terms relevant to the infectious disease domain, but that these terms and their organization are biased towards capturing information about clinical observations and about patients in patient records. Terms and relations describing pathogens and the host immune responses to these pathogens are correspondingly lacking. The emphasis on clinical findings and their attributes is not surprising given SNOMED CT's intended use for the documentation and reporting of clinical data, but this does handicap SNOMED CT in terms of its usefulness for translational medicine. This handicap could be overcome of SNOMED CT were developed in accordance with a set of rigorously applied principles sufficient to allow its interoperation with vocabulary resources from the biological domain.

The logical formalism underlying SNOMED CT has been evaluated previously [6,14,70]. Our evaluation based on the infectious disease-relevant content is consistent with these previous evaluations. We observed problems with SNOMED CT's classification hierarchies resulting primarily from the use of multiple modes of classification and a lack of adherence to basic principles of sound classification. The result is the assertion of type-supertype relations that do not hold. For example, the SNOMED class *Infectious disease* is asserted to have subclass *Abrasion AND/OR friction burn with infection*, where neither an abrasion nor a friction burn is itself an infectious disease. Similarly, *Incomplete illegal abortion with genital tract or pelvic infection* is a subtype of *Infectious disease* in SNOMED CT, asserting that a type of abortion is an infectious disease.

As SNOMED becomes more widely used, and begins to serve as a platform to ensure cross-language interoperability of clinical data, it will become ever more urgent that SNOMED meets the highest standards of logical coherence. The SNOMED International Health Terminology Standards Development Organization has recognized many of the above problems and is taking steps to correct them.

19.5.4 The Disease Ontology (DO)

The Disease Ontology was developed for the annotation of patient DNA samples collected with the patients' associated healthcare information. Broader motivations for the creation of the DO were to provide a public domain vocabulary resource for use in data mining against medical records and in annotating model organism phenotype data using terms for human disease.

The DO is organized as a taxonomy of diseases with terms, taken over primarily from ICD, referring to types of diseases. The hierarchy is intended to reflect

the *is_a* relation between disease types. Few terms are defined, but the definitions thus far included are natural language expressions, usually taken from MeSH, SNOMED CT, or the NCI thesaurus. The current DO hierarchy improves somewhat on ICD version 9, and plans for further improvements to the DO are based on a strategy of aligning DO to the SNOMED CT disease typology.

Despite the DO claim of organizing disease terms based on types using an *is_a* relation, the DO hierarchy is poorly organized, mixing not only types of infection with types of disease, but also mixing types based on anatomical location, properties of infection (e.g. latent), type of infectious agent, developmental stage, type of geographical area to which a disease is endemic, and properties of infectious agents (e.g. zoonotic). The mixing of modes of classification and the use of multiple inheritance results in the inheritance of properties that do not hold for a type. For example, *Tuberculosis* is a subtype of *Respiratory Tract Infections* in DO, but not all instances of tuberculosis infection are an infection of the respiratory tract. *Tuberculosis* is also a subtype of *Opportunistic Infections* which is a subtype of *Virus Diseases*, but Tuberculosis is not a viral disease. The DO has a limited utility as a general vocabulary resource for the infectious disease domain due to its limited scope and its disorganized classification hierarchy containing false assertions. The DO developers are, however, aware of these problems, and have initiated efforts towards realizing the necessary reforms.

19.5.5 General conclusions concerning clinical vocabularies

The most common use of clinical vocabulary resources thus far is as dictionaries with the potential to support forms of computer-aided retrieval of information. Vocabularies such as SNOMED CT also have in a certain logical structure, which means that they may be able to support more advanced services, including data integration (for example, the integration of public health data), patient status descriptions, providing codes for problem lists or drug adverse events, and support for text-mining [71]. In addition, they can support certain kinds of reasoning. They are increasingly used in association with basic biology vocabulary resources as tools for clinical and translational research, which are reviewed next.

19.5.6 The Gene Ontology (GO) and Other OBO Foundry Ontologies

We focus here on ontologies within the OBO Foundry, as these ontologies are being developed with the intention of broad interoperability and of their joint use for computation. Although there are still gaps in the domain jointly covered by Foundry ontologies, there is steady progress towards broad coverage of the biomedical domain, including both basic biological and clinical entities.

