doi:10.1093/bib/bbs053

Evaluation of research in biomedical ontologies

Robert Hoehndorf, Michel Dumontier and Georgios V. Gkoutos

Submitted: 15th May 2012; Received (in revised form): 19th July 2012

Abstract

Ontologies are now pervasive in biomedicine, where they serve as a means to standardize terminology, to enable access to domain knowledge, to verify data consistency and to facilitate integrative analyses over heterogeneous biomedical data. For this purpose, research on biomedical ontologies applies theories and methods from diverse disciplines such as information management, knowledge representation, cognitive science, linguistics and philosophy. Depending on the desired applications in which ontologies are being applied, the evaluation of research in biomedical ontologies must follow different strategies. Here, we provide a classification of research problems in which ontologies are being applied, focusing on the use of ontologies in basic and translational research, and we demonstrate how research results in biomedical ontologies can be evaluated. The evaluation strategies depend on the desired application and measure the success of using an ontology for a particular biomedical problem. For many applications, the success can be quantified, thereby facilitating the objective evaluation and comparison of research in biomedical ontology. The objective, quantifiable comparison of research results based on scientific applications opens up the possibility for systematically improving the utility of ontologies in biomedical research.

Keywords: *biomedical ontology; quantitative biology; ontology evaluation; evaluation criteria; ontology-based applications*

INTRODUCTION

Biomedical ontology is an emerging discipline that applies theories and methods from diverse disciplines such as philosophy, cognitive science, linguistics and formal logics to perform or improve biomedical applications. As a scientific discipline, it requires a research methodology that yields reproducible and comparable results that can be evaluated independently. Methodological progress in biomedical ontology will be recognized when different methods generate results that can be objectively compared, such that it becomes possible to evaluate whether the methods yield better results.

There is considerable debate about establishing metrics for evaluating research results in applied ontology as well as determining the perspective from which its results should be evaluated [1–3]. Many evaluation strategies are based on criteria stemmed from philosophy, knowledge representation, formal logics or 'common sense', while an empirical, repeatable and reproducible evaluation based on the domain of application is challenging to perform [4, 5]. The absence of commonly agreed criteria for evaluating research results in biomedical ontology leads to challenges in the development of an effective research methodology for the field of biomedical ontology: before a research methodology in any scientific field can be established, it is first necessary to determine what constitutes a research result, what constitutes a 'novel' research result (i.e. what does it mean that two research results are different) and what constitutes a better result than another (i.e. how can two competing results be compared and evaluated). Only after these questions

Corresponding author. Robert Hoehndorf, Department of Computer Science, Aberystwyth University, Aberystwyth, Ceredigion, SY23 3DB, UK. Tel: +44-1970-622950; Fax: +44-1970-628536; E-mail: leechuck@leechuck.de

Robert Hoehndorf is a research associate at the University of Cambridge and a visiting researcher at Aberystwyth University. His research is focused on using biomedical ontologies for computational analyses in integrative biology.

Michel Dumontier is a professor of Bioinformatics at Carleton University. His research interests involve the development of ontology-based applications for personalized medicine.

Georgios V. Gkoutos is a reader in Bioinformatics at the University of Aberystwyth. His research interests are in developing biomedical ontologies and applying them to the study of association between genotype and phenotype.

are answered will it be possible to design a research methodology in a scientific field than enables the field as a whole to make progress with respect to the evaluation criteria that the discipline has established.

Here, we review fundamental questions pertaining to research in biomedical ontologies. We will focus on the application of ontologies in basic and translational research and will not discuss the large field of applying ontologies in health care and medicine, which is discussed elsewhere [6–9]. First, we review major applications of ontologies in biomedical research. From the perspective of an ontology user, we then discuss the problem of the 'research question' of biomedical ontology, i.e. what is the 'scientific' problem that research in biomedical ontology addresses. Third, we characterize and classify different types of research results in biomedical ontology, and finally, we discuss in depth different ways for evaluating and comparing research results in biomedical ontology. Although we will primarily focus on ontologies as they are used in biomedicine, we believe that many of our arguments will hold for research in other areas of applied ontology as well.

USES OF ONTOLOGIES IN BIOMEDICAL INVESTIGATIONS

Biomedical applications of ontologies

At the end of the 1990s and early 2000s, genetics made a leap forward with the availability of the first genome sequences for several species [10]. The availability of genome sequences for multiple species enabled comparative genomic analyses and revealed that a large part of the genetic material in different species was conserved and that many of the genes in different organisms have similar functions. The Gene Ontology (GO) [11] was designed as a controlled vocabulary to provide stable names, textual definitions and identifiers to unify descriptions of functions, processes and cellular components across databases in biology. Today, with the rise of high-throughput sequencing technology, genome sequences for thousands of species are becoming available, and large international research projects, such as the 5000 genomes project (which aims to sequence the genomes of 5000 insects and other arthropods) [12] or the Genomes 10 k project (which aims to sequence the genomes of 10 000 vertebrate species) [13], will collect even more data in the near future. High-throughput technologies are not limited to genome sequencing, but influenced other areas in biology as well, from high-throughput phenotyping (to determine the observable characteristics of organisms, often resulting from targeted mutations) [14, 15] over microarray experiments (to determine gene expression) [16] to high-throughput screening in drug discovery [17, 18]. The amount of data produced in biology today makes the design of strategies for integration of data across databases, methods for retrieving the data and developing query languages and interfaces a central and important part of research in biology. The prime purpose of ontologies, such as the GO, is to address these arising challenges in biology and biomedicine and provide a means to integrate data across multiple heterogeneous databases.

To facilitate the integration of databases, retrieval of data and the provision of query languages, ontologies provided not only terms and textual definitions but also a basic structure. Initially, this structure was not expressed in a formal logic-based language. Instead, ontologies were seen as graph structures in which nodes represent terms and edges relations (such as 'is-a' or 'part-of') between them. Reasoning over these graphs was stated as operations on the graph, in particular the composition of edges and the transitive closure [11]. It was not until much later that formal languages were used to represent biomedical ontologies and recast the graph operation in terms of deductive inference over formal theories [19–22].

The graph structure of biomedical ontologies is not only a valuable feature to improve retrieval and querying but is also useful for other tasks, for example for Gene Set Enrichment Analysis (GSEA) [23] to analyse gene expression. GSEA utilizes the graph structure of the GO to determine whether a defined set of genes shows statistically significant, concordant differences between two biological states; it utilizes the annotation of sets of genes with GO terms and the GO graph structure and inference rules to statistically test for enriched GO terms. A large number of tools were developed to perform such enrichment analyses that have lead to discoveries of cancer mechanisms [23], evolutionary differences in primates [24], genes involved in particular functions, such as oxidative phosphorylation [25] and many more. GSEA is now a standard tool in many biological analyses, as evidenced by more than 3200 citations (based on Google Scholar, 5 April 2012) for the original paper. Similar enrichment analyses are now being performed using ontologies of other domains, such as the Human Disease Ontology [26].

The graph structure of ontologies is also widely utilized for semantic similarity analyses [27]. Semantic similarity measures apply a metric on an ontology in order to compare the similarity between data that are annotated with classes in the ontology [28–30]. Some metrics are based on the distance between two nodes in the ontologies' graph structure, while others compare sets of classes that are closed with respect to relations in the ontology [31–33]. In some cases, the metrics include further information, such as the information content that a class in an ontology has within a given domain. Importantly, however, all semantic similarity measures rely on the number and the kind of distinctions that the ontology developers have made explicit. Another application of ontologies is in text mining and literature search and retrieval [34, 35]. The availability of a common terminology throughout biology enables the task of named entity recognition, i.e. the identification of standardized terms in natural language texts [36, 37]. When terms from ontologies can reliably be detected in natural language texts, ontologies can be used for retrieving text documents from literature archives such as PubMed [38]. This task is made easier when terms in ontologies are widely used, and several biomedical ontologies have been evaluated based on how well their terms can be recognized in scientific literature [39]. Furthermore, identification of ontology term labels in text can be combined with analyses over the structure of ontologies (including similarity-based analyses and enrichment analyses) to improve text-mining results based on the ontology hierarchy.

