Published in final edited form as:
Yearb Med Inform. 2008 ; : 67–79.

Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support

O. Bodenreider
Lister Hill National Center for Biomedical Communications, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

Summary

Objectives—To provide typical examples of biomedical ontologies in action, emphasizing the role played by biomedical ontologies in knowledge management, data integration and decision support.

Methods—Biomedical ontologies selected for their practical impact are examined from a functional perspective. Examples of applications are taken from operational systems and the biomedical literature, with a bias towards recent journal articles.

Results—The ontologies under investigation in this survey include SNOMED CT, the Logical Observation Identifiers, Names, and Codes (LOINC), the Foundational Model of Anatomy, the Gene Ontology, RxNorm, the National Cancer Institute Thesaurus, the International Classification of Diseases, the Medical Subject Headings (MeSH) and the Unified Medical Language System (UMLS). The roles played by biomedical ontologies are classified into three major categories: knowledge management (indexing and retrieval of data and information, access to information, mapping among ontologies); data integration, exchange and semantic interoperability; and decision support and reasoning (data selection and aggregation, decision support, natural language processing applications, knowledge discovery).

Conclusions—Ontologies play an important role in biomedical research through a variety of applications. While ontologies are used primarily as a source of vocabulary for standardization and integration purposes, many applications also use them as a source of computable knowledge. Barriers to the use of ontologies in biomedical applications are discussed.

Keywords

Biomedical ontologies; knowledge management; data integration; decision support

1 Introduction

The need for standardizing biomedical vocabulary is not recent. As long ago as the 17th century, health authorities in London used a standard list of about 200 causes of death - later integrated into the International Classification of Diseases - to compile accurate health statistics known as the Bills of Mortality [1]. In addition to terms, scientists such as Linnaeus started formalizing the relations among biological entities, in order to represent and share their knowledge of the world [2].

The last decade has seen a marked increase in the number of artifacts created for representing biomedical entities, their terms and their relations, often referred to as vocabularies,

Correspondence to: Dr. Olivier Bodenreider, National Library of Medicine, 8600 Rockville Pike - MS 3841 (Bldg 38A, Rm B1N28U), Bethesda, MD 20894 - USA, Tel: +1 301 435 3246, Fax: +1 301 480 3035, E-mail: olivier@nlm.nih.gov.

terminologies and ontologies. As shown in Figure 1, the number of citations on ontologies and controlled vocabularies in the PubMed/MEDLINE database has grown by 600% to about 1200 per year[1]. While some authors have proposed definitions for these artifacts [3,4] and attempted to characterize the distinctions among them [5], in practice, these names are often used interchangeably. This phenomenon is reflected in part by the fact that 5–10% of the PubMed/MEDLINE citations indexed under the MeSH descriptor "Vocabulary, Controlled" also contain the word "ontology" (dark section of the histogram in Figure 1.) For the sake of simplicity, we henceforth refer to these various types of artifacts as ontologies. Another interesting trend in the past decade is the change in the relative importance of these ontologies, as measured by the number of mentions in PubMed/MEDLINE citations. As shown in Figure 2, the Gene Ontology (GO) has become the most cited ontology, with over 450 citations per year. In contrast, the footprint of Unif ied Medical Language System (UMLS) seems smaller now than ten years ago, although the number of citations has remained essentially constant throughout the decade.

A number of recent reviews have presented the major biomedical ontologies to various audiences, most often with an emphasis on their design and structural characteristics, mentioning their use only in passing [6–11]. Other reviews have presented the role played by several biomedical ontologies in specific applications, such as clinical decision support [12] and discovery applications [13], or in a specific domain, such as bioinformatics [14]. In contrast, one review provides a functional perspective on biomedical ontologies [15]. The interested user is referred to these reviews for more information about these ontologies.

In the present survey, we analyze some of the high-impact biomedical ontologies presented in [7] through the functional lens of [15], classifying their roles - somewhat arbitrarily - into three major categories: knowledge management, including the indexing and retrieval of data and information; data integration, exchange and semantic interoperability; and decision support and reasoning. The three categories, however, are not mutually exclusive and we examine the various roles played by each ontology. For example, LOINC is used as a source of standard vocabulary for retrieval purposes [16], for the integration and exchange of laboratory data [17,18], and for "reliable execution of decision logic in clinical decision support systems" [12]. More generally, reference ontologies are designed independently of any particular applications and expected to be useful in a variety of tasks [19,20].

The ontologies under investigation in this survey include SNOMED CT, a comprehensive concept system for healthcare [21–23]; the Logical Observation Identifiers, Names, and Codes (LOINC), a vocabulary for laboratory tests and clinical observations [24–26]; the Foundational Model of Anatomy (FMA), a domain ontology of structural human anatomy [20,27]; the Gene Ontology, a controlled vocabulary for the functional annotation of gene products across species [28–30]; RxNorm, a controlled vocabulary of normalized names and codes for clinical drugs [31–33]; the National Cancer Institute Thesaurus, a public domain terminology that provides broad coverage of the cancer domain [34–36]; the International Classification of Diseases, the 115-year-old medical terminology, now part of a family of health classifications [37,38]; the Medical Subject Headings (MeSH), a controlled vocabulary for the indexing and retrieval of the biomedical literature [39,40]; and the Unified Medical Language System (UMLS), a terminology integration system in which all the above ontologies are integrated (with the exception of the FMA, soon-to-be integrated) [41–43]. Some characteristics of these ontologies (based on information present in the UMLS) are shown in Table 1, including scope, number of entities, distribution of the number of terms per entity, and existence of a subsumption hierarchy.

[1]Citations indexed by the MeSH term "Diagnostic and Statistical Manual of Mental Disorders" (DSM) are excluded, because, in most articles, DSM is used not as a terminology, but as a source of diagnostic criteria for mental diseases.

Due to the large number of publications on the subject, this review will necessarily be superficial. Its objective is to provide not an exhaustive list of references, but rather examples, hope-fully typical, of biomedical ontologies in action. In order to reflect the state of the art, the selection of references is also somewhat biased towards recent journal articles.

2 Knowledge Management

One major role of biomedical ontologies is to serve as a source of vocabulary, i.e., a list of names for the entities represented in these ontologies. Strictly speaking, collecting names is the function of terminology, not ontology, and ontology languages such as OWL, the Web Ontology Language, treat names as labels or annotations [44]. In practice, however, most biomedical ontologies under investigation here (with the notable exception of LOINC) provide lists of names for the entities they accommodate, in addition to properties and relations for these entities. The terminological component of biomedical ontologies is an important resource for natural language processing systems [45] and supports knowledge management tasks such as annotation (or indexing) of resources, information retrieval, access to information and mapping across resources. However, the corpus of entity names present in biomedical ontologies covers only in part the lexicon of the domain (especially for languages other than English) and only forms the basis for managing term variation [46,47]. As shown in Table 1, the number of terms per entity varies largely among ontologies.

2.1 Annotating Data and Resources

Virtually every ontology in our survey serves as a source of vocabulary for the purpose of annotating data or indexing documents. Besides the prototypical examples of MeSH, used for indexing the biomedical literature [39], and the Gene Ontology, used for the functional annotation of gene products in several dozen model organisms [48], many other ontologies have also been used for annotation purposes.

Indexing is principally used in reference to the assignment of entries from a controlled vocabulary to documents, e.g., the biomedical literature. While the indexing of large collections such as PubMed/MEDLINE is still performed manually for the most part, automatic indexing systems have been developed (e.g., [49,50]). Although the goal is to assign MeSH descriptors, these systems often take advantage of the large set of terms and relations provided by the UMLS. Systems such as *GoPubMed* co-annotate the biomedical literature to both MeSH and the Gene Ontology [51].

The indexing of clinical documents is generally referred to as *coding* - and biomedical ontologies are sometimes called "code sets" [9]. The International Classification of Diseases (ICD) has been used for over a century for coding morbidity and mortality and, more recently, as a coding system for reimbursement purposes [52]. SNOMED CT is becoming adopted as a standard terminology for electronic health records by a growing number of countries [21,53] and has also been evaluated as a source of vocabulary for clinical research [54]. The UMLS Metathesaurus as a whole has also been used to support the coding of clinical documents, such as surgical pathology reports [55]. Like indexing, most coding is still performed manually. However, automatic techniques have been developed and evaluated (e.g., for ICD [56–58]), some of which exhibit high accuracy in limited domains.

In biology, the functional description of experimental data is usually referred to as *annotation*. Here again, (semi-)automatic methods for acquiring annotations from text have been investigated recently [59–63], but annotations are still most often the product of manual curation. Functional annotation is not limited the annotation of gene products to the Gene Ontology, but can be seen more generally as a "normalization" process applied to datasets, enabling further processing. For example, [64] used SNOMED CT and the NCI Thesaurus to

annotate tissue microarray data in the Stanford Tissue Microarray Database. Analogously, MeSH was used to annotate mentions of human diseases in the Gene Expression Omnibus, a public repository of gene expression data, in order to create gene-disease networks [65]. Related to the notion of indexing is that of *term recognition*, i.e., the process of automatically identifying mentions of entities of interest in text through natural language processing (NLP) techniques. A number of biomedical term recognition systems have been developed for the biomedical domain, exploiting the rich sources of vocabulary provided by biomedical ontologies [45]. UMLS-based systems include *MetaMap* [66] and *MetaPhrase* [67]. Developed more recently are systems such as *Termine* [68] and *Whatizit* [69], which cover genomics (e.g., gene and protein names) in addition to clinical medicine.