To fully support informatics-driven infectious disease research, prevention, and treatment, vocabulary resources that cover physiologic and pathologic entities are needed, and within each of those categories, resources are needed that cover: ob-

jects, such as molecules and cells; qualities, functions, and roles of the objects; and processes. The domain of physiologic objects is already well covered within the OBO Foundry by ontologies such as the many anatomy ontologies, the Cell Ontology, the Protein Ontology, and the GO Cellular Component Ontology. In addition, the domains of physiologic processes and molecular functions are also well covered by the GO Biological Process Ontology and the GO Molecular Function Ontology.

However, there are important gaps in the current coverage of the infectious diseases domain by OBO Foundry ontologies. In particular: terms for population-level processes, such as the epidemiological spread of disease; terms for cellular functions, such as the presentation of antigen to naïve T cells; terms for pathological anatomical entities, such as granulomas, and pathological processes, such as hematogenous seeding; terms for roles, such as host, pathogen, vector, carrier, and reservoir; terms for qualities, such as immunocompromised and virulent; and terms for relevant clinical entities, such as clinical phenotypes. In addition, important information is not captured, even about the entities already represented in Foundry ontologies, due to the restricted set of relations currently used. There are, however, large consortia of individuals committed to the development of Foundry ontologies, including the development of a set of ontologies developed specifically for the coverage of the infectious diseases domain (described below). Thus, we anticipate good coverage of the relevant entities in the near future.

Previous evaluations of the GO's implementation and underlying formalism found flaws [8,72,73], but the GO Consortium has responded by working to educate curators and make the necessary changes to the GO ontologies. For example, efforts are under way to create genus-differentiae definitions [1] for all terms, to standardize naming conventions, to utilize rigorous definitions of the GO's two relations, *is_a* and *part_of* [4], and also to add further relations, including relations spanning GO's three constituent ontologies. Development of OBO Foundry ontologies, including revisions and expansion to the GO, adheres to a set of guidelines (http://www.obofoundry.org/crit.shtml) that include the features outlined above and are designed to maximize the long-term utility of Foundry ontologies, in particular for computational applications.

19.5.7 Concluding remarks

The existing vocabulary resources in medicine, such as SNOMED CT, and many of the other source terminologies collected by the UMLS are highly valuable for purposes of data retrieval. However, they were independently developed by separate specialist groups, and thus manifest a low degree of interoperation. They use different naming conventions, different modes of classification, different relations, and different formalisms. Moreover, each has its own independently derived technical implementation. The resulting vocabulary resources are therefore inadequate for purposes of computational and translational medicine; their representations are lacking in both the needed formal rigor and in their coverage of the relevant biological domains. They fall short as cross-domain applications requiring high preci-

sion because they employ uneven standards of rigor. Thus, any information resource created using terms from these terminologies contains insufficient formalism for the sorts of reasoning applications needed for future biomedical and clinical research and translational medicine. Furthermore, the representation of information about the immunobiology and pathogenesis of infectious diseases has thus far been neglected in these terminologies, and this is so even for SNOMED CT, currently the medical terminology with the broadest coverage.

The medical vocabulary resources are also marked by a focus on billing, hospital management abd liability issues, and hence by a centrality in their organization on findings, observations and procedures, with associated epistemological problems. These factors hinder their interoperability with counterpart vocabulary resources developed in the basic biological sciences, where approaches to developing computable vocabulary resources have been developed and tested to a larger degree than in the clinical realm, primarily because the biological data are more highly structured and more readily accessible to researchers.

Biologically focused ontologies and terminologies accordingly employ a more rigorous formalism than do the medical terminologies. Even here, however, the biological content relevant to our purposes is lacking. Formal, computable representations of information about infectious diseases, immunology, and disease pathogenesis are thus still needed.

19.6 The Infectious Disease Ontology (IDO) Consortium

The last five years have seen a surge of interest in biomedical ontology, yet broad coverage, computable vocabulary resources for the infectious diseases domain are lacking. There is resulting in both an urgent need for ontology development in this field and there is an opportunity for a coordinated, community-wide development effort producing broad interoperability across the disease-specific specialties and across the clinical care, public health, and biomedical research domains.

To provide the foundation for such a community-wide ontology development effort, we have established a methodology for the development of ontology modules that together cover the entire infectious disease domain (http://www.infectiousdiseaseontology.org). The methodology relies on the use of a general Infectious Disease Ontology (IDO) that serves as a core for the development of domain-specific extensions (e.g. tuberculosis). This methodology offers many benefits. The core IDO ensures interoperability between the domain-specific extensions, while the modular approach allows for each module to be developed and maintained by researchers expert in that domain. The division of labor allows for rapid progress towards the needed set of ontologies, ensures the biological accuracy of the modules, and increases the likelihood of the broad adoption of the ontologies by the infectious disease research community.