Ontologies are also used as knowledge bases (or structured databases) which are primarily intended to store and expose information about a domain. Ontologies of this type are comparable to scientific databases, such as UniProt [40], in that they contain information for scientists that can be accessed on demand. Examples for this type of ontology include the various anatomy ontologies [41–47] and pathway knowledge bases such as EcoCyc and MetaCyc [48]. These ontologies can go into great detail; an ontology like the Foundational Model of Anatomy (FMA) [43] is likely the most comprehensive formal description of human anatomy and exceeds the information and the detail contained in most individual anatomy textbooks.

Ontologies as formalized theories of a domain

Although the applications of biomedical ontologies we discussed so far do not rely on formalized semantics, axioms, the use of knowledge representation languages, automated reasoning or philosophical foundations, the past years have seen a rapid increase in applying formal methods to biomedical ontologies. In particular, the Web Ontology Language [49] is now widely used to represent biomedical ontologies [19]. In some cases, more expressive languages such as first- and monadic second-order logic are used to specify ontologies, in particular for biological sequences [50, 51] and molecular structures and graphs [52]. Using these languages, knowledge about a domain is expressed following the axiomatic method [53], based on which axioms (i.e. statements that are considered to be true about the domain) are asserted and the consequences of these axioms are inferred using inference rules [54]. Automated reasoning is the process by which the inferences are deduced automatically.

The stated aims of applying philosophical foundations, the axiomatic method, knowledge representation languages and automated reasoning for biomedical ontologies are manifold and include the search for philosophical rigour and a foundation in particular philosophical theories [3, 5, 55], providing expressive and machine-readable documentation of the meaning of terms in a vocabulary [51], verifying the consistency of a data model [56, 57], verifying the consistency of data with respect to a data model [56, 58], enabling complex retrieval and querying through automated reasoning [59], integrating multiple ontologies [60, 61] and decreasing the cost of developing and maintaining an ontology [62–65]. Furthermore, the application of formal methods in biomedical ontologies has the potential to reveal mistakes in the design of ontologies [5, 21, 66, 67] or to improve their utility for scientific analyses [61, 68]. Several projects have started to axiomatize biomedical ontologies [57, 61, 69, 70], and these projects have led to changes in the ontologies and the detection and removal of contradictory statements [57, 60]. Other researchers have suggested changes to improve ontologies' structures and axioms based on applying formal, ontological and philosophical methods [21, 55, 66, 67, 71, 72], or they provide ontological interpretations of domain-specific knowledge by applying a formal ontological theory to some phenomena in a domain [67, 73–75]. Table 1 provides a list of use cases and examples for the application of formal methods in biomedical ontologies.

THE RESEARCH QUESTIONS OF BIOMEDICAL ONTOLOGY

The examples we discussed include the current major applications of ontologies in biomedical research, and additional ontology-based applications are developed frequently and range from novel scientific data analysis methods over the design of user interfaces to semantic publishing of scientific articles. One underlying commonality in the ontology-based applications reviewed here is that ontologies determine or guide the 'way' in which domain content is expressed. Research in ontology answers the questions of 'how' a proposed standard terminology should be built so that it satisfies the needs of multiple users in the domain, 'how' domain content must be expressed so that relevant retrieval operations or particular scientific analyses are supported and 'how' information must be formalized so that data and model consistency can be verified with regard to specific constraints. In most cases, there are multiple possibilities for structuring information within a domain and not all perform equally well. Additionally, it may be possible to identify common underlying principles of 'how' to structure information within a domain in order to serve particular applications. While these principles may originate from diverse disciplines, including philosophy, linguistics and cognitive science, it is their effect on biomedical applications that makes them either successful or unsuccessful choices. In this sense, the research area of 'ontology' is the bridge between theories originating from these diverse disciplines and the domain of application; ontology is about selecting the right way of modelling a domain for a particular application.

Following this understanding of 'ontology', we can distinguish between several different types of research results. First, ontologies themselves are research results in biomedical ontology. An ontology is an artefact that specifies a particular set of categories that are useful and applicable for certain tasks within a domain. More than 300 ontologies in the biomedical domain are listed in the BioPortal [82] alone, and their intended applications are highly diverse, covering all the use cases we discussed so far and more.

A second type of research result is an 'ontology design pattern', i.e. a 'way' to represent information so that it can be applied for a specific purpose [83]. Many of these patterns are currently implemented in domain ontologies, or arose from best practices in building ontologies. Most notably, relations in biomedical ontologies were a controversial topic for several years [66] until a set of ontology patterns was proposed that standardized the meaning of a large number of relations used in biomedical ontologies [21]. Similarly, in the biomedical domain, patterns have been proposed for expressing information about qualities [84], functions [85], dispositions [73], phenotypes [86, 87] and realizable entities [67]. Often, these patterns are motivated by theories taken from other scientific fields and applied to the field of biomedicine. For example, well-developed ontological theories of functions are available in philosophy [88], biology [89] and linguistics [90] and can be applied to formulate biological knowledge. Since not all of them will perform equally for all tasks, the evaluation of an ontology design patterns requires the application of the design pattern to a particular ontology and a measure of its impact in an ontology-based application.

We consider the 'application of a design pattern to an ontology' a third type of research result. Applying a design patterns often involves changing an ontology so that certain information is structured according to the design pattern. This can either be done on a single place in the ontology, in order to demonstrate the consequences of applying the design pattern, or throughout an ontology. In the first case, consequences can be measured on a single example and their effect on the whole ontology could be hypothesized. Only the second case will enable the direct evaluation of all consequences.

Finally, a fourth type of research result is a methodological result. Methodological advances in applied ontology may abstract from specific applications of ontologies and identify generic approaches that will lead to reproducible positive outcomes in certain scenarios. These approaches can eventually lead to guidelines for ontology quality with respect to certain application. For example, the OntoClean approach [91] is such a general method for building ontologies that are robust (i.e. re-usable across multiple applications) and comprehensible.

Table 1: Formal approaches to ontology research and their potential impact on biomedical applications and analyses

Task	Description	Potential impact on biomedical applications and analyses	Example
Philosophical foundations	A theory from philosophy is applied either to biomedical ontologies orro biomedical domains. It is then demonstrated that the philosophical theory can explain the distinctions within the domain. Furthermore, philosophical foundation theory can provide insights into the principles based on which scientists within a domain distinguish different kinds of entities and can provide a methodology for classifying domain entities. Formalizing aspects of the philosophical principles can enable verification of a domain theory with regard to these principles.	Increased coherence in representing knowledge and data, comprehensibility and interoperability	A study demonstrated that a particular perspective on philosophical realism can be used to describe chemical structures, even when the type of structure is known to be impossible to exist [55].
Provision of unambiguous, formal documentation	The use of formal languages can remove ambiguity from specifications (of a domain, the meaning of a term, etc.). Based on formal logics, consequences of a specification can then be determined by a mathematical proof, thereby avoiding potential misunderstandings based on natural language.	Increased coherence, increased clarity	The RNA ontology (RNAO) [51] is a biomedical ontology used to describe RNA structure. The core of RNAO is formalized in first- and second-order logic. These rich formalisms are used to precisely formalize basic notions, such as the meaning of 'molecule' within the context of the RNAO.
Provision of machine-readable documentation	Some aspects of the meaning of terms are formalized using a knowledge representation language so that automated systems can gain access to the meaning and process it.	Automated data processing, automated knowledge- and data-integration, semantic integration	The GO [11], as well as a large number of ontologies in the OBO Foundry [1], use the OBO Flatfile Format [19] to make 'some aspects' of term meanings explicit. For example, the GO contains a taxonomy as well as relations about parthood and regulation. Another project is aimed at providing richer formalized definitions for the GO [57] so that more information about the terms' meaning can be accessed automatically.
Consistency verification	Statements that are considered true in the domain (axioms) and term definitions are formalized and an automated reasoner is used to verify the consistency (i.e. the absence of contradictions) of the stated knowledge. Furthermore, 'satisfiability' of a class can be automatically verified (a class is satisfiable if it is possible for the class to have instances). Once a model of a domain is consistently formalized, it can be applied to verify data in this domain. For this purpose, an automated reasoner verifies whether data items satisfy the constraints expressed in the model of the domain Often, expressive automated reasoners such as OWL 2.0 reasoners are used to perform consistency verification.	Increased coherence, detection of modelling errors, detection of competing scientific theories, data coherence	Inconsistencies when combining anatomy and phenotype ontologies were detected [72] and resolved by explicitly distinguishing between normal and abnormal anatomy.