2.2 Accessing Biomedical Information

The main function of the indexing of large document collections such as MEDLINE is to support accurate retrieval, i.e., with high recall and high precision. With hierarchical controlled vocabularies such as MeSH [70] or the UMLS [71,72], queries can be expanded to the descendants of the original input term, in addition to being enriched with synonyms, which contributes to improving recall.

More generally, by providing lists of synonyms, relations among concepts, high-level categorization and co-occurrence information, the UMLS plays a major role in the retrieval of various types of documents, not only the biomedical literature in MEDLINE [73], but also medical textbooks available on the Internet [74], knowledge bases (e.g., of medical computational problems [75]) and medical images [76–78]. Because they provide terms in several languages, the UMLS and MeSH have also been used for cross-language information retrieval [79,80].

Several biomedical search engines exploit MeSH and the UMLS to provide access to the biomedical literature, including *SAPHIRE* [81], *Essie* [82] and *Textpresso* [83], as well as web resources for consumers (e.g., *WRAPIN* [84], *MedicoPort* [85]). Several specialized search engines have been created as well. Of particular interest are systems supporting evidence-based medicine and answering clinical questions. Such systems often exploit existing search engines (or term recognition systems) [86,87], and add specific constraints to the search [88,89].

Besides MeSH and the UMLS, other biomedical ontologies have been used for the retrieval of specific information. In addition to model organism databases, most microarray experiment databases can be searched by terms from the Gene Ontology [90], including ArrayExpress [91] and the Cancer gene expression database (CGED) [92]. The Stanford Tissue Microarray Database includes a NCI Thesaurus browser for searching disorders [93]; SNOMED CT is used in a system that helps patients find physicians with particular expertise [94]; and medical web resources are indexed with the International Classification of Diseases in the *HealthCyberMap* [95]. In the case of *Emily* [96], the ontology itself - here the Foundational Model of Anatomy - is used as the knowledge source for question answering purposes. Finally, some search engines such as *GoPubMed* organize the documents according to two ontologies and support searches on either ontology or both [51]. For example, a search on "COX-2" in *GoPubMed*, shows index terms from both MeSH (*Cyclooxygenase 2*) and the Gene Ontology (*cyclooxygenase pathway*).

The automatic classification of biomedical documents is also generally supported by ontologies. For example, the high-level categorization of UMLS Metathesaurus concepts with semantic types from the Semantic Network has been used for topic detection in medical texts [97], as well as document clustering [98]. The hierarchy of MeSH terms is used in [99] for the purpose of categorizing MEDLINE documents. Even when they do not exploit their structure, some document classification systems use the list of synonyms provided by ontologies such

as MeSH and the Gene Ontology to aggregate document features (i.e., using concepts as features instead of words) [100].

2.3 Mapping across Biomedical Ontologies

The availability of several dozen biomedical ontologies is both a blessing and a curse. On the one hand, users can choose from a variety of ontologies and select the artifact that best fits their purpose. On the other, resources annotated to different ontologies become more difficult to integrate, unless mappings are created among ontologies in order to identify equivalent concepts across ontologies. This issue was identified several decades ago and was in part the motivation for creating the Unified Medical Language System [42]. In effect, the UMLS Metathesaurus is a terminology integration system, in which synonymous terms from various terminologies are clustered into concepts, allowing for the seamless mapping between terms from different terminologies through a UMLS concept [41]. As mentioned earlier, these groupings of terms are often exploited for query expansion purposes in information retrieval. Some terminologies provide mapping information to other terminologies (e.g., SNOMED to ICD-9-CM), which, in some cases is recorded in the UMLS. Such features of the UMLS have been used for mapping between MeSH and SNOMED CT in the context of a digital library [101]. However, due to large differences in scope and granularity among vocabularies, direct mapping through synonymy and built-in mapping information fails to provide mapping for most concepts. In addition to these features, the hierarchical and associative relations among UMLS concepts have also been exploited for automatic mapping purposes (sometimes in combination with lexical mapping [102,103]), allowing concepts from one terminology to be mapped to more generic concepts in another terminology [104,105]. Other large ontologies such as SNOMED CT have also been exploited for mapping between clinical terminologies [106]. Analogously, the Foundational Model of Anatomy has been used as a reference for aligning anatomical ontologies [107]. Finally, medication reconciliation, i.e., the process of comparing a patient's medication orders to all of the medications that the patient has been taking, can be facilitated by the mapping among drug vocabularies realized in systems such as RxNorm and the UMLS [108], as is the exchange of medication information between federal agencies [109].

3 Data Integration, Exchange and Semantic Interoperability

Biomedical ontologies are often cited as an important element of semantic interoperability and information exchange in biomedicine, along with messaging standards and clinical information models [110]. For example, [111] notes the role of ontologies (called "standards") in the standardization of patients data to be exchanged across electronic health record (EHR) systems, contributing to connect "islands of data". Analogously, ontologies are key to clinical guideline models such as *SAGE* [112], where they standardize the representation of knowledge, thus facilitating maintenance, sharability and interoperability with EHR systems [112]. Ontologies also play a major role in the integration of heterogeneous data from disparate sources, which is a critical to translational research [113].

3.1 Information Exchange and Semantic Interoperability

The use of RxNorm, UMLS, and SNOMED CT is reported in [114] as part of a mediation strategy to exchange medication data between the Veterans Affairs (VA) and the Department of Defense (DoD) clinical information systems. LOINC is used widely in the exchange of laboratory data [18,115], often in conjunction with HL7 [116].

Semantic interoperability projects such as *BRIDG*, *CDA* and *caCORE* also rely on ontologies, although indirectly in most cases. The *BRIDG* model, developed by the Biomedical Research Integrated Domain Group, is an information model designed to "support practical application

and data interchange" for clinical research [117]. Semantic interoperability between clinical trials information systems is supported in *BRIDG* through semantic harmonization. Although *BRIDG* stopped short of binding the information model to specific ontologies, its developers acknowledge the role ontologies in semantic interoperability. (Methods for binding clinical terminologies to information models are presented in [118], and [119] has investigated the mapping of the Outcome and Assessment Data Set (OASIS-B1) to LOINC and other terminologies).

The HL7 Clinical Document Architecture, Release 2 (*CDA R2*) model is "richly expressive, enabling the formal representation of clinical statements", including clinical observations, medication administrations, and adverse events [120]. *CDA R2* associates the HL7 Reference Information Model with terminologies such as LOINC, SNOMED CT and RxNorm for representing the semantics of a clinical document.

The Common Ontologic Representation Environment (*caCORE*) is a model-driven infrastructure developed to support an interoperable biomedical information system for cancer research [121]. Ontologies, including the NCI Thesaurus [35], represent an important element of this infrastructure.

3.2 Information and Data Integration

Ontologies support data integration in two different ways, corresponding to two different approaches to data integration: warehousing and mediation [122]. One the one hand, by providing a controlled vocabulary in a given domain, ontology support the standardization required from *warehousing approaches* to data integration, in which the sources to be integrated are transformed into a common format and converted to a common vocabulary. For example, the integration of model organism databases is facilitated by the existence of the Gene Ontology, used - natively or after conversion - for the functional annotation of gene products in many species [48]. Analogously, the integration of data from microarray experiments benefits from the standardization of their description with ontologies [123].

On the other hand, *mediation-based approaches* use ontologies for defining a global schema (in reference to which queries are made) and mapping between the global schema and local schemas (the schemas of the sources to be integrated). *TAMBIS* [124], the *BioMediator* [125] and *OntoFusion* [126] provide examples of such systems. The UMLS is used (along with the Gene Ontology) for the creation of the global schema in *OntoFusion*. A similar approach, also based on the UMLS, is used in *ARIANE* [127], a system that provides access to heterogeneous medical databases.

More generally, ontologies facilitate the integration of datasets, often by providing a common reference for biomedical entities in several datasets. For example, LOINC has been used for integrating laboratory data with adverse events [128], the Foundational Model of Anatomy for the integration of genomic information sources [129], and SNOMED CT for the integration of disease and pathway information [130].

4 Decision Support and Reasoning

Ontologies represent domain knowledge in computable, reusable form [131]. Simple ontologies (e.g., limited to subsumption hierarchies) are useful for data aggregation and clustering. Rich ontologies comprise large networks of associative relations among the entities of a given domain. Such ontologies provide domain knowledge to applications and support the interpretation of relations identified in datasets through data mining processes based on linguistic or statistical techniques. Five broad kinds of applications of ontologies are discussed

next: data selection, data aggregation, decision support, natural language processing, and knowledge discovery.