IDO and its extensions are being built by relating terms from OBO Foundry ontologies using relations from the Foundry's relation ontology where possible, and

creating new terms and relations as needed. There are many benefits from building IDO and its extensions from OBO Foundry ontologies. In addition to the formalism underlying Foundry ontologies subsequently ensuring their support for sophisticated computation both within and between ontologies, building from Foundry ontologies means extensive use of existing ontology resources, both eliminating redundant effort and providing a significant head-start to ontology development. By building on OBO Foundry ontologies, IDO and its extensions are automatically interoperable with other ontologies that also build from Foundry ontologies as well as with the large information resources, such as UniProt and others mentioned above, that use Foundry ontologies for their wide base of existing annotations. Finally, as OBO Foundry ontologies, and in particular GO, are widely used, the use of Foundry ontologies in constructing IDO and its extensions improves the chances that IDO and its extensions will be accepted by the biological ontology and database communities.

To facilitate participation in the development and use of the infectious disease ontologies, we have established an Infectious Disease Ontology Consortium. In addition to development of the core IDO, consortium members are developing extensions for malaria, dengue fever, *Staphylococcus aureus* bacteremia, tuberculosis, brucellosis, influenza, HIV, and infective endocarditis. The Vaccine Ontology described earlier is also being developed as an IDO extension.

The IDO extensions are being tested for interoperability and for their use in a variety of computational applications. In response to these tests, the ontologies are refined for continued improvement. For example, the Vaccine Ontology is being applied to text-mining within the VOLIN project; the *Staphylococcus aureus* bacteremia ontology is being applied to the prediction of disease genes; the influenza ontology is being applied to influenza surveillance within the context of the Centers for Excellence in Influenza Research and Surveillance program; and the Dengue fever ontology is being utilized with the Dengue Decision Support System (DDSS).

The DDSS project (http://www.rams-aid.org) is the most developed and best demonstrates the long-term potential of computing with ontologies. The goal of the DDSS is to guide the implementation of locally appropriate Dengue and Dengue vector control programs. The DDSS makes use of the Mosquito Insecticide Resistance Ontology (http://www.obofoundry.org/), the Vector Surveillance Ontology, the Vector Control Ontology, and the Dengue ontology.

19.7 Conclusions

Here we have described the various types of vocabulary resources used to support informatics. We have emphasized the formal features of ontologies that enhance their utility for informatics applications relative to other types of vocabulary resources. We have discussed the current uses of vocabulary resources with a particular focus on the use of ontologies in the domain of infectious diseases. We have included a brief review of existing vocabulary resources relevant to the infec-

tious diseases domain, and have found that they are lacking in terms of their support of computational applications and translational medicine. We have described the Infectious Disease Ontology suite of ontologies and now invite all interested parties to participate in the development, testing, and refinement of these ontologies.

