Agrawal and Cui *BMC Med Inform Decis Mak* 2020, **20**(Suppl 10):301
https://doi.org/10.1186/s12911-020-01342-4

BMC Medical Informatics and Decision Making

INTRODUCTION Open Access

Quality assurance and enrichment of biological and biomedical ontologies and terminologies

Ankur Agrawal[1] and Licong Cui[2*]

Abstract

Biological and biomedical ontologies and terminologies are used to organize and store various domain-specific knowledge to provide standardization of terminology usage and to improve interoperability. The growing number of such ontologies and terminologies and their increasing adoption in clinical, research and healthcare settings call for effective and efficient quality assurance and semantic enrichment techniques of these ontologies and terminologies. In this editorial, we provide an introductory summary of nine articles included in this supplement issue for quality assurance and enrichment of biological and biomedical ontologies and terminologies. The articles cover a range of standards including SNOMED CT, National Cancer Institute Thesaurus, Unified Medical Language System, North American Association of Central Cancer Registries and OBO Foundry Ontologies.

Keywords: Quality assurance, Auditing, Semantic enrichment, Mapping, Ontology

Background

Ontologies and terminologies provide structured and unambiguous ways of representing domain information in biology and biomedicine. Examples of such ontologies and terminologies include Gene Ontology [1], SNOMED CT [2], and National Cancer Institute Thesaurus (NCIt) [3]. They have served as knowledge sources for a wide range of biomedical applications including data integration and exchange, natural language processing, reasoning, and decision support [4, 5]. These ontologies and terminologies tend to be large and are regularly maintained through revisions and modifications in their lifecycle, which may result in ambiguity, redundancy and modeling inconsistencies. As such, quality assurance and enrichment of these ontologies and terminologies become more and more important as they impact all the downstream applications that rely on them, and thus an active research area. In [6], Zhu et al. provided a comprehensive review of early works on the auditing methods of biomedical terminologies regarding various quality factors. In [7], Amith et al. surveyed more recent quality assurance approaches for biomedical ontologies. In [8], Zheng et al. performed a thorough review of methods for auditing the Unified Medical Language System (UMLS), as well as ontology enrichment and alignment techniques. Two special issues [9, 10] have been organized and published to showcase the state of the art in auditing and quality assurance of biomedical terminologies and ontologies in 2009 and 2018, respectively.

In this special issue supplement, we aim to capture the most recent work on quality assurance and enrichment of biological and biomedical ontologies and terminologies. Articles were invited by sending out calls for papers to major listservers. A total of nine papers were accepted for publication in this special issue after going through a rigorous, single-blind review process. Each article was reviewed by 2–3 reviewers. The reviewers included both the authors who submitted their work as well as other prominent researchers from this field. All the reviewers

*Correspondence: licong.cui@uth.tmc.edu
[2] School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA
Full list of author information is available at the end of the article

had extensive experience working in the field of quality assurance and enrichment of medical terminologies. The entire process from sending out the first call for papers to the final publication of articles took approximately eleven months.

Summary of papers in this special issue

In the paper "Missing lateral relationships in top-level concepts of an ontology" [11], the authors leverage their previous work on two types of abstraction networks, "area taxonomy" and "subtaxonomy," to audit non-hierarchical or lateral relationships. In this study, the authors focus on grouping high-level concepts, called "top areas," in the NCIt's *Biological Process* hierarchy and SNOMED CT's *Eye/vision finding* sub-hierarchy. Manual reviews by domain experts revealed missing lateral relationships, which were further used to validate their hypotheses that top areas with a large number of concepts and concepts in the deeper hierarchical depth of top areas have a high likelihood of containing missing relationships. As the authors stated, their methodology can be seen as a useful addition to the quality assurance tools available to ontology maintenance personnel.

The paper "Extending import detection algorithms for concept import from two to three biomedical terminologies" [12] introduces a novel topological pattern, called "fire ladder," to structurally compare three terminologies (two source terminologies and one target terminology) and detect candidate concepts from the source terminologies that could potentially be imported into the target terminology. The authors explored the fire ladder patterns in ten selected terminologies in the UMLS (2018 AB release) and identified a total of 55 candidate instances for concept import, among which 39 were agreed by two domain experts and 48 by at least one expert. This is an important work on concept enrichment by leveraging external terminologies, and it may further help in enhancing semantic harmonization among different terminologies.

In "Web-based interactive mapping from data dictionaries to ontologies, with an application to cancer registry" [13], the authors present an interactive web-based tool to map data dictionary elements to ontology concepts. This tool has a recommendation engine at its core that provides a list of recommended concepts from the target ontology for an unmapped data element from the source data dictionary. This recommendation is based on a fuzzy matching algorithm. A pilot-test of the mapping between North American Association of Central Cancer Registries (NAACCR) elements extracted from Kentucky Cancer Registry (KCR) and NCIt concepts showed 47 of the 301 data elements were mapped to NCIt concepts. Of these 47, 25 were found to be correct when manually verified by domain experts. The study is important as such mapping techniques can provide semantic enrichment and interoperability between data dictionaries and ontologies.