(continued)

Table I: Continued

Task	Description	Potential impact on biomedical applications and analyses	Example
Data classification	An ontology of a domain is applied to classify data in a domain. In this task, an automated reasoner uses the constraints in the domain ontology to automatically assign data items into ontology categories.	Classification, data analysis	A study [76] formalized human knowledge about the classification of protein phosphatases in an OWL ontologies and applied automated reasoning to automatically assign classes for human and *Aspergillus fumigatus* proteins. An evaluation showed that ontology-based classification matches, and sometimes exceeds, human judgment.
Supporting ontology development	Formal representations and automated reasoning can support ontology development by inferring information that is not explicitly stated. Possibly undesired consequences can be examined either manually or automatically and the statements leading to the undesired consequence can be corrected. Particularly useful is the automated construction of taxonomies based on axioms in an ontology. Complex statements and definitions are automatically transformed into a generalization hierarchy (a taxonomy) by the automated reasoner.	Decreased maintenance, detection of errors	The GULO software [77] uses automated reasoning over the axioms in an ontology to improve the taxonomic structure of an ontology. Furthermore, it enables ontology developers to validate the accuracy of their definitions.
Support querying	Based on the axioms about a domain, automated reasoners can infer a potentially infinite number of statements that are true if the axioms are true. Therefore, formal logics are ideally suited to encode knowledge about a domain so that it can support a wide variety of queries. Automated reasoners are capable of automatically determining the answers to the queries using the statements in the formalized theory. This is one of the most widely used application of automated reasoning in ontologies. To efficiently support querying in applications that require quick response times, highly optimized reasoners and low expressivity of the knowledge representation language are beneficial [74].	Support knowledge extraction, connect databases and domains	A web-based query tool used the ELK reasoner [78] to query the GO and the mouse phenotype ontology as well as their annotated data [79]. Another example is the FlyBase model organism database [80] which uses the Pellet reasoner [81] to perform data queries.

EVALUATION OF RESEARCH RESULTS IN BIOMEDICAL ONTOLOGY

Despite the large number of research projects that apply formal ontological theories to scientific domains, no common evaluation criteria are being applied in these studies. Similarly, the stated goals of such research are highly diverse and sometimes the impact of the research on scientific applications is not demonstrated or even discussed [2]. Examples of evaluation criteria for research in applied ontology include formal consistency of the developed theory [92], the identification of unsatisfiable classes [57, 60], conformance to a particular philosophical theory [3, 55, 93], user acceptance [94], conformance to naming conventions [95] or the recall of ontology class labels in scientific literature [39]. Only few of these criteria actually evaluate 'what ontologies do', while the majority of these criteria evaluate the research results based on philosophical, formal and technical criteria that lie within the domain of ontology or its underlying technologies themselves.

The selection and application of evaluation criteria provides the means to distinguish research in 'applied' ontology from research in 'non-applied' ontology. In 'applied' ontology, ontologies are being used for some task within a domain, and that task lies usually outside of the domain of ontology itself (A notable exception to this is when we apply ontological methods to the domain of ontology itself, and classify different kinds of ontologies, their parts, analyse the types of relations between classes, relations, instances and individuals, etc. Such an ontology could, for example, be used to provide the conceptual foundation of an ontology editor, to enable interoperability between different ontology learning algorithms, in portals providing access to different ontologies, or in an ontology evaluation framework.). Consequently, evaluation criteria for research results in 'applied' biomedical ontology will be derived based on the task to which the result is being applied, and not from the domain of ontology itself. The search for philosophical foundation and rigor, including the demonstration that a particular philosophical theory is capable of expressing distinctions that are being made within a domain, are examples of research goals of non-applied ontology, because the aims of the research and its evaluation will generally lie within the realm of ontology, not within the domain of application. Applying a particular philosophical theory can, in many cases, improve the utility of an ontology, and demonstrating that the application of a particular philosophical perspective improves the utility of an ontology for some task in a domain would constitute a result in applied biomedical ontology.

We can also distinguish between 'who' or 'what' directly benefits from a particular result of research in ontology: either the users and uses of an ontology, ontology-based applications and specific tasks to which ontologies are being applied, or the developers and maintainers of an ontology. Developers and maintainers of ontologies will benefit directly from decreased maintenance work, ease of construction and the availability of technical documentation, while users and applications of an ontology will only benefit indirectly from such research goals. Users and applications of ontologies benefit from the community agreement which ontologies can bring about and their resulting potential for ontology-based data annotation and integration, retrieval and querying, novel scientific analyses and in some cases consistency verification of data. Since, users of ontologies will benefit from something that ontologies can 'do', research in 'applied' ontology has to be measured based on how well ontologies 'do' their tasks.

One of the most widely cited applications of ontologies in science is their potential to facilitate community agreement of the meaning of terms in a domain. These terms are frequently used as meta-data in scientific databases and publications. Consequently, applying ontologies to standardize the vocabulary used as meta-data can enable the integration and interoperability of databases and research results. There are several possibilities for evaluating an ontology that is intended to effectively standardize the meaning of terms in a vocabulary and support interoperability and integration. Since the prime aim of such a research result is to achieve community agreement, an obvious evaluation criterion would be to conduct a user-study that evaluates whether different users can consistently apply terms within a standardized task such as the annotation of a data set with classes from an ontology. For this task, Kappa statistics can be applied and a κ value can be reported that measures the degree to which annotator agree [96, 97]. Kappa statistics is widely applied in computational linguistics [98], biomedical text mining [99], for the verification and disambiguation of biomedical resources [100], and to evaluate some consequences of biomedical ontologies [94].

The support of queries and retrieval of data is another task for which ontologies and their axioms are built. Information retrieval is a discipline in computer science for which rigorous quantitative evaluation criteria are available [101], often based on the comparison to a gold standard or a set of positive and negative examples based on which statistical measures can be applied. Quantitative measures include the F-measure (the harmonic mean between precision and recall) or the area-under-curve (AUC) in an analysis of the receiver operating characteristic (ROC) curve [102]. If an ontology, or axioms in an ontology, are intended for retrieval, measures that compare the inferences to a gold standard can be applied to demonstrate the success of the ontology. In many cases, axioms in ontologies are added in order to enable novel queries that make distinctions which could not be made before. For example, adding axioms about parthood to a purely taxonomic representation of anatomical structures enables new kind of queries based on the use of parthood relations. Such a result—the addition of new axioms to enable novel types of queries and retrieval operations—can be evaluated using the same quantitative measures as ontology-based retrieval. All of these descriptions assume that there is already some data which is being retrieved using queries over the ontology. In the absence of such data, e.g. when a new ontology is proposed within a domain with the intent to use this ontology to annotate data in the future, data could be simulated and then used in the evaluation.

Further applications of formalized ontologies include the verification of data with respect to certain constraints that are expressed within the ontology. For example, in the domain of biological pathways, the BioPax ontology [56] has been proposed, and one of its aims is to verify pathway data with respect to the model that the BioPax ontology provides. Similarly, a recent study used formal ontological analysis and automated reasoning to investigate the consistency of a database of computational models and identified a large number of incorrectly characterized database entries [58]. A quantitative measure of success would then be the number of inconsistencies that were identified in a data set.