References

1. Rosse C, Mejino JLV (2003) A reference ontology for bioinformatics: The Foundational Model of Anatomy. J Biomed Informatics 36: 478-500.
2. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, et al. (2007) Advancing translational research with the Semantic Web. BMC Bioinformatics 8 Suppl 3: S2.
3. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol 25: 1251-1255.
4. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, et al. (2005) Relations in biomedical ontologies. Genome Biol 6: R46.
5. Bechhofer S, van Harmelen F, Hendler J, Horrocks I, McGuinness DL, et al. (2004) OWL Web Ontology Language Reference.
6. Ceusters W, Smith B, Kumar A, Dhaen C (2004) Mistakes in medical ontologies: Where do they come from and how can they be detected? . In: Pisanelli D, editor. Ontologies in Medicine. Amsterdam: IOS Press. pp. 145-164.
7. Ceusters W, Smith B, Goldberg L (2005) A terminological and ontological analysis of the NCI thesaurus. Methods of Information in Medicine 44: 498-507.
8. Smith B, Köhler J, Kumar A (2004) On the application of formal principles to life science data: A case study in the Gene Ontology. Data Integration in the Life Sciences (DILS) 2004: Springer. pp. 79-94.
9. Yu AC (2006) Methods in biomedical ontology. J Biomed Inform 39: 252-266.
10. Bodenreider O, Stevens R (2006) Bio-ontologies: current trends and future directions. Brief Bioinform 7: 256-274.
11. Cimino JJ, Zhu X (2006) The practical impact of ontologies on biomedical informatics. Yearb Med Inform: 124-135.
12. Coonan KM (2004) Medical informatics standards applicable to emergency department information systems: making sense of the jumble. Acad Emerg Med 11: 1198-1205.
13. Michael J, Mejino JL, Jr., Rosse C (2001) The role of definitions in biomedical concept representation. Proc AMIA Symp: 463-467.
14. Bodenreider O, Smith B, Kumar A, Burgun A (2004) Investigating subsumption in DL-Based terminologies: A case study in Snomed-CT,. KR-MED Proceedings 2004. pp. 12-20.
15. Guarino N. Some ontological principles for designing upper level lexical resources. In: Rubio A GN, Castro R, Tejada A, editors, editor; 1998; Granada, Spai. pp. 527-534.
16. Baader F (2007) The description logic handbook : theory, implementation, and applications. Cambridge ; New York: Cambridge University Press.
17. Hill DP, Blake JA, Richardson JE, Ringwald M (2002) Extension and integration of the gene ontology (GO): combining GO vocabularies with external vocabularies. Genome Res 12: 1982-1991.
18. Bodenreider O (2008) Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform: 67-79.
19. Natale DA, Arighi CN, Barker WC, Blake J, Chang TC, et al. (2007) Framework for a protein ontology. BMC Bioinformatics 8 Suppl 9: S1.
20. Bard J, Rhee SY, Ashburner M (2005) An ontology for cell types. Genome Biol 6: R21.

21. Rubin DL, Shah NH, Noy NF (2008) Biomedical ontologies: a functional perspective. Brief Bioinform 9: 75-90.
22. Rickard KL, Mejino JL, Jr., Martin RF, Agoncillo AV, Rosse C (2004) Problems and solutions with integrating terminologies into evolving knowledge bases. Medinfo 11: 420-424.
23. Zhang S, Bodenreider O (2005) Alignment of multiple ontologies of anatomy: deriving indirect mappings from direct mappings to a reference. AMIA Annu Symp Proc: 864-868.
24. Pisanelli D (2004) If ontology is the solution, what is the problem? In: Pisanelli D, editor. Ontologies in Medicine. Amsterdam: IOS Press. pp. 1-19.
25. Spasic I, Ananiadou S, McNaught J, Kumar A (2005) Text mining and ontologies in biomedicine: making sense of raw text. Brief Bioinform 6: 239-251.
26. Nelson SJ, Johnston D, Humphreys BL (2001) Relationships in Medical Subject Headings. In: Bean CA, Green R, editors. Relationships in the organization of knowledge. Dordrecht ; Boston Norwell, MA: Kluwer Academic Publishers; Sold and distributed in North, Central, and S. America by Kluwer Academic Publishers. pp. ix, 232 p.
27. Ide NC, Loane RF, Demner-Fushman D (2007) Essie: a concept-based search engine for structured biomedical text. J Am Med Inform Assoc 14: 253-263.
28. Muller HM, Kenny EE, Sternberg PW (2004) Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2: e309.
29. Doms A, Schroeder M (2005) GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res 33: W783-786.
30. Schonbach C, Nagashima T, Konagaya A (2004) Textmining in support of knowledge discovery for vaccine development. Methods 34: 488-495.
31. Rajapakse M, Kanagasabai R, Ang WT, Veeramani A, Schreiber MJ, et al. (2008) Ontology-centric integration and navigation of the dengue literature. J Biomed Inform 41: 806-815.
32. Blake JA, Eppig JT, Bult CJ, Kadin JA, Richardson JE (2006) The Mouse Genome Database (MGD): updates and enhancements. Nucleic Acids Res 34: D562-567.
33. Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, et al. (1997) Genetic and physical maps of *Saccharomyces cerevisiae*. Nature 387: 67-73.
34. Grumbling G, Strelets V (2006) FlyBase: anatomical data, images and queries. Nucleic Acids Res 34: D484-488.
35. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, et al. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. Nucleic Acids Res 32: D262-266.
36. Blake JA, Hill DP, Smith B. Gene Ontology annotations: What they mean and where they come from.; 2007 July 20, 2007; Vienna. pp. 79-82.
37. Medigue C, Moszer I (2007) Annotation, comparison and databases for hundreds of bacterial genomes. Res Microbiol 158: 724-736.
38. Whetzel PL, Parkinson H, Stoeckert CJ, Jr. (2006) Using ontologies to annotate microarray experiments. Methods Enzymol 411: 325-339.
39. Butte AJ, Chen R (2006) Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics. AMIA Annu Symp Proc: 106-110.
40. Chabalier J, Mosser J, Burgun A (2007) Integrating biological pathways in disease ontologies. Stud Health Technol Inform 129: 791-795.
41. Baranzini SE, Wang J, Gibson RA, Galwey N, Naegelin Y, et al. (2009) Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis. Hum Mol Genet 18: 767-778.
42. Valouev A, Johnson DS, Sundquist A, Medina C, Anton E, et al. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. Nat Methods 5: 829-834.
43. Kim CH, Lillehoj HS, Hong YH, Keeler CL, Jr. (2008) Comparison of transcriptional changes associated with E. acervulina and E. maxima infections using cDNA microarray technology. Dev Biol (Basel) 132: 121-130.