4.1 Data Selection

Many clinical and epidemiological research studies involve the creation of groups (from an independent variable) whose characteristics (dependent variables) are examined for differences (e.g., survival rate at five years in breast cancer patients). By providing an abstraction of some domain, ontologies can help define groups from a high level value for the independent variable (e.g., breast cancer), instead of listing all possible values (e.g., cancer of upper-inner quadrant of breast, of lower-outer quadrant, etc.). The International Classification of Diseases is used pervasively for selecting groups of patients in association with a high-level disease category. For example, in a study of emergency department visits for supraventricular tachycardia (SVT), the selection of cases of SVT was based on the descendants of 2 high-level ICD codes. Analogously, [132] calculated survival risk ratios for trauma patients for various groups of hierarchically-defined diagnostic categories in ICD and [133] used high-level ICD codes in a study of stroke hospitalization over time. Many other ontologies are used for data selection purposes, including SNOMED CT, used for querying clinical data warehouses [134]. A hierarchical structure was added to LOINC in order to facilitate public health reporting [135].

4.2 Data Aggregation

In addition to data selection, ontologies are used for identifying the characteristics of groups obtained through various methods (e.g., the characteristics of patients in a group of long-term cancer survivors). Here again, ontologies support the aggregation of characteristics and ICD is often used for aggregating diagnoses. For example, in a study of the evolution over time of discharge diagnoses in emergency departments, the major categories of diagnoses investigated correspond to the top-level categories in ICD-10 [136]. (The accuracy of such studies, which depends on the quality of the coding and the type of study, is discussed in [137,138]).

In biology, microarray technologies for measuring gene expression typically identify groups of genes up-and down-regulated under certain circumstances [139]. The simultaneous activity (or inactivity) of genes in these groups represents only one clue into their participation in biological activities and such groups of genes generally require further characterization, especially through functional annotations [140]. Some fifty tools have been developed to date for the characterization of gene sets, exploiting Gene Ontology annotations [141] and other resources (e.g., *Onto-Express* [142] and *GoMiner* [143]). Some tools specifically take advantage of the hierarchical organization of terms in the Gene Ontology (e.g., [144]). Several tools also use MeSH descriptors for characterizing sets of genes [145], sometimes in combination with Gene Ontology terms [146,147]. The functional characterization of gene expression signatures is used widely. A search combining "gene ontology" and "gene expression" in PubMed/MEDLINE yields over 800 citations. A recent trend in gene expression profiling is co-clustering, i.e., the use of functional annotations not for the *post hoc* characterization of gene sets, but as part of the clustering process itself [148–150]. One limitation of data aggregation based on hierarchies is the heterogeneous density of terms throughout the ontology (i.e., some branches are more richly developed than others). Semantic similarity metrics based on information content have been developed to address this issue and successfully applied to the Gene Ontology [151]. These metrics provide a new approach to clustering genes [152].

4.3 Decision Support

Clinical decision support systems generally benefit from ontologies in two principal ways. First, as mentioned earlier, ontologies provide a standard vocabulary for biomedical entities, helping standardize and integrate data sources [12]. For example, a system for drug allergies

must be able to resolve drug names into standard codes and map between drug coding systems and the allergy knowledge base. Second, ontologies are a source of computable domain knowledge that can be exploited for decision support purposes, often in combination with business rules [153,154]. For example, in an alert system for drug allergies, allergy to betalactams can be represented efficiently if the system can access a classification of drugs (as opposed to direct links to specific drugs). The interested reader is referred to [155] for a discussion about the role of ontologies in specific clinical decision systems. Issues discussed earlier about knowledge management support for evidence-based medicine (2.2) and the role of ontologies in clinical guidelines (3.1) are also relevant to clinical decision support.

Besides clinical decision support, ontologies support reasoning in applications. The Foundational Model of Anatomy (FMA) was used as a source of anatomical knowledge for reasoning about penetrating injuries, more exactly for predicting the consequences of penetrating injury [156]. In this application, knowledge about spatial relations between the path of injury and vital organs is provided by the FMA. The availability of the NCI Thesaurus in OWL (Web Ontology Language) format makes it amenable to automatic processing by reasoners developed for OWL, enabling consistency checking and automatic classification. Leveraging such automatic reasoning services, [157] developed an automatic grading system for gliomas. Ontologies sometimes participate indirectly in reasoning processes. For example, [158] emphasizes the role of the Gene Ontology in the extraction of information required for creating an ontology of phosphatases. This ontology was subsequently used for reasoning about phosphatases. Although isolated, these examples illustrate the potential benefit of ontologies for decision support and reasoning.

4.4 Natural Language Processing Applications

As mentioned earlier, Natural Language Processing (NLP) techniques support term recognition, exploiting the vocabulary provided by biomedical ontologies. Ontologies also provide the domain knowledge necessary for advanced NLP applications, including information extraction for a specific task, relation extraction, document summarization, question answering, literature-based discovery, and more generally, text mining [45].

While term recognition systems merely identify entities in text, advanced systems identify specialized facts - sometimes on the basis of information provided by term recognition systems - used to guide specific applications (e.g., mentions of smoking in patient records, used for selecting cases [159]; medical problems from patient records, used to maintain problem lists [160]; respiratory findings from emergency department reports, for biosurveillance purposes [161]). Systems such as *BioCaster* [162] and *EpiSpider* [163] apply term recognition techniques to health news feeds and integrate the extracted information with other resources (including ontologies), creating what is known as "mashups". These resources can help track cases of, say, avian influenza and support biosurveillance and public health.

In addition to entity recognition, some systems extract relations (i.e., facts asserted in text), thus "interpreting" the text. Example of such systems exploiting the UMLS for processing clinical text and the biomedical literature include *SemRep* [164], *(Bio)MedLEE* [165,166] and commercial systems such as *Tessi* [167]. More specifically, *SemRep* draws on *MetaMap* for identifying entities in text and relies on the UMLS Semantic Network as its source of domain knowledge for the interpretation of the semantic predication it extracts [164].

The UMLS has been used in advanced NLP applications including question-answering systems [168–171] and the summarization of medical documents [172–174]. More generally, NLP techniques have evolved to support the high-throughput processing of the biomedical literature [175], in a similar fashion to the high-throughput processing of genomic data enabled by sequence alignment techniques. Massive amounts of data such as the MEDLINE database are

now routinely exploited, often in combination with ontologies, for hypothesis generation and knowledge discovery purposes. Literature-based discovery systems take advantage of the UMLS and MeSH as sources of knowledge, which are combined with the knowledge extracted from text to support the discovery process [176–179].

4.5 Knowledge Discovery

By supporting the high-throughput processing of biological and clinical data, ontologies are a component of the data-driven approach to biomedical research, synergistic with the traditional hypothesis-driven approach [180]. Moreover, data mining often operates on datasets resulting from the integration of heterogeneous resources, also supported by ontologies [181].

Because of the availability of datasets coded with the International Classification of Diseases (ICD), clinical data exploration often involves the mining of ICD codes, along with, for example, geographic data [182] or meteorological data [183]. The availability of large volumes of data makes it possible to detect rare events, such as adverse reactions to drugs (e.g., diabetic ketoacidosis [184] and hepatic toxicity [185]). In biology, the functional annotations of gene products from multiple model organisms to the Gene Ontology represent an important knowledge source, often mined in combination with sequence similarity [186,187], gene expression data [188,189], or both [190]. Predicting the molecular function or subcellular localization of uncharacterized genes is an active field of research. While most methods exploit the annotations of related gene products to the Gene Ontology, some also take advantage of the hierarchical structure of the Gene Ontology [191].

Finally, as mentioned earlier, ontologies have been used for identifying relations between genotype and phenotype, both for the vocabulary they provide [65,166], and for the relations among entities asserted in these ontologies [192]. Ontologies have also been used for creating and interpreting gene networks [193,194], as well as drug-target networks [195].

5 Discussion

Ontologies have become important resources for biomedical research and researchers have come to rely on ontologies such as the International Classification of Diseases and the Gene Ontology in a large variety of applications, taking their existence for granted. There are still barriers, however, to the use of ontologies in biomedical applications, including availability, discoverability, the formalisms used for their representation, integration and quality.

Availability

A large number of ontologies are freely available, including LOINC, the Foundational Model of Anatomy, the Gene Ontology, the NCI Thesaurus and MeSH. Because some of the ontologies integrated in the UMLS are subject to intellectual property restrictions, however, its users must sign a license agreement to get access to the UMLS content. RxNorm follows the same model, although the part of its content owned by the National Library of Medicine is made freely available through a browser and an application programming interface [196]. Finally, the availability of SNOMED CT to users depends on whether their country is a member of the International Health Terminology Standards Development Organization[2] (IHTSDO). Being freely available is one of the requirements for ontologies to be included in the Open Biomedical Ontology repository [197], as it is also expected from the ontologies used in the Semantic Web.

[2]http://www.ihtsdo.org/

Discoverability

With over 140 ontologies, the UMLS is the largest repository of biomedical ontologies (accessible through the *Knowledge Source Server* [198]), but its coverage is some-what biased towards healthcare applications. The National Center for Biomedical Ontology's *BioPortal* [199] provides access to about ninety ontologies, including those from the Open Biomedical Ontology (OBO) collection, with a bent for biological ontologies. Ontologies such as the Gene Ontology and the NCI Thesaurus are present in both collections. While useful, these two resources do not completely compensate for the lack of a registry allowing users to discover biomedical ontologies corresponding to their needs, which leads to both the underutilization of existing but unpublicized resources, and the development of roughly similar artifacts by independent groups.