Applications of ontology research in scientific analyses and in the process of making novel scientific discoveries are maybe the best evaluated contributions in applied ontology, since the contributions that ontology research can make in these areas is commonly subject to the same evaluation criteria as other contributions in the scientific domain of application. For example, the GSEA method was evaluated both using statistical measures and experimentally verified data that has been extensively studied [23, 25], and the use of semantic similarity measures to identify interacting proteins based on GO is rigorously evaluated and compared using ROC and correlation coefficient analysis [103]. In each case, the scientific domain to which ontology-based methods are being applied has established, and often demands, quantitative evaluation criteria that can ensure the objective and empirical evaluation and comparison of research results. Furthermore, an integrated scientific analysis of the data in multiple databases between which interoperability is intended to be achieved can be performed and evaluated on a scientific use case. For example, the development of formal definitions for phenotype ontologies [86] can be quantitatively evaluated by using these definitions to integrate multiple model organism databases and analyse the integrated knowledge with regard to its potential for revealing novel candidate genes for diseases [68].

There are several other tasks that may fall in the domain of applied ontology research. For example, formal ontological analysis can be applied to specify a conceptual model, verify its consistency and identify modelling choices that potentially lead to faulty results; or formal ontology can be applied to formally specify the meaning of terms in a vocabulary (e.g. to enable communication between autonomous intelligent agents). Some of these tasks can also be evaluated quantitatively: while consistency of a conceptual model is a binary quality that relies on a consistency proof, incorrect consequences can be estimated using predefined tests that aim to make inferences of a certain kind [104]. A formal specification of the meaning of a term using an ontology can be accompanied by a meta-theoretical analysis and a completeness proof for the ontology [105].

Depending on the application to which ontology-based research is applied, we can derive quality criteria, some of which are illustrated in Table 2. The heterogeneity of ontology-based applications prevents the application of a single quality and evaluation criterion. Instead, research results in biomedical ontology must be evaluated in conjunction with a task to which this result is being applied. For example, instead of evaluating the quality of an ontology *O* that represents biological pathways, we have

Table 2: Opportunities for the quantitative evaluation of research results in applied ontology

Application	Possible evaluation methods	Description	Quantifiable result	Example
Establish 'community agreement' about meanings of terms in a vocabulary. In a domain in which terms can have different meanings based on the background of a researcher, an ontology is developed to provide a reference for 'particular' meanings of terms.	User-study	Multiple people perform a task, such as determining the occurrence of terms from an ontology in a manuscript, independently. The goal is to achieve a high agreement between annotators about the ontology terms that have occurred in the manuscript.	Percentage agreement, κ statistics	A study was performed to evaluate the agreement between expert curators of the GO and found 'that there is 39% chance of curators exactly interpreting the text and selecting the same GO term, a 43% chance that they will extract a term from new/different lineage and a 19% chance that they will annotate a term from the same GO lineage' [106].
'Annotate data consistently' across multiple databases, user communities or domains.	User-study	Multiple people annotate the same data set using an ontology. The goal is to achieve a high agreement in the resulting annotations.	Percentage agreement, κ statistics	A study was performed to evaluate GO annotation consistency between human and mouse. The authors find that, out of a set of 3359 annotations, 2137 are matches and 1222 are mismatches (and potential annotation inconsistencies) [107].
'Integrate multiple databases' and provide a uniform view across.	User study, integrated analysis	An evaluation can perform an analysis of an integrated data set, or compare the integration results to a gold standard. Integrated analysis results can either be compared to a reference or tested based on a scientific use case.	Integrated data analysis results, precision, recall, F-measure	The phenotype data contained in multiple model organism databases were integrated and utilized for the task of prioritization of candidate genes for a disease. The results were compared against gene–disease associations in the OMIM database (gold standard) and quantitatively evaluated using ROC analysis [68].
'Answer queries over data' using the ontology as the conceptual model of a database (i.e. the classes and relations in the ontology are used to structure the database and functions as a vocabulary based on which queries can be built).	Test suite, comparison to gold standard	An evaluation can be based on a test suite (in which particular queries and the desired results are specified) or a gold standard, and use the ontology to perform test queries over the database and determine if the results conform to the desired outcome or compare the results to the gold standard. Additionally, a performance analysis can be used to determine the time and space required to implement the queries.	Number of tests passed, precision, recall, F-measure; complexity class, performance measurements	A study implemented an RDF-based query system over biomedical ontologies together with several relation axioms, demonstrating several queries that could not be answered before. The evaluation found that 'the answers to such a query are complete and they correspond to the logical meaning of the relation types as intended by the ontology engineers' [108].

(continued)

Table 2: Continued

Application	Possible evaluation methods	Description	Quantifiable result	Example
'Answer questions' over the knowledge contained in the ontology.	Test suite, content evaluation	Evaluation can take several directions. A test suite of questions can be designed, the ontology used to answer these questions, and the results compared to the outcome. Furthermore, the content of the ontology can be evaluated similarly to database content evaluation [109].	Number of tests passed, domain coverage (percentage), currency (number of times updated), expert evaluation	A study evaluated whether questions about existential restrictions in biomedical ontologies are correct as judged by experts in the field. The results show that, '[a]ccording to a rating done by four experts, 23% of all existential restrictions in OBO Foundry candidate ontologies are suspicious (Cohens' $\kappa = 0.78$)' [94].
Determine 'consistency of data' with respect to constraints in the ontology.	Test suite, performance measurement	A test suite of different types of data inconsistencies can be designed, and a performance evaluation used to measure the time and space complexity for identifying inconsistencies.	Number of tests passed (contradictions found); complexity class	A top-level ontology of computation models in systems biology (consisting of less than 10 classes and less than 10 relations) was formalized in OWL and the models in the BioModels database [110] were verified with regard to this ontology. As a consequence, the study detects several contradictions in the BioModels knowledge base, arising from annotations in 27 models [58].
Determine the 'consistency and accuracy of the conceptual model'.	Automated reasoning, test suite	Automated reasoning can be used to determine model consistency, and a test suite can be used to test accuracy of consequences following from the model.	Number of tests passed, number of inconsistencies found	A project to formalize the definitions of GO terms [57] has detected 7397 unsatisfiable classes in GOs definitions, 3487 in MPs definitions and 1017 in HPOs definitions. For example, 'system process' and 'cellular process' were declared as disjoint classes, but 'leukocyte activation' was inferred to be both a subclass of 'cellular process' and of (immune) 'system process' [60].
Enable 'novel scientific analyses', such as Gene Set Enrichment Analysis (GSEA) or semantic similarity, that rely on the type and the number of distinctions made in an ontology to analyse a data set.	Case-specific scientific validation	Evaluation must be based on the specific scientific problem and the standard established for the particular scientific discipline. An example for an evaluation could be to perform an experiment.	Various quantifiable results, including p-value, F-measure, ROC AUC	The novel method GSEA was proposed, that utilizes the annotations and the structure of GO to interpret gene expression data [23]. Results of GSEA were compared to published results and experimentally validated.

to evaluate O with respect to different tasks that it is intended to perform. For example, O may be used to achieve community agreement about the terms used to annotate pathway databases (task t_1), and we can evaluate O with respect to t_1. On the other hand, O may also be used to verify the consistency of biological pathway data (task t_2), and we can evaluate O with respect to t_2. A consequence could be that O achieves one task very well while its performance in a second task is poor.

'Robustness' can then be evaluated based on evaluating an ontology (or another research result in biomedical ntology) on multiple tasks: if the ontology performs well in multiple heterogeneous tasks, the ontology is 'robust'. Additionally, it becomes possible to evaluate how much the quantitative results change under changing application conditions.