44. Grinde B, Gayorfar M, Rinaldo CH (2007) Impact of a polyomavirus (BKV) infection on mRNA expression in human endothelial cells. Virus Res 123: 86-94.
45. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, et al. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 4: R28.
46. Zhang B, Schmoyer D, Kirov S, Snoddy J (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics 5: 16.
47. Djebbari A, Karamycheva S, Howe E, Quackenbush J (2005) MeSHer: identifying biological concepts in microarray assays based on PubMed references and MeSH terms. Bioinformatics 21: 3324-3326.
48. Bresell A, Servenius B, Persson B (2006) Ontology annotation treebrowser : an interactive tool where the complementarity of medical subject headings and gene ontology improves the interpretation of gene lists. Appl Bioinformatics 5: 225-236.
49. Osborne JD, Zhu LJ, Lin SM, Kibbe WA (2007) Interpreting microarray results with gene ontology and MeSH. Methods Mol Biol 377: 223-242.
50. Ochs MF, Peterson AJ, Kossenkov A, Bidaut G (2007) Incorporation of gene ontology annotations to enhance microarray data analysis. Methods Mol Biol 377: 243-254.
51. Lee JA, Sinkovits RS, Mock D, Rab EL, Cai J, et al. (2006) Components of the antigen processing and presentation pathway revealed by gene expression microarray analysis following B cell antigen receptor (BCR) stimulation. BMC Bioinformatics 7: 237.
52. Brameier M, Wiuf C (2007) Co-clustering and visualization of gene expression data and gene ontology terms for Saccharomyces cerevisiae using self-organizing maps. J Biomed Inform 40: 160-173.
53. Huang D, Wei P, Pan W (2006) Combining gene annotations and gene expression data in model-based clustering: weighted method. OMICS 10: 28-39.
54. Liu J, Wang W, Yang J (2004) Gene Ontology friendly biclustering of expression profiles. Proc IEEE Comput Syst Bioinform Conf: 436-447.
55. Wolting C, McGlade CJ, Tritchler D (2006) Cluster analysis of protein array results via similarity of Gene Ontology annotation. BMC Bioinformatics 7: 338.
56. Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, et al. (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res 33: 1544-1552.
57. Detwiler LT, Chung E, Li A, Mejino JL, Jr., Agoncillo A, et al. (2004) A relation-centric query engine for the Foundational Model of Anatomy. Stud Health Technol Inform 107: 341-345.
58. Rubin DL, Dameron O, Bashir Y, Grossman D, Dev P, et al. (2006) Using ontologies linked with geometric models to reason about penetrating injuries. Artif Intell Med 37: 167-176.
59. Racunas SA, Shah NH, Albert I, Fedoroff NV (2004) HyBrow: a prototype system for computer-aided hypothesis evaluation. Bioinformatics 20 Suppl 1: i257-264.
60. Schurink CA, Lucas PJ, Hoepelman IM, Bonten MJ (2005) Computer-assisted decision support for the diagnosis and treatment of infectious diseases in intensive care units. Lancet Infect Dis 5: 305-312.
61. Thursky K (2006) Use of computerized decision support systems to improve antibiotic prescribing. Expert Rev Anti Infect Ther 4: 491-507.
62. Sintchenko V, Coiera E, Gilbert GL (2008) Decision support systems for antibiotic prescribing. Curr Opin Infect Dis 21: 573-579.
63. Pestotnik SL (2005) Expert clinical decision support systems to enhance antimicrobial stewardship programs: insights from the society of infectious diseases pharmacists. Pharmacotherapy 25: 1116-1125.
64. Coleman M, Sharp B, Seocharan I, Hemingway J (2006) Developing an evidence-based decision support system for rational insecticide choice in the control of African malaria vectors. J Med Entomol 43: 663-668.