Formalism

The ontologies integrated in the UMLS are all converted to the so-called RRF format, regardless of their native representation formalism. RRF supports the representation of both the terms and relations natively present in these ontologies and the concept-oriented view superimposed by the UMLS. On the other hand, most ontologies available on *BioPortal* are represented in OBO format, the others being in frame-based Protégé or OWL format. Despite the availability of converters between OBO and OWL (e.g., [200]) and terminology servers supporting multiple formats, such as LexGrid [201], the multiplicity of formats remains an impediment to the use of biomedical ontologies.

Integration

There are basically two approaches to integrating ontologies. On the one hand, the UMLS realizes the *post hoc* integration of ontologies, from the bottom up, without interfering with the development process or governance of the ontologies being integrated. On the other, the OBO Foundry promotes a model of coordinated development of ontologies [202]. Both approaches are useful to data integration. By integrating existing ontologies "as is", the former only links them to the extent possible (as they might show limited compatibility), but has the advantage of facilitating the integration of the vast datasets annotated to these ontologies (e.g., ICD, MeSH). On the other hand, the top-down approach of the OBO Foundry model ensures consistency *ab initio*, but is virtually impossible to apply retrospectively to large, widely used, legacy ontologies.

Quality

Intuitively, the poor quality of some ontologies might result in inaccuracies in the applications they support. In practice, assessing the quality of biomedical ontologies with intrinsic criteria is difficult and might be futile if disconnected from practical applications [203]. On the one hand, the evaluation of quality can be seen as the responsibility of users, who can share their experience with other users by commenting on the usefulness of a given ontology (or part thereof) from the perspective of their application. This constitutes a democratic approach to quality evaluation. For others, the determination of the quality of ontologies should be based solely on science and left to an oligarchy of specialists [204]. While the accuracy of statements in ontologies is important, other factors such as installed base (how many users does it have?) and governance (who makes decisions about development and maintenance?) also need to be taken into account when selecting an ontology for a given application.

6 Conclusions

Ontologies play an important role in biomedical research through a variety of applications. They provide the controlled vocabulary required for the annotation of biological datasets, the

biomedical literature and patient records, facilitating the retrieval of and, more generally, access to information. Such standardization also facilitates the exchange of information and contributes to semantic interoperability among systems. By providing a representation of a domain, ontologies are also used in the mediation approach to integrating datasets. Finally, many applications use ontologies as a source of computable domain knowledge, including natural language processing applications and decision support systems. Ontologies are also critical to hypothesis generation and knowledge discovery in a data-driven approach to biomedical research.

Acknowledgments

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. Gersenovic M. The ICD family of classifications. Methods Inf Med 1995;34(1–2):172–175. [PubMed: 9082128]
2. McCray AT. Conceptualizing the world: lessons from history. J Biomed Inform 2006;39(3):267–273. [PubMed: 16243005]
3. Chute CG. Clinical classification and terminology: some history and current observations. J Am Med Inform Assoc 2000;7(3):298–303. [PubMed: 10833167]
4. Smith, B.; Kusnierczyk, W.; Schober, D.; Ceusters, W. Towards a reference terminology for ontology research and development in the biomedical domain. In: Bodenreider, O., editor. Proceedings of the Second International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2006); 2006. p. 57-65.
5. Ontology, Taxonomy, Folksonomy: Understanding the Distinctions. http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2007_Communique
6. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. Brief Bioinform 2006;7(3):256–274. [PubMed: 16899495]
7. Cimino JJ, Zhu X. The practical impact of ontologies on biomedical informatics. Yearb Med Inform 2006:124–135. [PubMed: 17051306]
8. Coonan KM. Medical informatics standards applicable to emergency department information systems: making sense of the jumble. Acad Emerg Med 2004;11(11):1198–1205. [PubMed: 15528585]
9. Giannangelo, K., editor. Healthcare code sets, clinical terminologies, and classification systems. Chicago, Ill: American Health Information Management Association; 2006.
10. Yu AC. Methods in biomedical ontology. J Biomed Inform 2006;39(3):252–266. [PubMed: 16387553]
11. Bodenreider, O.; Burgun, A. Biomedical ontologies. In: Chen, H.; Fuller, S.; Hersh, WR.; Friedman, C., editors. Medical informatics: Advances in knowledge management and data mining in biomedicine. New York: Springer-Verlag; 2005. p. 211-236.
12. Huff, SM. Ontologies, vocabularies, and data models. In: Greenes, RA., editor. Clinical desision support: The road ahead. Amsterdam: Academic Press; 2007. p. 307-324.
13. Lussier, YA.; Bodenreider, O. Clinical ontologies for discovery applications. In: Baker, CJO.; Cheung, KH., editors. Semantic Web: Revolutionizing knowledge discovery in the life sciences. New York: Springer; 2007. p. 101-119.
14. Stevens, R.; Wroe, C.; Lord, P.; Goble, C. Ontologies in bioinformatics. In: Staab, S.; Studer, R., editors. Handbook on ontologies. Berlin ; New York: Springer; 2004. p. 635-657.
15. Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. Brief Bioinform 2008;9(1):75–90. [PubMed: 18077472]
16. Hsu, C.; Goldberg, HS. Knowledge-mediated retrieval of laboratory observations; Proc AMIA Symp; 1999. p. 809-813.

17. Baorto, DM.; Cimino, JJ.; Parvin, CA.; Kahn, MG. Using Logical Observation Identifier Names and Codes (LOINC) to exchange laboratory data among three academic hospitals; Proc AMIA Annu Fall Symp; 1997. p. 96-100.

18. Baorto DM, Cimino JJ, Parvin CA, Kahn MG. Combining laboratory data sets from multiple institutions using the logical observation identifier. Int J Med Inform 1998;51(1):29–37. [PubMed: 9749897]

19. Burgun A. Desiderata for domain reference ontologies in biomedicine. J Biomed Inform 2006;39(3): 307–313. [PubMed: 16266830]

20. Rosse C, Mejino JL Jr. A reference ontology for biomedical informatics: the Foundational Model of Anatomy. J Biomed Inform 2003;36(6):478–500. [PubMed: 14759820]

21. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. Stud Health Technol Inform 2006;121:279–290. [PubMed: 17095826]

22. Wang, AY.; Sable, JH.; Spackman, KA. The SNOMED clinical terms development process: refinement and analysis of content; Proc AMIA Symp; 2002. p. 845-849.

23. SNOMED CT. (Systematized Nomenclature of Medicine-Clinical Terms). http://www.ihtsdo.org/our-standards/

24. Huff SM, Rocha RA, McDonald CJ, De Moor GJ, Fiers T, Bidgood WD Jr, et al. Development of the Logical Observation Identifiers Names and Codes (LOINC) vocabulary. J Am Med Inform Assoc 1998;5(3):276–292. [PubMed: 9609498]

25. McDonald CJ, Huff SM, Suico JG, Hill G, Leavelle D, Aller R, et al. LOINC, a universal standard for identifying laboratory observations: a 5-year update. Clin Chem 2003;49(4):624–633. [PubMed: 12651816]

26. Logical Observation Identifiers Names and Codes (LOINC). www.regenstrief.org/loinc/

27. Foundational Model of Anatomy (FMA). http://sig.biostr.washington.edu/projects/fm/

28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000 May;25(1):25–29. [PubMed: 10802651]

29. Lomax J. Get ready to GO! A biologist's guide to the Gene Ontology. Brief Bioinform 2005;6(3): 298–304. [PubMed: 16212777]

30. Gene Ontology. http://www.geneontology.org/

31. Liu S, Wei M, Moore R, Ganesan VAGV, Nelson SANS. RxNorm: prescription for electronic drug information exchange. IT Professional 2005;7(5):17–23.

32. Nelson, SJ.; Brown, SH.; Erlbaum, MS.; Olson, N.; Powell, T.; Carlsen, B.; Carter, J.; Tuttle, MS.; Hole, WT. A semantic normal form for clinical drugs in the UMLS: early experiences with the VANDF; Proc AMIA Symp; 2002. p. 557-561.

33. RxNorm. http://www.nlm.nih.gov/research/umls/rxnorm/

34. de Coronado S, Haber MW, Sioutos N, Tuttle MS, Wright LW. NCI Thesaurus: using science-based terminology to integrate cancer research results. Stud Health Technol Inform 2004;107(Pt 1):33–37. [PubMed: 15360769]

35. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform 2007;40(1): 30–43. [PubMed: 16697710]

36. NCI Thesaurus. http://www.nci.nih.gov/cancerinfo/terminologyresources

37. Jakob R, Ustun B, Madden R, Sykes C. The WHO Family of International Classifications. Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz 2007;50(7):924–931. [PubMed: 17581728]

38. International Classification of Diseases (ICD). http://www.who.int/classifications/icd/en/

39. Nelson, SJ.; Johnston, D.; Humphreys, BL. Relationships in Medical Subject Headings. In: Bean, CA.; Green, R., editors. Relationships in the organization of knowledge. New York: Kluwer Academic Publishers; 2001. p. 171-184.

