Journal of Biomedical Informatics 44 (2011) 3–7

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Guest Editorial

Ontologies for clinical and translational research: Introduction

1. Background

Ontologies are representational artifacts that are being used in many different ways by researchers in almost every life science discipline. Their use in the annotation of both clinical and experimental data is now a common approach for knowledge representation in support of integrative translational research. When high quality ontologies are used correctly and consistently, then the description of the relevant entities and the semantic framework for capturing their relationships support not only the retrieval and integration of data, but also algorithmic reasoning, and semantic enhancement of the published literature and database records. The expanded use of ontologies is therefore being encouraged by recent mandates promulgated by the National Institutes of Health (NIH) and other funding agencies requiring researchers to provide plans to ensure reusability of the data deriving from funded research.

This special issue on Ontologies for Clinical and Translational Research focuses on ways in which ontologies can contribute to breaking down the barriers between different sorts of information relevant to the understanding and treatment of disease, ranging from information deriving from experimental biology and model organism research to clinical trial data and information of the sort contained in electronic health records. It is clear that the development and use of well-formulated ontologies are still in their early stages; the manuscripts presented in this special issue represent both the state of the art and works in progress. We still have a considerable way to go to reach the level of domain coverage and semantic consistency sought by those engaged in information-driven clinical and translational research. Thus the inclusion in this issue of the *Journal of Biomedical Informatics* of papers discussing one or other ontology should not be seen as an unconditional endorsement for the use of the ontologies discussed.

From the combined experience of the groups reporting their results here and elsewhere, several common themes emerge that will be helpful in guiding future development and use of ontologies.

1. The Basic Formal Ontology (BFO) [http://www.ifomis.uni-saarland.de/bfo/] has been found to provide a valuable framework for the development of a wide range of different biomedical ontologies by over 75 biomedical ontologist groups. The initial classification of entities to be represented in an ontology into the three main axes of the BFO – independent continuants, dependent continuants and occurrents – establishes a consistent framework for the resulting subsumption hierarchy. This relatively simple step has had a dramatic impact on the overall quality and consistency of the resultant ontologies.
2. The principles proposed by the Open Biomedical Ontology (OBO) Foundry [http://obofoundry.org] have been found to be useful in providing guidance to ontology developers about best practices in ontology development and reducing the proliferation of overlapping ontologies. Rather than supporting the independent development of alternative ontologies covering the same biomedical domain, the OBO Foundry has encouraged the establishment of collaborative groups that come together to work on a common ontology for a given domain that is constructed to support the needs of all of the involved stakeholders [1]. In addition, this approach has allowed individuals in need of terms to represent entities that cross many different biomedical domains to contribute to the relevant ontology development projects in a way that ensures a high degree of cross-domain consistency.
3. It has become clear that a critical component of ontology-based annotation is the use of a consistent set of relations that establishes the semantic framework for knowledge representation. It has also become clear that in order to serve this function, the proliferation of new relations must be tightly controlled and coordinated. The Relation Ontology (RO) [2] has emerged as a critical OBO Foundry reference ontology for this purpose. By tightly controlling the number of relations, the RO can be incorporated into inferencing and other reasoning algorithms that can then traverse the resulting semantic network.
4. There is much value to be gained by reusing terms from existing ontologies in formulating logical definitions and compound terms, and by importing terms and definitions from existing ontologies using the "Mireoting" strategy [3]. This not only makes the process of covering new biomedical domains more efficient, it is also helpful in ensuring the coherence of the entire semantic network.

At the same time, several significant challenges have been noted by those faced with more sophisticated terminological needs. The first is the lack of complete domain coverage. While many groups of researchers may recognize the value of incorporating well-formulated ontologies into their representational framework, they must also determine how they will deal with the many sorts of entities that are not currently found in the relevant source ontologies. While the developers of existing OBO Foundry ontologies have committed themselves to be responsive to term requests submitted by members of user communities, provision of new terms can be a lengthy process, not only because the development and validation of accurate definitions for single terms is not always a simple process, but also because the need for careful

1532-0464/$ - see front matter
doi:10.1016/j.jbi.2011.01.002

positioning of new terms in the existing ontology hierarchy may imply the need for the creation of new sub-branches of additional new terms. In addition, resource limitations mean that for some biomedical domains ontologies do not yet exist. In many cases, therefore, a user will need to decide what kind of interim solution to adopt while waiting for acceptable domain coverage.