The paper "Detecting missing IS-A relations in the NCI Thesaurus using an enhanced hybrid approach" [14] presents a structural-lexical-based methodology to identify potentially missing IS-A relationships in NCIt by using lexical features and role definitions of biomedical concept names. Missing IS-A relationships can result in erroneous output by applications that rely on NCIt as their underlying vocabulary. The authors explain this with an example where someone is searching for patients with "Cystic Neoplasm" using an NCIt powered search engine. However, "Dermoid Cyst" is currently not listed as one of the descendants (i.e., a missing IS-A relation) of "Cystic Neoplasm". As a result, patients with "Dermoid Cyst" will be missing from the search result. The proposed method involves computing non-lattice subgraphs and identifying candidate pairs of concepts that are currently not linked by IS-A relations, modeling concepts utilizing role definitions and lexical features to represent the meaning of concepts, and performing subsumption checking for candidate pairs of concepts.The authors applied their approach to the 19.08d version of NCIt and their algorithm found 55 potentially missing IS-A relationships. Domain experts from NCI Enterprise Vocabulary Services confirmed 29 of the 55 suggested anomalies as valid and were implemented in the newer versions of the thesaurus. NCIt is widely used as a reference terminology in cancer related research and in clinical care, and studies such as this can prove to be a useful tool to improve the quality of NCIt.

In "Friend of a Friend with Benefits Ontology (FOAF+): Extending a Social Network Ontology for Public Health" [15], the authors build a social network-related ontology for use in the field of public health to logically infer dyadic social networks between individuals. This ontology which the authors call Friend of a Friend with Benefits (FOAF+) ontology, is constructed to describe the domain of social and sexual behavior as it pertains to STI transmission between individuals. FOAF+ has 713 classes, 137 object properties, 130 data properties, and 312 instances. The authors compare FOAF+ with VIVO and FOAF using semiotic metrics produced by their automated tool OntoKeeper and their evaluation found the tool to be adequate as a prototype release. Social network ontologies such as FOAF+ can prove to be helpful in aiding machines to understand and interpret social network data, to identify missing links and to discover new relational links from network data.

In "Evaluation of lexical clarification by patients reading their clinical notes: A quasi experimental interview study" [16], the authors evaluate the functionality of a patient portal at the Dutch university medical centre (UMC Utrecht) that helps patients in understanding the clinical terms in free-text medical data. A Dutch medical terminology system is used to explain the terms to the patients using synonyms and definitions. A survey of 15 participants found the functionality easy to use as well as useful, albeit with low coverage of clarification of terms. The study is important as understanding of clinical notes helps patients remember their discussion with physicians and take better care of themselves as stated by the authors in the paper.

The paper "Analysis of readability and structural accuracy in SNOMED CT" [17] presents readability and structural accuracy metrics to provide a quantitative description of the structural aspects of an ontology and possible detection of missing semantic relations in the ontology. The underlying assumption of the authors is that an ontology should be friendly for both humans and machines and the correspondence between the contents for humans and machines should provide information regarding the quality of the ontology. The authors apply their metrics to different versions of SNOMED CT to provide useful insights about its modeling and evolution over time. The study is significant as the proposed metrics can be used to improve the effectiveness of the quality assurance process by identifying areas of an ontology with low readability and structural accuracy.

In the paper "Outlier concepts auditing methodology for a large family of biomedical ontologies" [18], the authors apply an abstraction network technique called "partial-area taxonomy" to SNOMED CT's *Specimen* hierarchy and NCIt's *Gene* hierarchy. They validate their hypothesis that concepts in small partial-areas of the partial-area taxonomy have statistically significantly more errors than concepts in large partial-areas, which is consistent with previous studies on four hierarchies from the same family of ontologies with outgoing lateral relationships. This further proves the scalability of the small partial-area technique to be potentially applied for auditing the larger family of biomedical ontologies in BioPortal.

In the paper "Towards semantic interoperability: finding and repairing hidden contradictions in biomedical ontologies" [19], the authors explore a technique to identify and repair unsatisfiable classes by combining ontologies from the Open Biomedical Ontologies (OBO) Foundry and the OBO ontologies. The study found 636 unsatisfiable classes in the nine OBO Foundry ontologies and over 300,000 unsatisfiable across 123 OBO ontologies. The authors also present a semi-automatic repair algorithm to identify axioms that result in these unsatisfiable classes which when removed, resolves the unsatisfiable classes. Applying this algorithm, the authors identified a small set of only 117 axioms that could be removed or modified to correct all the issues that were identified across all the ontologies. Consistency is an important key towards interoperability of ontologies and this study presents an effective approach to producing consistent and coherent ontologies.