65. Buckeridge DL (2007) Outbreak detection through automated surveillance: a review of the determinants of detection. J Biomed Inform 40: 370-379.
66. Buckeridge DL, Burkom H, Campbell M, Hogan WR, Moore AW (2005) Algorithms for rapid outbreak detection: a research synthesis. J Biomed Inform 38: 99-113.
67. Veenema TG, Toke J (2006) Early detection and surveillance for biopreparedness and emerging infectious diseases. Online J Issues Nurs 11: 3.
68. Cornet R, de Keizer N (2008) Forty years of SNOMED: a literature review. BMC Med Inform Decis Mak 8 Suppl 1: S2.
69. Tveit H, Mollestad T, Laegreid A (2004) The alignment of the Medical subject headings to the Gene Ontology and its application in Gene annotation. Lecture notes in computer science 3066: 798-804.
70. Ceusters W, Smith B, Kumar A, Dhaen C (2004) Ontology-based error detection in SNOMED-CT. MedInfo 11: 482-486.
71. Bodenreider O (2006) Lexical, terminological and ontological resources for biological text mining. In: S A, J M, editors. Text mining for biology and biomedicine: Artech House. pp. 43-66.
72. Kohler J, Munn K, Ruegg A, Skusa A, Smith B (2006) Quality control for terms and definitions in ontologies and taxonomies. BMC Bioinformatics 7: 212.
73. Smith B, Kumar A (2004) On controlled vocabularies in bioinformatics: A case study in the Gene Ontology. BIOSILICO: Drug Discovery Today 2: 246-252.

Glossary

Analysis workflow: The transformation of raw data into biological evidence by applying algorithms, tools and services in a certain order

Annotation: The routine process of assignment of functions to genes in a sequenced genome or the extraction of biological knowledge from raw nucleotide sequences

Antisense: Nucleic acid molecules that bind a complimentary strand of nucleic acid to modify gene expression

Assembly: construction of longer sequences, such as contigs or genomes, from shorter sequences, such as sequence reads with or without prior knowledge on the order of the reads or reference to a closely related sequence

Bayes' rule: A mathematical identity [Pr(x|y)=Pr(y|x) Pr(x)/Pr(y)] that allows one to swap variables in a conditional probability expression

Bioinformatics: The application of molecular biology as an information science, especially the use of computational tools and algorithms in genomics research

Biomarker: A biological characteristic which is objectively measured and evaluated as an indicator of normal or pathological processes or host responses to a therapeutic intervention

BLAST: (Basic logical alignment and search tool): A computer program for finding sequences in databases that have identity to a query sequence

Clade: A group of organisms that shares a common ancestor to the exclusion of the other considered taxa

Cladistics: A school of thought that emphasizes reconstructing evolutionary relationships solely through recognizing clades by a set of specific criteria for inference

Clone: Clone can be identified using molecular epidemiological methods. Strains belong to a clonal claster if they share at least five out of seven housekeeping genes according to multilocus sequence typing

Core genome: The set of genes found in all members of a single species

Data: Any and all complex data entities from observations, experiments, simulations, models and higher order assemblies, along with the associated documentation needed to describe and interpret them

Data integration: The process of combining disparate data and providing a unified view of these data

Data mining: Automatically searching large volumes of data for patterns or associations

Data warehouse: An information infrastructure that enables researchers and clinicians to access and analyse detailed data and trends. Created by collecting databases and linking them using common data elements

***De novo* gene prediction**: An approach to gene prediction in which the only inputs are genome sequences; no evidence derived from RNA is used

DNA sequencing: Biochemical methods for determining the order of the nucleotide bases, adenine, guanine, cytosine and thymine, in a DNA oligonucleotide

Electronic laboratory reporting (ELR): The automated reporting of notifiable disease data via a secure, electronic connection by laboratories to state and local health departments or public health authorities

Electronic medical record (EMR): Computer-based patient medical record

Epitope: The regions of an antigen that bind to antigen-specific membrane receptors on lymphocytes