Finally, much of the current knowledge derived from biomedical research and biomedical informatics has been represented using legacy terminologies that did not incorporate a consistent semantic framework. Rather than discarding this legacy knowledge, most groups would prefer to adopt some kind of strategy for mapping between legacy vocabularies and the emerging ontologies. Unfortunately, few automated mapping strategies have been developed, making this a largely manual process that does not scale well to the current corpus of available knowledge. For the broad adoption of these new ontologies, it will be critical to develop better strategies for importing legacy information into more modern ontological frameworks that are both logically consistent and biologically coherent.

2. Contents of this issue

The manuscripts included in this special issue include work on both the development and use of ontologies. It has become clear that for successful ontology initiatives, development and use must go hand in hand. Ontologies created on the basis of a consistent logical framework must be tested and refined through large-scale application in data and literature annotation if they are to be of high utility in advancing data integration and reuse across a broad sweep of further applications, and if they are to support a variety of secondary uses not anticipated when the ontology was originally conceived.

In order to guide readers through this issue, we summarize the key characteristics of each paper below.

3. Original research

- "The ACGT Master Ontology and Its Applications – Towards an Ontology-Driven Cancer Research and Management System" [4].

Here Brochhausen et al. report on the development and use of an application ontology to support cancer research. The goal is to provide the semantics for a grid-based services infrastructure that will enable efficient execution of discovery-driven workflows in the context of multi-center, post-genomic clinical trials. The ontology is designed on the basis of the following principles: (1) adoption of a radically restrictive definition of the term 'ontology' in compliance with the principles of ontological realism; (2) enforcement of a strict subsumption hierarchy, based on a formally specified *is_a* relation; (3) avoidance of multiple inheritance in the hierarchy of universals; (4) avoidance of the types of confusions between ontology and epistemology illustrated by terms such as 'unknown X', 'unlocalized Y', and so forth; (5) use of BFO as the upper ontology; and (6) use of the OBO Relation Ontology (RO) for the semantic structure. A prior analysis revealed that SNOMED-CT, the NCI Thesaurus, and the UMLS, did not meet these requirements and therefore could not be used for the intended purpose. The ACGT ontology here presented is still marked by compromises designed to address 'clinical needs', but its authors are aware that work is still needed to ensure that these needs are addressed in a principled way that will allow consistency with other OBO Foundry candidate ontologies.

- "Towards an Ontological Theory of Substance Intolerance and Hypersensitivity" [5].

Here William Hogan outlines a realist approach to the ontology of substance intolerance. Building further on the Ontology for General Medical Science [6], he characterizes substance intolerance as a disease whose pathological processes are realized upon exposure to a quantity of substance of a particular type, and such that this quantity would normally not cause the realization of the pathological processes in question. His theory makes a careful distinction between a disposition to undergo particular processes, and the processes themselves, a considerable improvement over what is said in terminological artifacts such as SNOMED CT and MedDRA, which blur this important distinction. He also reviews and incorporates the three major axes on which these diseases are typically classified: the pathological process to which the organism is disposed, the location within the organism where the pathological process occurs, and the substance that induces the pathological process.

- "Towards an Ontological Representation of Resistance: The Case of MRSA" [7].

Here Goldfain et al. provide a characterization of the phenomenon of resistance in terms of what they call 'blocking dispositions,' by which they mean collections of mutually coordinated dispositions which are such that they cannot undergo simultaneous realization within a single bearer. The approach, which builds on BFO and on the Gene Ontology (GO), introduces some additional principles, of which the "nonproliferation of new relations and terms" is a most welcome alternative to the more common solutions in which relations are generated *ad libitum*, without serious ontological analysis. As the authors argue, it would indeed be easier to invent the relation "resistant to" and use it to describe every instance of a resistance phenomenon. But this would hide the complexity of the mechanisms of resistance working at a smaller scale, and eliminate many important inferences about resistance. The applicability of the approach is demonstrated by examples of drug resistance in Methicillin-Resistant Staphylococcus aureus (MRSA), HIV and malaria.

- "A Set of Ontologies to Drive Tools for the Control of Vector-Borne Diseases" [8].

This paper, by Pantelis Topalis and colleagues, describes work thus far on an ontology for vector-borne diseases based on the Infectious Disease Ontology (IDO). The paper describes a survey of the available resources and tools, and of some of the potential applications of such an ontology, particularly in the consistent design of disease databases in a way which can allow the construction of decision support systems (DSS) to control malaria, dengue, yellow fever and other diseases of global significance.

- "Toward an Ontology-Based Framework for Clinical Research Databases" [9].