40. Medical Subject Headings (MeSH). http://www.nlm.nih.gov/mesh/

41. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32(Database issue):D267–D270. [PubMed: 14681409]
42. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med 1993;32(4):281–291. [PubMed: 8412823]
43. Unified Medical Language System (UMLS). http://www.nlm.nih.gov/research/umls/
44. OWL Web Ontology Language Reference. http://www.w3.org/TR/owl-ref/
45. Bodenreider, O. Lexical, terminological and ontological resources for biological text mining. In: Ananiadou, S.; McNaught, J., editors. Text mining for biology and biomedicine. Boston: Artech House; 2006. p. 43-66.
46. Aronson, AR. The effect of textual variation on concept based information retrieval; Proc AMIA Annu Fall Symp; 1996. p. 373-377.
47. Kolarik C, Hofmann-Apitius M, Zimmermann M, Fluck J. Identification of new drug classification terms in textual resources. Bioinformatics 2007;23(13):i264–i272. [PubMed: 17646305]
48. Blake JA, Bult CJ. Beyond the data deluge: data integration and bio-ontologies. J Biomed Inform 2006;39(3):314–320. [PubMed: 16564748]
49. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Stud Health Technol Inform 2004;107(Pt 1):268–272. [PubMed: 15360816]
50. Zhang, D.; Roderer, NK.; Huang, G.; Zhao, X. Developing a UMLS-based indexing tool for health science repository system; AMIA Annu Symp Proc; 2006. p. 1157
51. Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res 2005;33(Web Server issue):W783–W786. [PubMed: 15980585]
52. Alexander S, Conner T, Slaughter T. Overview of inpatient coding. Am J Health Syst Pharm 2003;60:S11–S14. [PubMed: 14619128]
53. Daniel-Le Bozec C, Steichen O, Dart T, Jaulent MC. The role of local terminologies in electronic health records. The HEGP experience. Stud Health Technol Inform 2007;129(Pt 1):780–784. [PubMed: 17911823]
54. Richesson RL, Andrews JE, Krischer JP. Use of SNOMED CT to represent clinical research data: a semantic characterization of data items on case report forms in vasculitis research. J Am Med Inform Assoc 2006;13(5):536–546. [PubMed: 16799121]
55. Liu, K.; Mitchell, KJ.; Chapman, WW.; Crowley, RS. Automating tissue bank annotation from pathology reports - comparison to a gold standard expert annotation set; AMIA Annu Symp Proc; 2005. p. 460-464.
56. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 2004;11(5):392–402. [PubMed: 15187068]
57. Pakhomov SV, Buntrock JD, Chute CG. Automating the assignment of diagnosis codes to patient encounters using example-based and machine learning techniques. J Am Med Inform Assoc 2006;13 (5):516–525. [PubMed: 16799125]
58. Pestian, JP.; Brew, C.; Matykiewicz, P.; Hovermale, D.; Johnson, N.; Cohen, KB., et al. Biological, translational, and clinical language processing; 2007 2007-06. Prague: Czech Republic: Association for Computational Linguistics; 2007. A shared task involving multi-label classification of clinical free text; p. 97-104.
59. Blaschke C, Leon EA, Krallinger M, Valencia A. Evaluation of BioCreAtIvE assessment of task 2. BMC Bioinformatics 2005;6:S16. [PubMed: 15960828]
60. Couto FM, Silva MJ, Lee V, Dimmer E, Camon E, Apweiler R, Kirsch H, Rebholz-Schuhmann D. GOAnnotator: linking protein GO annotations to evidence text. J Biomed Discov Collab 2006 Dec;: 1–19. [PubMed: 17181854]
61. Crangle, CE.; Zbyslaw, A. Identifying gene ontology concepts in natural-language text; Conf Proc IEEE Eng Med Biol Soc; 2004. p. 2821-2823.
62. Daraselia N, Yuryev A, Egorov S, Mazo I, Ispolatov I. Automatic extraction of gene ontology annotation and its correlation with clusters in protein networks. BMC Bioinformatics 2007;8:243. [PubMed: 17620146]
63. Srinivasan P, Qiu XY. GO for gene documents. BMC Bioinformatics 2007;8:S3. [PubMed: 18047704]

64. Shah, NH.; Rubin, DL.; Supekar, KS.; Musen, MA. Ontology-based annotation and query of tissue microar ray data; AMIA Annu Symp Proc; 2006. p. 709-713.

65. Butte, AJ.; Chen, R. Finding disease-related genomic experiments within an international repository: first steps in translational bioinformatics; AMIA Annu Symp Proc; 2006. p. 106-110.

66. Aronson, AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program; Proc AMIA Symp; 2001. p. 17-21.

67. Tuttle MS, Olson NE, Keck KD, Cole WG, Erlbaum MS, Sherertz DD, et al. Metaphrase: an aid to the clinical conceptualization and formalization of patient problems in healthcare enterprises. Methods Inf Med 1998;37(4–5):373–383. [PubMed: 9865035]

68. Termine. http://www.nactem.ac.uk/software/termine/

69. Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A. Text processing through web services: calling Whatizit. Bioinformatics 2008;24:296–298. [PubMed: 18006544]

70. Srinivasan P. Retrieval feedback in MEDLINE. J Am Med Inform Assoc 1996;3(2):157–167. [PubMed: 8653452]

71. Aronson, AR.; Rindflesch, TC. Query expansion using the UMLS Metathesaurus; Proc AMIA Annu Fall Symp; 1997. p. 485-489.

72. Hersh, W.; Price, S.; Donohoe, L. Assessing thesaurus-based query expansion using the UMLS Metathesaurus; Proc AMIA Symp; 2000. p. 344-348.

73. Hersh, W.; Hickam, DH.; Haynes, RB.; McKibbon, KA. Evaluation of SAPHIRE: an automated approach to indexing and retrieving medical literature; Proc Annu Symp Comput Appl Med Care; 1991. p. 808-812.

74. Brandt C, Nadkarni P. Web-based UMLS concept retrieval by automatic text scanning: a comparison of two methods. Comput Methods Programs Biomed 2001;64(1):37–43. [PubMed: 11084231]

75. Bratsas, C.; Koutkias, V.; Kaimakamis, E.; Bamidis, P.; Maglaveras, N. Ontology-based vector space model and fuzzy query expansion to retrieve knowledge on medical computational problem solutions; Conf Proc IEEE Eng Med Biol Soc; 2007. p. 3794-3797.

76. Lowe HJ, Antipov I, Hersh W, Smith CA, Mailhot M. Automated semantic indexing of imaging reports to support retrieval of medical images in the multimedia electronic medical record. Methods Inf Med 1999;38(4–5):303–307. [PubMed: 10805018]

77. Ruiz, ME. Combining image features, case descriptions and UMLS concepts to improve retrieval of medical images; AMIA Annu Symp Proc; 2006. p. 674-678.

78. Yu H, Lee M. Accessing bioscience images from abstract sentences. Bioinformatics 2006;22 (14):e547–e556. [PubMed: 16873519]

79. Hersh, WR.; Donohoe, LC. SAPHIRE International: a tool for cross-language information retrieval; Proc AMIA Symp; 1998. p. 673-677.

80. Liu, F.; Ackerman, M.; Fontelo, P. BabelMeSH: development of a cross-language tool for MEDLINE/ PubMed; AMIA Annu Symp Proc; 2006. p. 1012

81. Hersh, W.; Leone, TJ. The SAPHIRE server: a new algorithm and implementation; Proc Annu Symp Comput Appl Med Care; 1995. p. 858-862.

82. Ide NC, Loane RF, Demner-Fushman D. Essie: a concept-based search engine for structured biomedical ext. J Am Med Inform Assoc 2007;14(3):253–263. [PubMed: 17329729]

83. Muller HM, Kenny EE, Sternberg PW. Textpresso: an ontology-based information retrieval and extraction system for biological literature. PLoS Biol 2004;2(11):e309. [PubMed: 15383839]

84. Gaudinat A, Ruch P, Joubert M, Uziel P, Strauss A, Thonnet M, et al. Health search engine with edocument analysis for reliable search results. Int J Med Inform 2006;75(1):73–85. [PubMed: 16377235]

85. Can AB, Baykal N. MedicoPort: a medical search engine for all. Comput Methods Programs Biomed 2007;86(1):73–86. [PubMed: 17321002]

86. Fontelo P, Liu F, Leon S, Anne A, Ackerman M. PICO Linguist and BabelMeSH: development and partial evaluation of evidence-based multilanguage search tools for MEDLINE/PubMed. Stud Health Technol Inform 2007;129(Pt 1):817–821. [PubMed: 17911830]

87. Sneiderman CA, Demner-Fushman D, Fiszman M, Ide NC, Rindflesch TC. Knowledge-based methods to help clinicians find answers in MEDLINE. J Am Med Inform Assoc 2007;14(6):772–780. [PubMed: 17712086]

88. Haynes RB, McKibbon KA, Wilczynski NL, Walter SD, Werre SR. Optimal search strategies for retrieving scientifically strong studies of treatment from Medline: analytical survey. BMJ 2005;330 (7501):1179. [PubMed: 15894554]

89. Huang, X.; Lin, J.; Demner-Fushman, D. Evaluation of PICO as a knowledge representation for clinical questions; AMIA Annu Symp Proc; 2006. p. 359-363.