ONTOLOGY PEER REVIEW

Several evaluation methods for research in applied ontology have been proposed, and multiple studies have attempted to evaluate the quality of ontologies in biomedicine. Currently, there is little emphasis on the need for objective, quantitative evaluation criteria for applied ontology research; on the contrary, many quality criteria are derived from philosophical and social considerations. In particular, several studies emphasize the need to treat ontologies similarly to scientific publications and propose an evaluation strategy similar to scientific peer review. For example, Obrst *et al.* [4] aim to identify 'meaningful, theoretically grounded units of measure in [ontology]' and perform an extensive review of previous ontology evaluation attempts, including a brief discussion of application-based evaluation approaches and quantifiable results. However, Obrst *et al.* dismiss application-based evaluation strategies since they are 'expensive to carry out', and instead propose ontology evaluation by humans based on principles derived from common sense, formal logics or philosophy (especially in the form of philosophical realism). A similar route is being taken by Smith who suggests that peer review of ontologies should become standard practice, since '[p]eer review provides an impetus to the improvement of scientific knowledge over time' [1, 5]. The criteria for peer review of ontology-based research results proposed by Smith [5], Orbst *et al.* [4] and others [111], are largely derived from 'common sense' or philosophical positions and do not rely on an objective, empirical demonstration that the criteria improve the performance of ontologies in any biomedical application. Such a peer review system is intended to be adopted by the OBO Foundry ontology community [1, 5].

The OBO Foundry principles (accepted and proposed principles can be found on http://obofoundry.org/crit.shtml) form some of the most widely used criteria for ontology development in biology. The majority of the OBO Foundry criteria are intrinsically social and highly valuable for enabling wide access to the content of the ontologies, serving scientific discourse about and investigations into the ontologies and their content. To evaluate ontologies based on social criteria, peer review is valuable. Some criteria could be further extended by asking for empirical, quantifiable evidence. For example, while the inclusion of textual definitions (criterion 6) and documentation (criterion 8) can improve comprehensibility of ontologies, comprehensibility will primarily depend on the quality of the textual definitions and documentation: not all definitions and all documentations are equally well suited. User-studies can be used to evaluate and quantify the quality of the definitions and even compare them against automated methods to generate textual definitions [112].

BUILDING AND EVALUATING ONTOLOGIES FOR INTEGRATIVE RESEARCH

The development of a systematic evaluation strategy grounded in real biomedical data will help to further improve the utility of ontologies for integrative biomedical research. To develop such a strategy, different approaches for evaluating ontologies can be combined. The direct evaluation of ontologies (see Figure 1), such as facilitated by ontology peer review, is an approach for evaluating ontologies that ensures availability of the ontologies, compliance with good scientific practices and reporting standards, the use of standard formats in distributing ontologies, and other valuable criteria.

However, a review of an ontology alone does not immediately evaluate the ontology's suitability for particular applications and analyses. Therefore, ontology evaluation can be further substantiated by an application-based evaluation (see Figure 2). In such an evaluation, an ontology is not assessed directly, but rather by means of an application that makes use of the ontology. Depending on what type of application the ontology is used for, a large variety of evaluation criteria can be applied to report, compare and quantify the results. Some of these criteria are listed in Table 2.

One major type of application in a research setting is to facilitate the integrated analysis of scientific data. In such a scenario, the 'application' that is used to evaluate an ontology is an integrated scientific analysis (see Figure 3). Evaluation criteria in such a scenario follow the established criteria in the scientific domain and range from comparisons with a gold standard to experimental validation.

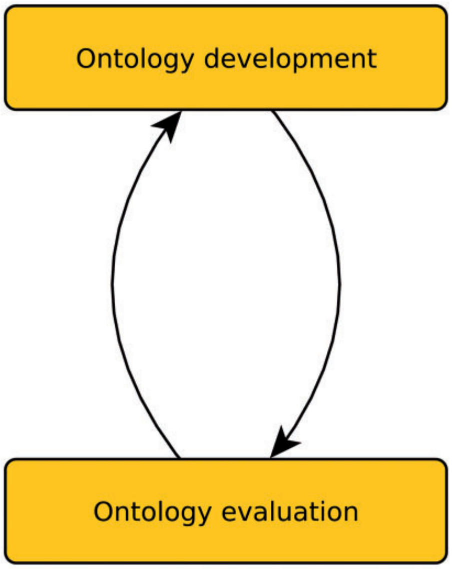

Figure 1: A direct evaluation of an ontology can assess intrinsic properties of the ontology such as consistency, expressivity, or the inclusion of natural language definitions and labels. Furthermore, the evaluating person can examine definitions and axioms of the ontology and either agree or disagree with their content.

These three strategies for evaluating research in ontology are complementary and ensure different aspect of an ontology's quality. Peer review can assure social criteria as well as adherence to scientific reporting standard, application-based evaluation ensures that ontologies can be used efficiently and the evaluation using a scientific analysis ensures that ontologies lead to verifiable novel insights in science. An adoption of this combined evaluation methodology shifts the research focus in ontology research from building better ontologies towards systematically improving the ontologies with regard to ontology-based applications and ontology-based scientific analyses, and thereby paves the way for the critical role that ontologies will continue to play in the future.

In particular, an area that will benefit from integrated ontology-based data analysis methods include experiment design [113]. In experiment design, ontologies can be used to relate experimental assays to the biological phenomena that are recorded by the assay catering for the experiments to be then designed so that they can test specific hypotheses about the scientific domain [114].

Furthermore, ontologies are now being used to annotate large data sets, including those originating from high-throughput technologies in all areas of biology, and it is a major challenge of biology to synthesize the available information into an understanding of whole organisms and their interactions

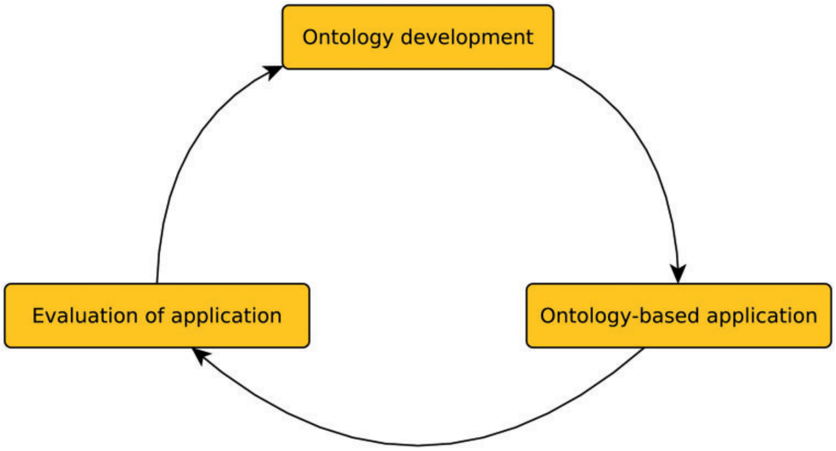

Figure 2: An application-based evaluation does not directly assess an ontology, but rather evaluates an application that utilizes an ontology for its operations.

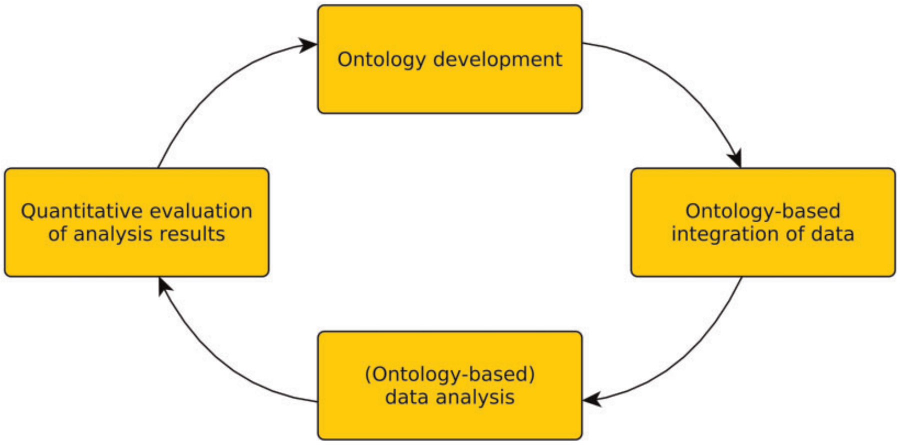
Ontology development
Ontology-based integration of data
(Ontology-based) data analysis
Quantitative evaluation of analysis results

Figure 3: An analysis-based evaluation performs a scientific data analysis that relies on an ontology and evaluates the success of the analysis using criteria established in the scientific domain.

with the environment as well as to transform this information into knowledge that can benefit human health. Ontologies will play a crucial role in this integration process because they provide the means to integrate data not only within domains, but also across domains, across species and across levels of granularity.