In this paper Kong et al. describe a data model for clinical research data which is designed around the logical structure of the BFO and the Ontology for Biomedical Investigations (OBI). The model is designed to simplify the development of data dictionaries based on ontologies from the OBO Foundry. Existing clinical data standards, all centered around CDISC, were analyzed and found to fall short in several respects. The paper presents a practical application of OBO Foundry ontologies for the design of an extensible database schema to capture and manage data from a wide

range of different clinical and translational research projects supported by the US National Institute of Allergy and Infectious Diseases (NIAID).

- "NanoParticle Ontology for Cancer Nanotechnology Research" [10].

This paper by Nathan Andrew Baker discusses the design and development of an ontology relating to the preparation, chemical composition, and characterization of nanomaterials involved in cancer research. While the ontology is in part developed within the framework of the BFO, it does not yet satisfy all of the associated principles. For example, the authors employ many relations that do not conform to the rules set forth in the Relation Ontology (RO); there are also confusions between dependent and independent continuants, specifically at the level of realizable entities, disjointness of classes is not enforced, and so forth. This work serves nonetheless as an important first step towards the needed nanoparticle ontology, and it clearly illustrates the potential applications of such an ontology in supporting information-driven cancer research.

- "Hematopoietic Cell Types: Prototype for a Revised Cell Ontology" [11].

This paper, by Diehl et al., describes the major modifications that have been introduced into those portions of the Cell Ontology (CL) dealing with hematopoietic cells. This revision is part of a larger initiative to bring the CL up to current standards for biomedical ontologies, both in its structure and its coverage of various subfields of biology, to transform the ontology into an OBO Foundry reference ontology. The achievements obtained include the elimination of multiple inheritance in the asserted hierarchy and the groundwork for structuring the hematopoietic cell type terms as cross-products incorporating logical definitions built from relationships to external ontologies, such as the Protein Ontology and the Gene Ontology.

- "Cross-Product Extensions of the Gene Ontology" [12].

This paper by Mungall et al. provides preliminary results of ongoing work to normalize the GO by providing definitions for Gene Ontology (GO) terms in a logical form that can be used by reasoners. These definitions draw on a partitioning of terms into mutually exclusive sets, corresponding for example to the OBO Foundry candidate ontologies for chemical entities, proteins, biological qualities and anatomical entities. The advantage of these logical definitions is that they have the potential to allow the automation of many aspects of ontology development, of detecting errors and of filling in missing relationships. These definitions also enhance the GO by weaving it into the fabric of a wider collection of interoperating ontologies, increasing opportunities for data integration and enhancing genomic analyses. A novelty in the approach is that the traditional ontology development scenario, in which reasoners are used to infer the subsumption hierarchy of composite classes based on properties described in terms of simpler classes, is complemented by a form of inverse, abductive reasoning, in which inferences are drawn from the GO to referenced ontologies such as the CL or ChEBI. This makes it possible to find inconsistencies within the GO and between the GO and other ontologies, and to uncover a number of fundamental differences between classifications in ChEBI and in the implicit chemical entity ontology in GO.

- "Evolution of the Sequence Ontology Terms and Relationships" [13].

Here Mungall et al. report on recent improvements in the Sequence Ontology, focusing on new relationships included in the ontology in order to better define the mereological, spatial and temporal aspects of biological sequences. Although definitions for these new relationships are provided, these do not follow the format proposed by the RO; for instance they do not require that instance-level relationships between continuants should be time-indexed, thereby raising the question whether molecules, because of the ways we refer to them using chemical formulae, are governed by different rules governing changes such as gain and loss of parts from those which govern entities such as cells and organisms.

- "Desiderata for Ontologies to Be Used in Semantic Annotation of Biomedical Documents" [14].

In this paper Bada and Hunter report on their effort manually to annotate 97 full-text biomedical journal articles with terms derived predominantly from OBO ontologies. They argue that these ontologies contain infelicities with respect to their use in semantic annotation of biomedical documents, and propose desiderata whose implementation could, in their view, improve their utility for this purpose. The desiderata include integration of overlapping terms across OBO ontologies, the resolution of OBO-specific ambiguities, the integration of BFO with the OBO ontologies, the use of mid-level ontologies, the inclusion of non-canonical instances, and the expansion of relations and realizable entities. Their work demonstrates clearly the need for principles of the sort advocated by the OBO Foundry, adherence to which has the potential to avoid many of the problems they identify.

4. Applications

- vSPARQL: A View Definition Language for the Semantic Web" [15].