90. Whetzel PL, Parkinson H, Stoeckert CJ Jr. Using ontologies to annotate microarray experiments. Methods Enzymol 2006;411:325–339. [PubMed: 16939798]

91. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, et al. ArrayExpress--a public database of microarray experiments and gene expression profiles. Nucleic Acids Res 2007;35(Database issue):D747–D750. [PubMed: 17132828]

92. Kato K, Yamashita R, Matoba R, Monden M, Noguchi S, Takagi T, et al. Cancer gene expression database (CGED): a database for gene expression profiling with accompanying clinical information of human cancer tissues. Nucleic Acids Res 2005;33(Database issue):D533–D536. [PubMed: 15608255]

93. Marinelli RJ, Montgomery K, Liu CL, Shah NH, Prapong W, Nitzberg M, et al. The Stanford Tissue Microarray Database. Nucleic Acids Res 2008;36(Database issue):D871–D877. [PubMed: 17989087]

94. Cole CL, Kanter AS, Cummens M, Vostinar S, Naeymi-Rad F. Using a terminology server and consumer search phrases to help patients find physicians with particular expertise. Stud Health Technol Inform 2004;107(Pt 1):492–496. [PubMed: 15360861]

95. Boulos MN. A first look at HealthCyberMap medical semantic subject search engine. Technol Health Care 2004;12(1):33–41. [PubMed: 15096685]

96. Detwiler LT, Chung E, Li A, Mejino JL Jr, Agoncillo A, Brinkley J, Rosse C, Shapiro L. A relation-centric query engine for the Foundational Model of Anatomy. Stud Health Technol Inform 2004;107 (Pt 1):341–345. [PubMed: 15360831]

97. Lee M, Wang W, Yu H. Exploring supervised and unsupervised methods to detect topics in biomedical text. BMC Bioinformatics 2006;7:140. [PubMed: 16539745]

98. Yamamoto Y, Takagi T. Biomedical knowledge navigation by literature clustering. J Biomed Inform 2007;40(2):114–130. [PubMed: 16996316]

99. Darmoni SJ, Neveol A, Renard JM, Gehanno JF, Soualmia LF, Dahamna B, et al. A MEDLINE categorization algorithm. BMC Med Inform Decis Mak 2006;6:7. [PubMed: 16464249]

100. Ruch P. Automatic assignment of biomedical categories: toward a generic approach. Bioinformatics 2006;22(6):658–664. [PubMed: 16287934]

101. Robinson J, de Lusignan S, Kostkova P, Madge B. Using UMLS to map from a library to a clinical classification: Improving the functionality of a digital library. Stud Health Technol Inform 2006;121:86–95. [PubMed: 17095807]

102. Fung KW, Bodenreider O, Aronson AR, Hole WT, Srinivasan S. Combining lexical and semantic methods of inter-terminology mapping using the UMLS. Stud Health Technol Inform 2007;129(Pt 1):605–609. [PubMed: 17911788]

103. Sun Y. Methods for automated concept mapping between medical databases. J Biomed Inform 2004;37(3):162–178. [PubMed: 15196481]

104. Bodenreider, O.; Nelson, SJ.; Hole, WT.; Chang, HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies; Proc AMIA Symp; 1998. p. 815-819.

105. Cimino, JJ.; Johnson, SB.; Peng, P.; Aguirre, A. From ICD9-CM to MeSH using the UMLS: a how-to guide; Proc Annu Symp Comput Appl Med Care; 1993. p. 730-734.

106. Brown SH, Husser CS, Wahner-Roedler D, Bailey S, Nugent L, Porter K, Bauer BA, Elkin PL. Using SNOMED CT as a reference terminology to cross map two highly pre-coordinated classification systems. Stud Health Technol Inform 2007;129(Pt 1):636–639. [PubMed: 17911794]

107. Zhang, S.; Bodenreider, O. Alignment of multiple ontologies of anatomy: deriving indirect mappings from direct mappings to a reference; AMIA Annu Symp Proc; 2005. p. 864-868.

108. Cimino JJ, Bright TJ, Li J. Medication reconciliation using natural language processing and controlled terminologies. Stud Health Technol Inform 2007;129(Pt 1):679–683. [PubMed: 17911803]
109. Parrish, F.; Do, N.; Bouhaddou, O.; Warnekar, P. Implementation of RxNorm as a terminology mediation standard for exchanging pharmacy medication between federal agencies; AMIA Annu Symp Proc; 2006. p. 1057
110. Mead CN. Data interchange standards in healthcare IT--computable semantic interoperability: now possible but still difficult, do we really need a better mousetrap? J Healthc Inf Manag 2006;20(1): 71–78. [PubMed: 16429961]
111. McDonald CJ. The barriers to electronic medical record systems and how to overcome them. J Am Med Inform Assoc 1997;4(3):213–221. [PubMed: 9147340]
112. Tu SW, Campbell JR, Glasgow J, Nyman MA, McClure R, McClay J, et al. The SAGE Guideline Model: achievements and overview. J Am Med Inform Assoc 2007;14(5):589–598. [PubMed: 17600098]
113. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, et al. Advancing translational research with the Semantic Web. BMC Bioinformatics 2007;8:S2. [PubMed: 17493285]
114. Bouhaddou O, Warnekar P, Parrish F, Do N, Mandel J, Kilbourne J, et al. Exchange of Computable Patient Data Between the Department of Veterans Affairs (VA) and the Department of Defense (DoD): Terminology Standards Strategy. J Am Med Inform Assoc. 2007
115. Khan AN, Griffith SP, Moore C, Russell D, Rosario AC Jr, Bertolli J. Standardizing laboratory data by mapping to LOINC. J Am Med Inform Assoc 2006;13(3):353–355. [PubMed: 16501183]
116. Dolin RH. Advances in data exchange for the clinical laboratory. Clin Lab Med 1999;19(2):385–419. [PubMed: 10421962]vii
117. Fridsma DB, Evans J, Hastak S, Mead CN. The BRIDG Project: A Technical Report. J Am Med Inform Assoc 2008;15(2):130–137. [PubMed: 18096907]
118. Rector, A.; Qamar, R.; Marley, T. Binding ontologies & coding Systems to electronic mealth records and messages. In: Bodenreider, O., editor. Proceedings of the Second International Workshop on Formal Biomedical Knowledge Representation (KR-MED 2006); 2006. p. 11-19.
119. Choi J, Jenkins ML, Cimino JJ, White TM, Bakken S. Toward semantic interoperability in home health care: formally representing OASIS items for integration into a concept-oriented terminology. J Am Med Inform Assoc 2005;12(4):410–417. [PubMed: 15802480]
120. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, et al. HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc 2006;13(1):30–39. [PubMed: 16221939]
121. Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragoso G, et al. caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability. J Biomed Inform 2008;41(1):106–123. [PubMed: 17512259]
122. Hernandez T, Kambhampati S. Integration of biological sources: Current systems and challenges ahead. Sigmod Record 2004;33(3):51–60.
123. Brooksbank C, Quackenbush J. Data standards: a call to action. OMICS 2006;10(2):94–99. [PubMed: 16901212]
124. Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, et al. TAMBIS: transparent access to multiple bioinformatics information sources. Bioinformatics 2000;16(2):184–185. [PubMed: 10842744]
125. Louie, B.; Mork, P.; Shaker, R.; Kolker, N.; Kolker, E.; Tarczy-Hornoch, P. Integration of data for gene annotation using the BioMediator system; AMIA Annu Symp Proc; 2005. p. 1036
126. Perez-Rey D, Maojo V, Garcia-Remesal M, Alonso-Calvo R, Billhardt H, Martin-Sanchez F, et al. ONTOFUSION: ontology-based integration of genomic and clinical databases. Comput Biol Med 2006;36(7–8):712–730. [PubMed: 16144697]
127. Joubert M, Dufour JC, Aymard S, Falco L, Fieschi M. Designing and implementing health data and information providers. Int J Med Inform 2005;74(2–4):133–140. [PubMed: 15694618]
128. Brandt, CA.; Lu, CC.; Nadkarni, PM. Automating identification of adverse events related to abnormal lab results using standard vocabularies; AMIA Annu Symp Proc; 2005. p. 903
129. Gennari, JH.; Silberfein, A.; Wiley, JC. Integrating genomic knowledge sources through an anatomy ontology; Pac Symp Biocomput; 2005. p. 115-126.