For example, personalizing the treatment of disease based on the background of the individual patient requires integration of large amounts of data across domains, including information about genetic variation and their associations with phenotypes and drug response [115], genomic, transcriptomic, proteomic, metabolomic, and autoantibody information [116, 117], environmental factors [118], and the patient's medical history [119]. Another example in which ontologies will increasingly be applied is to bridge the gap between basic research results and clinical applications. In the last years, several pioneering studies used ontologies as a means to understand, diagnose and find treatment strategies for human diseases [68, 120–122]. Again, it is the potential of ontologies to connect data from different scientific domains and disciplines on a large scale that has enabled such analyses, and it is one of the most promising future applications of ontologies in biomedicine.

One of the great challenges in using ontologies to facilitate integrative, translational biomedical analyses is to connect ontologies that cover basic research domains, such as the GO [11], with medical ontologies, such as SNOMED CT [123] and the repository of ontologies that are within the Unified Medical Language System (UMLS) [124]. The medical ontologies provide access to data in health care and medical knowledge while the biological ontologies enable access to findings from basic research in biology, and the integration of both types of ontologies has the potential to enable analyses that connect basic research with clinical applications and support the personalization of medical treatment. A strategy for evaluating the ontologies involved in such a task, as well as assessing the ontology-based integration results, is a crucial step to facilitate this goal.

CONCLUSIONS

Research results in biomedical ontology should always be evaluated against a biomedical task for which the ontologies are intended. Whether the research result is an ontology, an ontology design pattern, or a method to formulate biomedical phenomena, the benefit ontologies can bring cannot be evaluated based on the ontology alone; instead, any evaluation criteria must evaluate the whole system consisting of the ontology and the tasks to which they are applied. Many ontology-based applications are amenable to quantitative evaluation criteria. Quantitative measures enable the objective comparison of research results and play a crucial role in their evaluation. These quantitative measures can be adopted in addition to already established qualitative evaluation criteria, and they can also serve to justify and refine existing qualitative measures. Furthermore, with the application of quantitative measures, ontology development methodologies can be evaluated with respect to how well they ensure or improve the performance of research results in particular tasks within a domain. More importantly, objective evaluation criteria for research results are the next step in developing a research methodology for the field of biomedical ontologies. A research methodology based on quantitative evaluation with respect to biomedical applications will improve the ontologies' utility in data and knowledge integration and thereby increase their potential to improve integrative biology and translational research.

Key Points

- Ontologies are used in biomedicine to standardize terminology, to enable access to domain knowledge, to verify data consistency and to facilitate integrative analyses over heterogeneous biomedical data.
- Biomedical ontologies must be evaluated with respect to the purpose for which they are built.
- Ontology-based applications can be evaluated quantitatively.
- Quantitative evaluation can lead to developing a methodology for systematically improving biomedical ontologies.

ACKNOWLEDGEMENTS

We are grateful to Prof. Heinrich Herre, who contributed to this work over many years through discussions and critical remarks.

FUNDING

Funding for RH was provided by the European Commission's 7th Framework Programme, RICORDO project, grant number 248502. Funding for MD was provided by a National Sciences and Engineering Research Council of Canada Discovery Grant. Funding for GVG was provided by the National Institutes of Health, grant number R01 HG004838-02. The open access publication was funded by the European Commission's 7th Framework Programme, RICORDO project, grant number 248502.

References

1. Smith B, Ashburner M, Rosse C, *et al*. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech* 2007;**25**:1251–5.
2. Merrill GH. Ontological realism: methodology or misdirection? *Appl Ontol* 2010;**5**:79–108.
3. Smith B, Ceusters W. Ontological realism: a methodology for coordinated evolution of scientific ontologies. *Appl Ontol* 2010;**5**:139–88.
4. Obrst L, Ceusters W, Mani I, *et al*. The Evaluation of Ontologies: Toward Improved Semantic Interoperability. In: Baker CJO, Cheung K-H, (eds). *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences* 2007. Springer Science+Business Media, New York, NY, USA.
5. Smith B. *Proceeding of the 2008 Conference on Formal Ontology in Information Systems: Proceedings of the Fifth International Conference (FOIS 2008)*. pp. 21–35. IOS Press, Amsterdam, The Netherlands, 2008.
6. Rector AL, Qamar R, Marley T. Binding ontologies and coding systems to electronic health records and messages. *Appl Ontol* 2009;**4**:51–69.
7. Rogers J, Rector AL. Terminological Systems: bridging the generation gap. *Proceedings of the AMIA Annual Fall Symposium* 1997;**1997**:610–4.
8. Fung K, Bodenreider O. Knowledge representation and ontologies. In: Richesson R, Andrews J, (eds). *Clinical Research Informatics*. London, UK: Springer Verlag, 2012; 255–75.
9. Nohama P, Pacheco E, Andrade R, *et al*. Quality issues in thesaurus building: a case study from the medical domain. *Braz J Biomed Eng* 2012;**28**:11–22.
10. Freimer N, Sabatti C. The human phenome project. *Nat Genet* 2003;**34**:15–21.
11. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. *Nat Genet* 2000; **25**:25–9.
12. Robinson GE, Hackett KJ, Purcell-Miramontes M, *et al*. Creating a buzz about insect genomes. *Science* 2011;**331**: 1386.
13. Genome 10K Community of Scientists. Genome 10K: a proposal to obtain whole-genome sequence for 10000 vertebrate species. *J Hered* 2009;**100**:659–74.
14. Collins FS, Finnell RH, Rossant J, *et al*. A new partner for the international knockout mouse consortium. *Cell* 2007; **129**:235.
15. Skarnes WC, Rosen B, West AP, *et al*. A conditional knockout resource for the genome-wide study of mouse gene function. *Nature* 2011;**474**:337–42.
16. Parkinson H, Sarkans U, Kolesnikov N, *et al*. ArrayExpress update: an archive of microarray and high-throughput sequencing-based functional genomics experiments. *Nucleic Acids Res* 2011;**39**:D1002–4.
17. Munos B. Lessons from 60 years of pharmaceutical innovation. *Nat Rev Drug Discov* 2009;**8**:959–68.
18. Ekins S, Williams AJ, Krasowski MD, *et al*. In silico repositioning of approved drugs for rare and neglected diseases. *Drug Discov Today* 2011;**16**:298–310.
19. Horrocks I. *OBO Flat File Format Syntax and Semantics and Mapping to OWL Web Ontology Language Tech. Rep.* Manchester, UK: University of Manchester, 2007. http://www.cs.man.ac.uk/~horrocks/obo/.
20. Golbreich C, Horrocks I. The OBO to OWL mapping, GO to OWL 1.1!. In: Golbreich C, Kalyanpur A, Parsia B, (eds). *Proceedings of OWL: Experiences and Directions 2007 (OWLED-2007)*. Aachen, Germany: CEUR-WS.org, 2007.
21. Smith B, Ceusters W, Klagges B, *et al*. Relations in biomedical ontologies. *Genome Biol* 2005;**6**:R46.
22. Hoehndorf R, Oellrich A, Dumontier M, *et al*. Relations as patterns: bridging the gap between OBO and OWL. *BMC Bioinformatics* 2010;**11**:441+.
23. Subramanian A, Tamayo P, Mootha VK, *et al*. Gene set enrichment analysis: a knowledge-based approach for

interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005;**102**:15545–50.