The paper by Marianne Shaw et al. describes a view definition language, vSPARQL, that allows for the specification of subsets of data/information (views) represented in RDF or OWL for access through semantic web technologies. In addition, vSPARQL also allows for the reorganization and modification of the source content to meet specific use cases not easily supported by the native data structures. The authors demonstrate the use of vSPARQL for the extraction and modification of data from the NCI Thesaurus, Reactome, Ontology of Physics for Biology and the Foundational Model of Anatomy to support a series of biological use cases (e.g. generate a liver anatomy sub-ontology from the FMA that excludes all relations other than is_a or part_of for use in the annotation of radiology images). Finally, the authors compare the use of vSPARQL with other existing RDF query languages against the defined requirements.

- "An Ontology-Based Measure to Compute Semantic Similarity in Biomedicine" [16].

This paper by Montserrat Batet and collaborators addresses the problem of automatic knowledge extraction from text, focusing on the issue of measuring semantic similarity between word pairs. It surveys existing approaches to this issue in the field of biomedicine, and proposes a new approach which uses the taxonomical structure of biomedical ontologies and vocabularies such as SNOMED CT. The proposal is shown to enhance accuracy in identifying semantically similar word pairs as compared to some existing approaches.

- "Using an ECG reference ontology for Semantic Interoperability of ECG data" [17].

Here Bernardo Gonçalves et al. test the hypothesis that a domain reference ontology for electrocardiograms can be used for semantic integration of ECG data standards (e.g. Physionet, SCP-ECG and HL7 aECG). Integration is achieved by mapping the individual representations to a common reference ontology developed by the authors in order to provide a consistent realism-based conceptualization of the entities of interest. The uses and advantages of this approach for data integration remain to be demonstrated.

- "The Biomedical Resource Ontology (BRO) to Enable Resource Discovery in Clinical and Translational Research" [18].

This paper by Tenenbaum et al. describes the development and use of the Biomedical Resource Ontology (BRO) and of associated software tools. The BRO is designed to facilitate semantically based search and discovery of funding, material, software, training and other resources for biomedical research. The ontology contains also a terminological component dealing with areas of research and with activities such as community engagement and device development.

- "Multiple Ontologies in Action: Composite Annotations for Biosimulation Models" [19].

Here Gennari et al. report on their proposal to use what they call "composite annotations" to access multiple ontologies in a way that will capture the physics-based meaning of model variables. They argue that these composite annotations can "provide the semantic expressivity needed to disambiguate the often-complex features of biosimulation models, and can be used to assist with model merging and interoperability". To that end, they provide a simple juxtaposition grammar and describe a tool based on this grammar which allows users to select terms from various ontologies that then are used as elementary building blocks for the desired composite annotations.

- "Ontology Modularization to Improve Semantic Medical Image Annotation" [20].

Here Pinar Wennerberg and colleagues describe an ontology-based strategy for image annotation that enables images and clinical reports to be linked via common annotations. In part because of the large size of clinical ontologes, the creation of such descriptions involves a considerable investment of effort. The authors propose a modularization strategy to address this problem, based on identification of ontology fragments relevant to particular sets of images, They illustrate this strategy by showing how it can identify terms in the Foundational Model of Anatomy [21] relevant for annotating medical images from patients suffering from lymphoma.

5. Methodological review

- "Natural Language Processing Methods and Systems for Biomedical Ontology Learning" [22].

This paper by Liu et al. is a thorough review of methods to address the increasingly pressing problems for clinical and translational ontologies that arise as a result of the use of manual, time-consuming, and often error-prone process methods of ontology development. The authors survey multiple techniques for automating the enrichment of an ontology from free-text documents. They conclude that, while fully automated acquisition of ontology by machines is not likely in the near future, there is potential value to be gained from semi-automatic ontology learning approaches that include human intervention.

6. Conclusion

Considerable progress toward the goal of the development and use of well-formulated ontologies has been made in the last decade. As the user community applies the resulting ontologies to addresses data annotation and data mining challenges, their experience will feed back to the ontology development community to further improve the structures represented in the relevant ontologies. While there is still a long way to go to realize this goal, we believe that the work reported here indicates that we are heading in the right direction.

Acknowledgments

The majority of the contributions to this issue are based on papers presented at the first International Conference on Biomedical Ontology held in Buffalo, NY in July 2009. We are grateful for the support of the National Institutes of Health and of the National Human Genome Research Institute through the NIH Roadmap for Medical Research Grant 1 U 54 HG004028 and through conference Grant 1 R13 HG005049-01. We also gratefully acknowledge the many helpful contributions provided by Janine Burch at all stages in the preparation of this special issue.