130. Chabalier J, Mosser J, Burgun A. Integrating biological pathways in disease ontologies. Stud Health Technol Inform 2007;129(Pt 1):791–795. [PubMed: 17911825]

131. Musen MA. Dimensions of knowledge sharing and reuse. Comput Biomed Res 1992;25(5):435–467. [PubMed: 1395522]

132. Bergeron E, Simons R, Linton C, Yang F, Tallon JM, Stewart TC, et al. Canadian benchmarks in trauma. J Trauma 2007;62(2):491–497. [PubMed: 17297340]

133. Fang J, Alderman MH, Keenan NL, Croft JB. Declining US stroke hospitalization since 1997: National Hospital Discharge Survey, 1988–2004. Neuroepidemiology 2007;29(3–4):243–249. [PubMed: 18176081]

134. Lieberman, MI.; Ricciardi, TN.; Masarie, FE.; Spackman, KA. The use of SNOMED CT simplifies querying of a clinical data warehouse; AMIA Annu Symp Proc; 2003. p. 910

135. Steindel, S.; Loonsk, JW.; Sim, A.; Doyle, TJ.; Chapman, RS.; Groseclose, SL. Introduction of a hierarchy to LOINC to facilitate public health reporting; Proc AMIA Symp; 2002. p. 737-741.

136. Gunnarsdottir OS, Rafnsson V. Seven-year evolution of discharge diagnoses of emergency department users. Eur J Emerg Med 2007;14(4):193–198. [PubMed: 17620908]

137. Noyes K, Liu H, Holloway R, Dick AW. Accuracy of Medicare claims data in identifying Parkinsonism cases: comparison with the Medicare current beneficiary survey. Mov Disord 2007;22 (4):509–514. [PubMed: 17230477]

138. Schneeweiss S, Robicsek A, Scranton R, Zuckerman D, Solomon DH. Veteran's affairs hospital discharge databases coded serious bacterial infections accurately. J Clin Epidemiol 2007;60(4): 397–409. [PubMed: 17346615]

139. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998;95(25):14863–14868. [PubMed: 9843981]

140. Draghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. Global functional profiling of gene expression. Genomics 2003;81(2):98–104. [PubMed: 12620386]

141. Gene Ontology Tools. http://geneontology.org/GO.tools.shtml

142. Khatri P, Sellamuthu S, Malhotra P, Amin K, Done A, Draghici S. Recent additions and improvements to the Onto-Tools. Nucleic Acids Res 2005;33(Web Server issue):W762–W765. [PubMed: 15980579]

143. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol 2003;4(4):R28. [PubMed: 12702209]

144. Zhang B, Schmoyer D, Kirov S, Snoddy J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. BMC Bioinformatics 2004;5:16. [PubMed: 14975175]

145. Djebbari A, Karamycheva S, Howe E, Quackenbush J. MeSHer: identifying biological concepts in microarray assays based on PubMed references and MeSH terms. Bioinformatics 2005;21(15): 3324–3326. [PubMed: 15919728]

146. Bresell A, Servenius B, Persson B. Ontology annotation treebrowser : an interactive tool where the complementarity of medical subject headings and gene ontology improves the interpretation of gene lists. Appl Bioinformatics 2006;5(4):225–236. [PubMed: 17140269]

147. Osborne JD, Zhu LJ, Lin SM, Kibbe WA. Interpreting microarray results with gene ontology and MeSH. Methods Mol Biol 2007;377:223–242. [PubMed: 17634620]

148. Brameier M, Wiuf C. Co-clustering and visualization of gene expression data and gene ontology terms for Saccharomyces cerevisiae using self-organizing maps. J Biomed Inform 2007;40(2):160–173. [PubMed: 16824804]

149. Huang D, Wei P, Pan W. Combining gene annotations and gene expression data in model-based clustering: weighted method. OMICS 2006;10(1):28–39. [PubMed: 16584316]

150. Liu, J.; Wang, W.; Yang, J. Gene Ontology friendly biclustering of expression profiles; Proc IEEE Comput Syst Bioinform Conf; 2004. p. 436-447.

151. Lord, PW.; Stevens, RD.; Brass, A.; Goble, CA. Semantic similarity measures as tools for exploring the gene ontology; Pac Symp Biocomput; 2003. p. 601-612.

152. Wolting C, McGlade CJ, Tritchler D. Cluster analysis of protein array results via similarity of Gene Ontology annotation. BMC Bioinformatics 2006;7:338. [PubMed: 16836750]
153. Achour SL, Dojat M, Rieux C, Bierling P, Lepage E. A UMLS-based knowledge acquisition tool for rule-based clinical decision support system development. J Am Med Inform Assoc 2001;8(4): 351–360. [PubMed: 11418542]
154. Kashyap, V.; Morales, A.; Hongsermeier, T. On implementing clinical decision support: achieving scalability and maintainability by combining business rules and ontologies; AMIA Annu Symp Proc; 2006. p. 414-418.
155. Greenes, RA. Clinical decision support : the road ahead. Amsterdam ; Boston: Elsevier Academic Press; 2007.
156. Rubin DL, Dameron O, Bashir Y, Grossman D, Dev P, Musen MA. Using ontologies linked with geometric models to reason about penetrating injuries. Artif Intell Med 2006;37(3):167–176. [PubMed: 16730959]
157. Marquet, G.; Dameron, O.; Saikali, S.; Mosser, J.; Burgun, A. Grading glioma tumors using OWL-DL and NCI Thesaurus; AMIA Annu Symp Proc; 2007 Oct. p. 508-512.
158. Wolstencroft KJ, Stevens R, Tabernero L, Brass A. PhosphaBase: an ontology-driven database resource for protein phosphatases. Proteins 2005;58(2):290–294. [PubMed: 15558746]
159. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008;15(1):14–24. [PubMed: 17947624]
160. Meystre S, Haug PJ. Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. J Biomed Inform 2006;39(6):589–599. [PubMed: 16359928]
161. Chapman WW, Fiszman M, Dowling JN, Chapman BE, Rindflesch TC. Identifying respiratory findings in emergency department reports for biosurveillance using MetaMap. Medinfo 2004;11(Pt 1):487–491.
162. BioCaster. http://biocaster.nii.ac.jp/
163. EpiSpider. http://www.epispider.org/
164. Rindflesch, TC.; Fiszman, M.; Libbus, B. Semantic interpretation for the biomedical literature. In: Chen, H.; Fuller, S.; Hersh, WR.; Friedman, C., editors. Medical informatics: Advances in knowledge management and data mining in biomedicine. Springer-Verlag; 2005. p. 399-422.
165. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161–174. [PubMed: 7719797]
166. Lussier, Y.; Borlawsky, T.; Rappaport, D.; Liu, Y.; Friedman, C. PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing; Pac Symp Biocomput; 2006. p. 64-75.
167. Tessi. http://www.landcglobal.com/pages/tessi_indexing.php
168. Jacquemart P, Zweigenbaum P. Towards a medical question-answering system: a feasibility study. Stud Health Technol Inform 2003;95:463–468. [PubMed: 14664030]
169. Terol RM, Martinez-Barco P, Palomar M. A knowledge based method for the medical question answering problem. Comput Biol Med 2007;37(10):1511–1521. [PubMed: 17374369]
170. Wedgwood, J. MQAF: a medical question-answering framework; AMIA Annu Symp Proc; 2005. p. 1150
171. EAGLi. http://eagl.unige.ch/EAGLi/
172. Fiszman M, Rindflesch TC, Kilicoglu H. Summarization of an online medical encyclopedia. Stud Health Technol Inform 2004;107(Pt 1):506–510. [PubMed: 15360864]
173. Reeve LH, Han H, Brooks AD. Biomedical text summarisation using concept chains. International Journal of Data Mining and Bioinformatics 2007;1(4):389–407. [PubMed: 18402049]
174. Whalen, G. Medical textbook summarization and guided navigation using statistical sentence extraction; AMIA Annu Symp Proc; 2005. p. 814-818.
175. Lussier, YA.; Li, J. Terminological mapping for high throughput comparative biology of phenotypes; Pac Symp Biocomput; 2004. p. 202-213.
176. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. Int J Med Inform 2005;74(2–4):289–298. [PubMed: 15694635]

177. Torvik VI, Smalheiser NR. A quantitative model for linking two disparate sets of articles in MEDLINE. Bioinformatics 2007;23(13):1658–1665. [PubMed: 17463015]

178. Yetisgen-Yildiz M, Pratt W. Using statistical and knowledge-based approaches for literature-based discovery. J Biomed Inform 2006;39(6):600–611. [PubMed: 16442852]

179. Weeber M, Vos R, Klein H, De Jong-Van Den Berg LT, Aronson AR, Molema G. Generating hypotheses by discovering implicit associations in the literature: a case report of a search for new potential therapeutic uses for thalidomide. J Am Med Inform Assoc 2003;10(3):252–259. [PubMed: 12626374]

180. Weinstein JN. 'Omic' and hypothesis-driven research in the molecular pharmacology of cancer. Curr Opin Pharmacol 2002;2(4):361–365. [PubMed: 12127867]

181. Gopalacharyulu PV, Lindfors E, Miettinen J, Bounsaythip CK, Oresic M. An integrative approach for biological data mining and visualisation. International Journal of Data Mining and Bioinformatics 2008;2(1):54–77. [PubMed: 18399328]

182. Uccelli R, Binazzi A, Altavista P, Belli S, Comba P, Mastrantonio M, et al. Geographic distribution of amyotrophic lateral sclerosis through motor neuron disease mortality data. Eur J Epidemiol 2007;22(11):781–790. [PubMed: 17874192]