24. Prufer K, Muetzel B, Do H-H, *et al*. FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* 2007;**8**:41+.
25. Mootha VK, Lindgren CM, Eriksson K-F, *et al*. PGC-1alpharesponsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003;**34**:267–73.
26. LePendu P, Musen M, Shah N. Enabling enrichment analysis with the human disease ontology. *J Biomed Inform* 2011; **44(Suppl 1)**:S31–8.
27. Resnik P. Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J Artif Intell Res* 1999;**11**: 95–130.
28. Du Plessis L, Kunca N, Dessimoz C. The what, where, how and why of gene ontology: a primer for bioinformaticians. *Brief Bioinform* 2011;**12**:723–35.
29. Benabderrahmane S, Smail-Tabbone M, Poch O, *et al*. IntelliGO: a new vector-based semantic similarity measure including annotation origin. *BMC Bioinformatics* 2010;**11**: 588.
30. Schlicker A, Lengauer T, Albrecht M. Improving disease gene prioritization using the semantic similarity of Gene Ontology terms. *Bioinformatics* 2010;**26**:i561–7.
31. Guzzi PH, Mina M, Guerra C, *et al*. Semantic similarity analysis of protein data: assessment with biological features and issues. *Brief Bioinform* 2011, in press.
32. Bodenreider O, Stevens R. Bio-ontologies: current trends and future directions. *Brief Bioinform* 2006;**7**: 256–74.
33. Lord PW, Stevens RD, Brass A, *et al*. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 2003; **19**:1275–83.
34. Rubin DL, Shah NH, Noy NF. Biomedical ontologies: a functional perspective. *Brief Bioinform* 2008;**9**:75–90.
35. Andronis C, Sharma A, Virvilis V, *et al*. Literature mining, ontologies and information visualization for drug repurposing. *Brief Bioinform* 2011;**12**:357–68.
36. Blaschke C, Leon EA, Krallinger M, *et al*. Evaluation of BioCreAtIvE assessment of task 2. *BMC Bioinformatics* 2005;**6(Suppl. 1)**:S16.
37. Leser U, Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. *Brief Bioinform* 2005;**6**:357–69.
38. Doms A, Schroeder M. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res* 2005;**33**:783–6.
39. Yao L, Divoli A, Mayzus I, *et al*. Benchmarking ontologies: bigger or better? *PLoS Comput Biol* 2011;**7**:e1001055.
40. Uniprot C. The Universal Protein Resource (UniProt). *Nucleic Acids Res* 2007;**35**:D193–7.
41. Bodenreider O, Hayamizu TF, Ringwald M, *et al*. Of mice and men: aligning mouse and human anatomies. *AMIA Annu Symp Proc* 2005;**2005**:61–65.
42. Dameron O, Rubin D, Musen M. Challenges in converting frame-based ontology into OWL: the foundational model of anatomy case-study. *AMIA Annu Symp Proc* 2005;**2005**:181.
43. Rosse C, Mejino JLV. A reference ontology for biomedical informatics: the foundational model of anatomy. *J Biomed Inform* 2003;**36**:478–500.
44. Haendel M, Neuhaus F, Osumi-Sutherland D, *et al*. CARO-the common anatomy reference ontology. In*: Anatomy Ontologies for Bioinformatics: Principles and Practice*. London, UK: Springer Verlag, 2007.
45. Lee RYN, Sternberg PW. Building a cell and anatomy ontology of Caenorhabditis elegans. *Comp Funct Genomics* 2003;**4**:121–6.
46. Mungall C, Torniai C, Gkoutos G, *et al*. Uberon, an integrative multispecies anatomy ontology. *Genome Biol* 2012; **13**:R5.
47. Segerdell E, Bowes J, Pollet N, *et al*. An ontology for Xenopus anatomy and development. *BMC Dev Biol* 2008; **8**:92.
48. Keseler IM, Collado-Vides J, Santos-Zavaleta A, *et al*. EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic Acids Res* 2011;**39**:D583–90.
49. Grau B, Horrocks I, Motik B, *et al*. OWL 2: the next step for OWL. *Web Semant Sci Serv Agent World Wide Web* 2008;**6**: 309–22.
50. Hoehndorf R, Kelso J, Herre H. The ontology of biological sequences. *BMC Bioinformatics* 2009;**10**:377+.
51. Hoehndorf R, Batchelor C, Bittner T, *et al*. The RNA Ontology (RNAO): An ontology for integrating RNA sequence and structure data. *Appl Ontol* 2011;**6**: 53–89.
52. Hastings J, Ceusters W, Smith B, *et al*. In *Proceedings of the 7th International and Interdisciplinary Conference on Modeling and Using Context*. Karlsruhe, Germany: Springer-Verlag, 2011; 119–123.
53. Hilbert D. Axiomatisches Denken. *Mathematische Annalen* 1918;**78**:405–15.
54. Barwise J, Etchemendy J. *Language, Proof and Logic*. Center for the Study of Language and Inf, 2002.
55. Hastings J, Batchelor CR, Steinbeck C, *et al*. What are chemical structures and their relations? In: Formal Ontology in Information Systems: Proceedings of the Sixth International Conference (FOIS 2010). Amsterdam, The Netherlands: IOS Press, 2010;257–70.
56. Demir E, Cary MP, Paley S, *et al*. The BioPAX community standard for pathway data sharing. *Nat Biotechnol* 2010;**28**: 935–42.
57. Mungall CJ, Bada M, Berardini TZ, *et al*. Cross-product extensions of the gene ontology. *J Biomed Inform* 2011;**44**: 80–6.
58. Hoehndorf R, Dumontier M, Gennari JH, *et al*. Integrating systems biology models and biomedical ontologies. *BMC Syst Biol* 2011;**5**:124+.
59. Ruttenberg A, Clark T, Bug W, *et al*. Advancing translational research with the semantic web. *BMC Bioinformatics* 2007;**8**:S2+.
60. Hoehndorf R, Dumontier M, Oellrich A, *et al*. Interoperability between biomedical ontologies through relation expansion, upper-level ontologies and automatic reasoning. *PLOS One* 2011;**6**:e22006.
61. Mungall CJ. *OBO Flat File Format 1.4 Syntax and Semantics [DRAFT] tech. rep*. Lawrence Berkeley National Laboratory, 2011. http://berkeleybop.org/~cjm/obo2owl/obo-syntax.html.

62. Goble C, Stevens R. State of the nation in data integration for bioinformatics. *J Biomed Inform* 2008;**41**:687–93.

63. Bada M, Stevens R, Goble C, *et al*. A short study on the success of the Gene Ontology. *Web Semant Sci Serv Agents World Wide Web* 2004;**1**:235–40.

64. Rector AL. Modularisation of domain ontologies implemented in description logics and related formalisms including OWL. In *K-CAP '03: Proceedings of the 2nd International Conference on Knowledge Capture*. New York, NY, USA: ACM Press, 2003;121–128.

65. Rector A. *Barriers, Approaches and Research Priorities for Integrating Biomedical Ontologies Tech. Rep*. Manchester, UK: University of Manchester, 2008.

66. Smith B, Williams J, Schulze-Kremer S. The ontology of the Gene Ontology. *AMIA Annu Symp Proc* 2003;**2003**: 609–613.

67. Schulz S, Stenzhorn H, Boeker M, *et al*. Strengths and limitations of formal ontologies in the biomedical domain. *RECIIS Electron J Commun Inf Innov Health* 2009; **3**:31–45.

68. Hoehndorf R, Schofield PN, Gkoutos GV. PhenomeNET: a whole-phenome approach to disease gene discovery. *Nucleic Acids Res* 2011;**39**:e119.

69. Mungall CJ, Batchelor C, Eilbeck K. Evolution of the sequence ontology terms and relationships. *J Biomed Inform* 2011;**44**:87–93.

70. Schindelman G, Fernandes J, Bastiani C, *et al*. Worm phenotype ontology: integrating phenotype data within and beyond the *C. elegans* community. *BMC Bioinformatics* 2011;**12**:32.

71. Hoehndorf R, Ngomo A-CN, Kelso J. Applying the functional abnormality ontology pattern to anatomical functions. *J Biomed Semant* 2010;**1**:4.

72. Hoehndorf R, Loebe F, Kelso J, *et al*. Representing default knowledge in biomedical ontologies: application to the integration of anatomy and phenotype ontologies. *BMC Bioinformatics* 2007;**8**:377.