References

[1] Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnol 2007;25(11):1251–5. PMID: 17989687. PMC2814061.
[2] Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biol 2005;6(5):R46. PMC: 1175958.
[3] Courtot M, Gibson F, Lister A, Malone J, Schober D, Brinkman R, et al. MIREOT: the minimum information to reference an external ontology term. In: Proceedings of international conference on biomedical ontology. Nature proceedings; 2009. <http://precedings.nature.com/documents/3576>.
[4] Brochhausen M, Spear AD, Cocos C, Weiler G, Martìn L, Anguita A, et al. The ACGT master ontology and its applications – towards an ontology-driven cancer research and management system. J Biomed Inform 2011;44(1):8–25.
[5] Hogan WR. Towards an ontological theory of substance intolerance and hypersensitivity. J Biomed Inform 2011;44(1):26–34.
[6] Scheuermann RH, Ceusters W, Smith B. Toward an ontological treatment of disease and diagnosis. In: AMIA 2009 summit on translational bioinformatics proceedings; 2009. p. 116–20..
[7] Goldfain A, Smith B, Cowell LG. Towards an ontological representation of resistance: the case of MRSA. J Biomed Inform 2011;44(1):35–41.
[8] Topalis P, Dialynas E, Mitraka E, Deliyanni E, Siden-Kiamos I, Louis C. A set of ontologies to drive tools for the control of vector-borne diseases. J Biomed Inform 2011;44(1):42–7.
[9] Kong YM, Dahlke C, Xiang Q, Qian Y, Karp D, Scheuermann RH. Toward an ontology-based framework for clinical research databases. J Biomed Inform 2011;44(1):48–58.
[10] Baker NA. Nanoparticle ontology for cancer nanotechnology research. J Biomed Inform 2011;44(1):59–74.
[11] Diehl D, Deckhut -Augustine A, Blake JA, Cowell LG, Gold ES, Gondré-Lewis TA, et al. Hematopoietic cell types: prototype for a revised cell ontology. J Biomed Inform 2011;44(1):75–9.
[12] Mungall CJ, Bada M, Berardini TZ, Deegan J, Ireland A, Harris MA, et al. Cross-product extensions of the gene ontology. J Biomed Inform 2011;44(1):80–6.
[13] Mungall CJ, Batchelor C, Eilbeck K. Evolution of the Sequence Ontology terms and relationships. J Biomed Inform 2011;44(1):87–93.
[14] Bada M, Hunter L. Ontology desiderata for their use in semantic annotation of biomedical documents. J Biomed Inform 2011;44(1):94–101.
[15] Shaw M, Detwiler LT, Noy N, Brinkley J, Suciu D. vSPARQL: a view definition language for the semantic web. J Biomed Inform 2011;44(1):102–17.
[16] Batet M, Sanchez D, Valls A. An ontology-based measure to compute semantic similarity in biomedicine. J Biomed Inform 2011;44(1):118–25.
[17] Gonçalves B, Guizzardi G, Pereira Filho JG. Using an ECG reference ontology for semantic interoperability of ECG data. J Biomed Inform 2011;44(1):126–36.

[18] Tenenbaum JD, Whetzel P, Anderson K, Borromeo CD, Dinov ID, Gabriel D, et al. The Biomedical Resource Ontology (BRO) to enable resource discovery in clinical and translational research. J Biomed Inform 2011;44(1):137–45.
[19] Gennari J, Neal M, Galdzicki M, Cook D. Multiple ontologies in action: composite annotations for biosimulation models. J Biomed Inform 2011;44(1):146–54.
[20] Wennerberg P, Schulz K, Buitelaar P. Ontology modularization to improve semantic medical image annotation. J Biomed Inform 2011;44(1):155–62.
[21] Rosse C, Mejino JLV. The foundational model of anatomy ontology. In: Burger A et al., editors. Anatomy ontologies for bioinformatics: principles and practice. New York: Springer; 2007. p. 59–117.
[22] Liu K, Hogan WR, Crowley RS. Natural Language Processing methods and systems for biomedical ontology learning. J Biomed Inform 2011;44(1):163–79.

Barry Smith
University at Buffalo,
135 Park Hall, Buffalo, NY 14260,
United States
E-mail address: phismith@buffalo.edu

Richard H. Scheuermann
University of Texas Southwestern Medical Center,
5323 Harry Hines Blvd.
Dallas, TX 75390-9072,
United States
E-mail addresses: richard.scheuermann@utsouthwestern.edu