183. Linares C, Diaz J. Impact of high temperatures on hospital admissions: comparative analysis with previous studies about mortality (Madrid). Eur J Public Health. 2007

184. Ramaswamy K, Kozma CM, Nasrallah H. Risk of diabetic ketoacidosis after exposure to risperidone or olanzapine. Drug Saf 2007;30(7):589–599. [PubMed: 17604410]

185. Jinjuvadia K, Kwan W, Fontana RJ. Searching for a needle in a haystack: use of ICD-9-CM codes in drug-induced liver injury. Am J Gastroenterol 2007;102(11):2437–2443. [PubMed: 17662100]

186. Choy KW, Wang CC, Ogura A, Lau TK, Rogers MS, Ikeo K, et al. Molecular characterization of the developmental gene in eyes: through datamining on integrated transcriptome databases. Clin Biochem 2006;39(3):224–230. [PubMed: 16427038]

187. Jones CE, Baumann U, Brown AL. Automated methods of predicting the function of biological sequences using GO and BLAST. BMC Bioinformatics 2005;6:272. [PubMed: 16288652]

188. Carmona-Saez P, Chagoyen M, Rodriguez A, Trelles O, Carazo JM, Pascual-Montano A. Integrated analysis of gene expression by Association Rules Discovery. BMC Bioinformatics 2006;7:54. [PubMed: 16464256]

189. Zhou Y, Young JA, Santrosyan A, Chen K, Yan SF, Winzeler EA. In silico gene function prediction using ontology-based pattern identification. Bioinformatics 2005;21(7):1237–1245. [PubMed: 15531612]

190. Huang, JC.; Frey, BJ.; Morris, QD. Comparing sequence and expression for predicting microRNA targets using GenMiR3; Pac Symp Biocomput; 2008. p. 52-63.

191. Barutcuoglu Z, Schapire RE, Troyanskaya OG. Hierarchical multi-label prediction of gene function. Bioinformatics 2006;22(7):830–836. [PubMed: 16410319]

192. Sahoo SS, Zeng K, Bodenreider O, Sheth A. From "glycosyltransferase" to "congenital muscular dystrophy": integrating knowledge from NCBI Entrez Gene and the Gene Ontology. Stud Health Technol Inform 2007;129(Pt 2):1260–1264. [PubMed: 17911917]

193. Camargo A, Azuaje F. Linking gene expression and functional network data in human heart failure. PLoS ONE 2007;2(12):e1347. [PubMed: 18094754]

194. Chabalier J, Mosser J, Burgun A. A transversal approach to predict gene product networks from ontology-based similarity. BMC Bioinformatics 2007;8:235. [PubMed: 17605807]

195. Yildirim MA, Goh KI, Cusick ME, Barabasi AL, Vidal M. Drug-target network. Nat Biotechnol 2007;25(10):1119–1126. [PubMed: 17921997]

196. RxNav. http://mor.nlm.nih.gov/download/rxnav/

197. Open Biomedical Ontology. http://www.obofoundry.org/

198. UMLS Knowledge Source Server. http://umlsks.nlm.nih.gov/

199. BioPortal. http://www.bioontology.org/tools/portal/bioportal.html

200. Moreira DA, Musen MA. OBO to OWL: a protege OWL tab to read/save OBO ontologies. Bioinformatics 2007;23(14):1868–1870. [PubMed: 17496317]

201. LexGrid. http://informatics.mayo.edu/LexGrid/

202. 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):1251–1255. [PubMed: 17989687]

203. Rogers JE. Quality assurance of medical ontologies. Methods Inf Med 2006;45(3):267–274. [PubMed: 16685334]

204. Smith, B. The evaluation of ontologies: editorial review vs. democratic ranking; Proceedings of the First Interdisciplinary Ontology Meeting; 2008. p. 29-36.

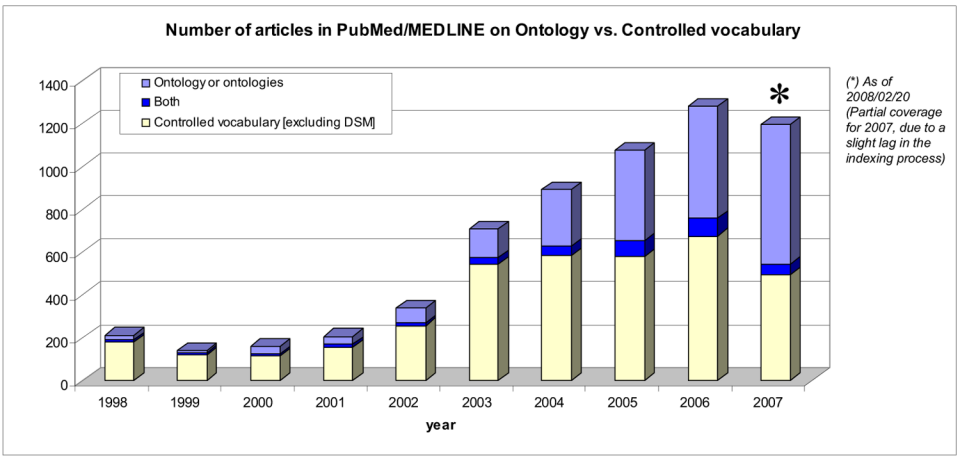

Fig. 1.

Evolution of the number of citations in PubMed/MEDLINE on ontologies and controlled vocabularies over the past 10 years (excluding DSM, the Diagnostic and Statistical Manual of Mental Disorders)

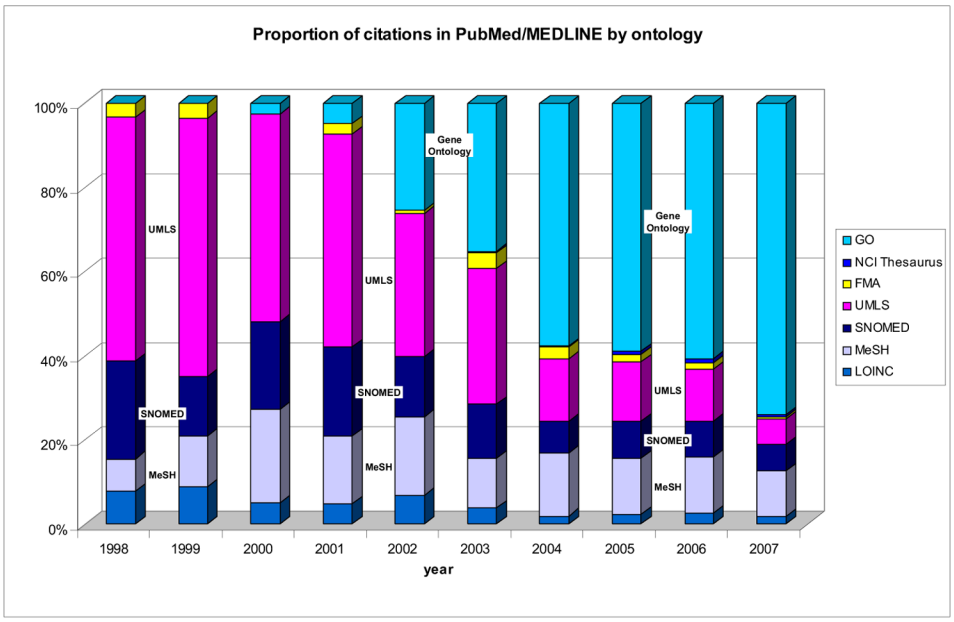

Fig. 2.
Evolution of the proportion of citations in PubMed/MEDLINE by ontology.

Table 1

Characteristics of some biomedical ontologies (including scope, number of entities, distribution of the number of terms per entity [minimum, maximum, median and average], and existence of a subsumption hierarchy), based on information present in the UMLS (2007AC)

Name	Ref.	Scope	# concepts	# concept names				Subs. Hier.	Version / Notes
				Min	Max	Med	Avg		
SNOMED CT	[21]	Clinical medicine (patient records)	310,314	1	37	2	2.57	yes	July 31, 2007
LOINC	[24]	Clinical observations and laboratory tests	46,406	1	3	3	2.85	no	Version 2.21(no “natural language” names)
FMA	[25]	Human anatomical structures	~72,000	1	?	?	~1.50	yes	(not yet in the UMLS)
Gene Ontology	[28]	Functional annotation of gene products	22,546	1	24	1	2.15	yes	Jan. 2, 2007
RxNorm	[31]	Standard names for prescription drugs	93,426	1	2	1	1.10	no	Aug. 31, 2007
NCI Thesaurus	[34]	Cancer research, clinical care, public information	58,868	1	100	2	2.68	yes	2007_05E
ICD-10	[36]	Diseases and conditions (health statistics)	12,318	1	1	1	1.00	no	1998 (tabular)
MeSH	[38]	Biomedicine (descriptors for indexing the literature)	24,767	1	208	5	7.47	no	Aug. 27, 2007
UMLS Meta.	[41]	Terminology integration in the life sciences	1,4 M	1	339	2	3.77	n/a	2007AC (English only)