Conclusions

With advances in health information technologies and their widespread adoption, ontologies and terminologies in biology and biomedicine have become ever more important to capture patient data in a consistent and standardized manner and for their effective transmission and communication. This has resulted in a call for more advances and research studies in the field of quality assurance and enrichment of these ontologies and terminologies. While research in this field has gained momentum over the past decade, the guest editors believe that advances will continue towards delivering more automated techniques for quality assurance and enrichment of ontologies and terminologies.

Acknowledgements
We would like to sincerely thank the authors for their scientific contribution and the reviewers for their valuable feedback.

About this supplement
This article has been published as part of BMC Medical Informatics and Decision Making, Volume 20 Supplement 10 2020: Quality Assurance and Enrichment of Biological and Biomedical Ontologies and Terminologies. The full contents of the supplement are available at https://bmcmedinformdecismak.biomedcentral.com/articles/supplements/volume-20-supplement-10.

Authors' contributions
Both AA and LC contributed to the writing of the manuscript. All the authors read and approved the final manuscript.

Funding
LC is supported in part by the University of Texas Health Science Center at Houston (UTHealth) startup. Publication costs are funded by LC's UTHealth startup.

Availability of data and materials
Not applicable.

Ethics approval and consent to participate
Not applicable.

Consent to publish
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Department of Computer Science, Manhattan College, New York, USA.
[2] School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA.

Published: 15 December 2020

References

1. Consortium GO. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res. 2019;47(D1):D330–8.
2. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. In: Bos L, Roa L, Yogesan K, O'connell B, Marsh A, Blobel B, editors. Studies in health technology and informatics, vol. 121. Amsterdam: IOS Press; 2006. p. 279.
3. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform. 2007;40(1):30–43.
4. Bodenreider O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. Yearb Med Inform. 2008;17(01):67–79.
5. Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: a functional perspective. Brief Bioinform. 2015;16(6):1069–80.
6. Zhu X, Fan JW, Baorto DM, Weng C, Cimino JJ. A review of auditing methods applied to the content of controlled biomedical terminologies. J Biomed Inform. 2009;42(3):413–25.
7. Amith M, He Z, Bian J, Lossio-Ventura JA, Tao C. Assessing the practice of biomedical ontology evaluation: gaps and opportunities. J Biomed Inform. 2018;80:1–13.
8. Zheng L, He Z, Wei D, Keloth V, Fan JW, Lindemann L, et al. A review of auditing techniques for the Unified Medical Language System. J Am Med Inform Assoc. 2020;27(10):1625–38.
9. Geller J, Perl Y, Halper M, Cornet R. Guest editorial: special issue on auditing of terminologies. J Biomed Inform. 2009;42(3):407–11.
10. Geller J, Perl Y, Cui L, Zhang GQ. Quality assurance of biomedical terminologies and ontologies. J Biomed Inform. 2018;86:106.
11. Zheng L, Chen Y, Min H, Hildebrand PL, Liu H, Halper M, et al. Missing lateral relationships in top-level concepts of an ontology. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01319-3.
12. Keloth VK, Geller J, Chen Y, Xu J. Extending import detection algorithms for concept import from two to three biomedical terminologies. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01290-z.
13. Tao S, Zeng N, Hands I, Mueller JH, Durbin EB, Cui L, Zhang GQ. Web-based interactive mapping from data dictionaries to ontologies, with an application to cancer registry. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01288-7.
14. Zheng F, Abeysinghe R, Sioutos N, Whiteman L, Remennik L, Cui L. Detecting missing IS-A relations in the NCI Thesaurus using an enhanced hybrid approach. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01289-6.
15. Amith M, Fujimoto K, Mauldin R, Tao C. Friend of a friend with benefits ontology (FOAF+): extending a social network ontology for public health. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01287-8.
16. van Mens H, van Eysden M, Nienhuis R, van Delden J, de Keizer N, Cornet R. Evaluation of lexical clarification by patients reading their clinical notes: a quasi-experimental interview study. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01286-9.
17. Abad-Navarro F, Quesada-Martınez M, Duque-Ramos A, Tomas F-B. Analysis of readability and structural accuracy in SNOMED CT. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01291-y.
18. Zheng L, Min H, Chen Y, Keloth VK, Geller J, Perl Y, et al. Outlier concepts auditing methodology for a large family of biomedical ontologies. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01311-x.
19. Slater L, Gkoutos G, Hoehndorf R. Towards semantic interoperability: finding and repairing hidden contradictions in biomedical ontologies. BMC Med Inform Decis Mak. 2020. https://doi.org/10.1186/s12911-020-01336-2.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.