Free text: Data that has no particular structure other than normal grammar; may show substantial variation between records

Functional genomics: Exploration of the function of genes and other parts of the genome

Genome: The complete set of genetic information in an organism. In bacteria, this includes the chromosome(s) and extrachromosomal genetic information, e.g., plasmids

Genome-level characters: Features of a genome or its products other than the linear sequences of nucleotides or amino acids that can be assessed for phylogenetic analysis

Genomics: The study of the entire genome of an organisms; structural genomics includes whole-genome sequencing, whereas functional genomics aims to determine the functions of all genes

Genotype: The entire genetic constitution of an organism or the genetic composition at a specific gene locus or set of loci

Grid: A fully distributed, dynamically reconfigurable, scalable and autonomous inftrastructure to provide location independent, secure and efficient access to a coordinated set of services encapsulating and virtualizing resources

Informed consent: A legal term referring to a situation where a person can be said to have given his or her consent based upon an appreciation and understanding of the facts and implications of an action

Health Level 7 (HL7): A health data interchange standard designed to facilitate the transfer of health data resident on different and disparate computer systems in a health care setting

Homoplasy: A pattern of character states that supports an alternative to the true, accepted or most parsimonious evolutionary tree that is generally caused by evolutionary changes

Horizontal gene transfer: Any process in which an organism transfers genetic material to another cell that is not its offspring. This process is in contrast to more common vertical gene transfer, which occurs when genetic information is passed from parent to offspring

Infectome: System of networks of interacting host and pathogen's genes, proteins and metabolites involved in a process of infection and disease

Intron: Portions of a gene between the coding exons that are also transcribed, but are enzymatically removed from the mRNA before its translation into a protein

Knowledge base: A repository for the knowledge used by a knowledge system

Knowledge based system: A computer system that represents and uses knowledge to carry out a task

Metagenomics: The high-throughput study of sequences from multiple genomes recovered from samples that contain mixed microbial populations

Metadata: Data about data; may be regarded as a subset of data which adds relevance and purpose to data and enables the identification of similar data in different data collections

Microbiome: Collective system of genomes of all microbial flora of the human

Middleware: A software stack composed of security, resource management, data access and other services and applications, users and resource providers to operate effectively

Network: Series of points or nodes interconnected by edges, edges can have direction or different weights

Next-generation sequencing: Novel approaches to DNA sequencing that dispense with the need to create libraries of clones sequences in bacteria and holds the promise of providing faster and cheaper sequencing

Ontology: The systematic description of a given phenomenon, which often includes a controlled vocabulary and relationships, captures nuances in meaning and enables knowledge sharing and reuse. Typically, ontology defines data entities, data attributes, relations and possible functions and operations

Pan-genome: The set of all genes found in members of a single species

Ontology: A formal description of set of entities within a body of knowledge and the relationships between those entities, used to reason about the entities. Usually is represented as hierarchical, and often richly interconnected, set of objects, concepts and other entities that embody knowledge about the field

Orthologous: Homologous genes in two or more organisms that are related only by lineage splitting and not by gene duplication

Parsing: A segmentation of a string of letters together with a labelling of the segments

Phenetics: Phylogenetic reconstruction based on measures of overall similarity

Quorum sensing: The communication and coordination of bacteria through signalling molecules

Single nucleotide polymorphism (SNP): Sites in the genome where individual organisms differ in their DNA sequence, often by a single base, usually with very low population frequencies

Standard vocabulary: Systems of names that are assigned to concepts or entities that can create order within databases

System biology: Integrative discipline that seeks to explain the properties and behavior of complex biological systems in terms of their components and their interactions

Systematized Nomenclature of Medicine (SNOMED): A standard vocabulary system for medical databases; contains more than 144,000 terms and is available in at least two languages. Developed by the College of American Pathologists

Terminal branch: The part of an evolutionary tree that lead only to the taxon considered (not internode branches)

Virulence factor: A protein or a gene that is required for a pathogen to cause disease

Universal genetic code: a misnomer based on an earlier, incorrect belief that all genomes share the same code for specifying amino acids from triplets of nucleotides

Whole-genome shortgun sequencing: An approach to determine the sequence of a genome in which the genome is broken into numerous small fragments. These fragments are then assembled en masse. The individual sequences are assembled into larger sequences (known as contigs) that correspond to substantial portions of the genome.