73. Rohl J, Jansen L. Representing dispositions. *J Biomed Semant* 2011;**2**:S4.

74. Schulz S, Suntisrivaraporn B, Baader F, *et al*. SNOMED reaching its adolescence: Ontologists' and logicians' health check. *Int J Med Inform* 2009;**78**:S86–94.

75. Schulz S, Stenzhorn H, Boeker M. The ontology of biological taxa. *Bioinformatics* 2008;**24**:i313.

76. Wolstencroft K, Lord P, Tabernero L, *et al*. Protein classification using ontology classification. *Bioinformatics* 2006;**22**: e530–8.

77. Kohler S, Bauer S, Mungall C, *et al*. Improving ontologies by automatic reasoning and evaluation of logical definitions. *BMC. Bioinformatics* 2011;**12**:418.

78. Kazakov Y, Kroetzsch M, Simancik F. Consequence-Driven Reasoning for Horn SHIQ Ontologies. In *Proceedings of the 23rd International Workshop on Description Logics (DL'10)*. Aachen, Germany: CEUR-WS.org, 2011.

79. Jupp S, Stevens R, Hoehndorf R. Logical Gene Ontology Annotations (GOAL): exploring gene ontology annotations with OWL. *J Biomed Semant* 2012;**3**:S3.

80. Tweedie S, Ashburner M, Falls K, *et al*. FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res* 2009;**37**:D555–9.

81. Sirin E, Parsia B. Pellet: An OWL DL Reasoner. In: Haarslev V, Moeller R, (eds). *Proceedings of the 2004 International Workshop on Description Logics, DL2004*. Whistler, British Columbia, Canada, 104 CEUR-WS.org, Aachen, Germany, 2004.

82. Noy NF, Shah NH, Whetzel PL, *et al*. BioPortal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res* 2009;**37**:W170–3.

83. Gangemi A. Ontology Design Patterns for Semantic Web Content. In *Proceedings of the Fourth International Semantic Web Conference*. Berlin: Springer-Verlag, 2005;262–76.

84. Gkoutos GV, Green EC, Mallon A-MM, *et al*. Using ontologies to describe mouse phenotypes. *Genome Biol* 2005;**6**:R8.

85. Burek P, Hoehndorf R, Loebe F, *et al*. A top-level ontology of functions and its application in the open biomedical ontologies. *Bioinformatics* 2006;**22**:e66–73.

86. Mungall C, Gkoutos G, Smith C, *et al*. Integrating phenotype ontologies across multiple species. *Genome Biol* 2010; **11**:R2+.

87. Hoehndorf R, Oellrich A, Rebholz-Schuhmann D. Interoperability between phenotype and anatomy ontologies. *Bioinformatics* 2010;**26**:3112–8.

88. Searle JR. *The Construction of Social Reality*. London, UK: Penguin Group, 1995.

89. Wright L. Functions. *Philos Rev* 1973;**82**:139–168.

90. Millikan RG. *Language, Thought, and Other Biological Categories: New Foundations for Realism*. Cambridge, MA, USA: MIT Press, 1988.

91. Guarino N, Welty CA. An overview of OntoClean. In: Staab S, Studer R, (eds). *Handbook on Ontologies*. Berlin, Germany: Springer Verlag, 2004;151–72.

92. Kutz O, Mossakowski T. A modular consistency proof for DOLCE. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence,* 2011.

93. Smith B, Ceusters W, Temmerman R. Wuesteria. *Stud Health Technol Inform* 2005;**116**:647–52.

94. Boeker M, Tudose I, Hastings J, *et al*. Unintended consequences of existential quantifications in biomedical ontologies. *BMC Bioinformatics* 2011;**12**:456.

95. Schober D, Smith B, Lewis S, *et al*. Survey-based naming conventions for use in OBO Foundry ontology development. *BMC Bioinformatics* 2009;**10**:125.

96. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measure* 1960;**20**:37–46.

97. Artstein R, Poesio M. Inter-coder agreement for computational linguistics. *Comput Linguist* 2008;**34**:555–96.

98. Cohen KB, Hunter LE, Palmer M. In *Proceedings of the 5th Linguistic Annotation Workshop*, pp. 82–91. Association for Computational Linguistics, Portland, Oregon, 2011.

99. Jourde J, Manine A-P, Veber P, *et al*. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pp. 65–73. Association for Computational Linguistics, Portland, Oregon, 2011.

100. Stevenson M, Guo Y. Disambiguation in the biomedical domain: the role of ambiguity type. *J Biomed Inform* 2010; **43**:972–81.

101. Van Rijsbergen CJ. *Information Retrieval*. London, UK: Butterworths, 1979.

102. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006;**27**:861–874.

103. Xu T, Du L, Zhou Y. Evaluation of GO-based functional similarity measures using S. cerevisiae protein interaction and expression profile data. *BMC Bioinformatics* 2008;**9**:472.

104. Huizinga D, Kolawa A. *Automated Defect Prevention: Best Practices in Software Management*. Hoboken, New Jersey, USA: Wiley-Interscience, 2007.

105. Baumann R, Loebe F, Herre H. Ontology of time in GFO. In *Formal Ontology in Information Systems: Proceedings of the Seventh International Conference,*. Amsterdam, The Netherlands: IOS Press, 2012;293–306.

106. Camon E, Barrell D, Dimmer E, *et al*. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinformatics* 2005;**6**:S17.

107. Dolan ME, Ni L, Camon E, *et al*. A procedure for assessing GO annotation consistency. *Bioinformatics* 2005;**21**: i136–43.

108. Blond W, Mironov V, Venkatesan A, *et al*. Reasoning with bio-ontologies: using relational closure rules to enable practical querying. *Bioinformatics* 2011;**27**:1562–8.

109. Jacso P. Content evaluation of databases. *Annu Rev Inform Sci Technol* 1997;**32**:231–67.

110. Li C, Donizelli M, Rodriguez N, *et al*. BioModels Database: an enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol* 2010;**4**:92+.

111. Jansen L, Schulz S. The Ten Commandments of Ontological Engineering. In: Herre H, Hoehndorf R, Loebe F, (eds). *OBML 2011 Workshop Proceedings*. Leipzig, Germany: Markus Loeffler, 2011. p G-11.

112. Stevens R, Malone J, Williams S, *et al*. Automating generation of textual class definitions from OWL to English. *J Biomed Semantics,* 2011;**2(Suppl 2)**:S5.

113. Brown S, Chambon P, de Angelis MH, *et al*. EMPReSS: standardized phenotype screens for functional annotation of the mouse genome. *Nat Genet* 2005;**37**:1155.

114. Gkoutos GV, Green EC, Mallon AM, *et al*. Building mouse phenotype ontologies. *Pac Symp Biocomput* 2004; 178–89.

115. Hernandez-Boussard T, Whirl-Carrillo M, Hebert JM, *et al*. The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. *Nucleic Acids Res* 2008;**36**:D913–D918.

116. Chen R, Mias GI, Li-Pook-Than J, *et al*. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* 2012;**148**:1293–307.

117. Samwald M, Coulet A, Huerga I, *et al*. Semantically enabling pharmacogenomic data for the realization of personalized medicine. *Pharmacogenomics* 2012;**13**:201–12.

118. Patel CJ, Bhattacharya J, Butte AJ. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS One* 2010;**5**:e10746.

119. Barabasi A-LL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet* 2011;**12**:56–68.

120. Gottlieb A, Stein GY, Ruppin E, *et al*. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol Syst Biol* 2011;**7**.

121. Tatonetti NP, Ye PP, Daneshjou R, *et al*. Data-driven prediction of drug effects and interactions. eng. *Sci Transl Med* 2012;**4**:125ra31.

122. Washington NL, Haendel MA, Mungall CJ, *et al*. Linking human diseases to animal models using ontology-based phenotype annotation. *PLoS Biol* 2009;**7**:e1000247.

123. Cornet R, de Keizer N. Forty years of SNOMED: a literature review. *BMC Med Inform Decis Mak* 2008;**8**:S2.

124. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;**32**:D267–D270.