MEDINFO 2019: Health and Wellbeing e-Networks for All
L. Ohno-Machado and B. Séroussi (Eds.)

doi:10.3233/SHTI190519

Characterizing the Scope of Exposome Research Through Topic Modeling and Ontology Analysis

Guillermo Lopez-Campos[a,b], Philip Kiossoglou[b], Ann Borda[b], Christopher Hawthorne[a], Kathleen Gray[b,c], Karin Verspoor[b,c]

[a] *Wellcome-Wolfson Institute for Experimental Medicine, Queen's University of Belfast, Belfast, Northern Ireland, United Kingdom,*
[b] *Health and Biomedical Informatics Centre, The University of Melbourne, Parkville, Victoria, Australia,*
[c] *School of Computing and Information Systems, The University of Melbourne, Parkville, Victoria, Australia*

Abstract

Exposomics is a field of research which is receiving growing attention. In this work, we characterize the exposome research landscape and update our previous study of formal knowledge representation approaches to this field. We applied a deductive analysis using the National Center for Biomedical Ontology Recommender for comparability of the results generated from a literature dataset and newly available ontologies with our previously published work. We highlight the changes in ontology recommendations.

Keywords:

Biomedical Ontologies, Data Mining, Environmental Exposure, Publications

Introduction

Environmental factors and exposures remain the major risk factors for most diseases and certainly for those that are a major burden for the health system. The concept of the *exposome* was introduced in 2005 when C.P. Wild coined the term, defining it as an exposure-oriented analogue of the genome [1]. The concept has since gained significant attention, particularly with the recent establishment of several large research collaborations and because of the demonstrated relevance to the new paradigm of "precision medicine" [2].

The increasing corpus of literature focused on the exposome provides opportunities to characterize the breadth of the field. As an interdisciplinary and evolving field, exposome research covers a broad range of topics and methodologies that could be related to different areas such as public health and biomarker detection.

Formal knowledge representation is an important tool for organization of information in many different areas and disciplines. The exposome poses complex challenges for biomedical informatics and requires the adaptation and development of suitable methods and tools [3]. Knowledge representation methods, such as ontologies, have also significantly supported "omics" research; examples of well-adopted informatics tools include the Gene Ontology [4] and the Human Phenotype Ontology, both of which provide formal vocabularies and represent relationships among concepts within these vocabularies.

In an associated work on which we base this study [5], we proposed a methodology to characterize this evolving exposome landscape.

The deductive analysis approach of the prior work used the National Center for Biomedical Ontology (NCBO) Recommender [6], which allowed for contextualization of current formal knowledge representation of the exposome. However, only a small portion of the overall recommended ontologies could be deemed as broadly useful in the field of exposome research at that point. Importantly, the results showed that none of the frequently recommended ontologies had major associations with the science of environment or exposures, suggesting that there was a need to bridge interdisciplinary research gaps to build knowledge in this field.

This study seeks to re-apply our previously developed deductive analysis to analyze the evolution of knowledge representation tools covering exposome research, as well as to ensure comparability with new publication results, indicated by the growth of the corpus of exposome literature, which has more than doubled in the last two years (2017-2018).

Methods

We analyzed the recommended ontologies using an expanded dataset of papers up to October 2018. The new dataset used a similar search strategy to our previous work [5]. Briefly, we considered papers published between January 2005 and October 2018 with the term "exposom*" in the title, abstract or keyword fields.

The abstracts from the retrieved documents were subsequently submitted for ontology recommendation using the NCBO Ontology Recommender tool API. Default parameters were used to ensure the comparability of the results generated using the literature dataset up to 2018 and the new ontologies, with those previously published.

For topic modeling, we used the R package "text2vec" (v0.5.1) on the subset of documents that contain the word exposome in their abstract.

Results

The results from the ontology analyses using NCBO's Ontology Recommender for the search 2005-2018 resulted in the identification of 171 ontologies for 2018 and 165 for 2016. Of those just 20 ontologies were present in more than 50% of the documents (50 ORF – ontology recommended fraction). The comparison of these ontologies with the results previously published shows that there is a significant overlap among the ontologies recommended for the three datasets in Table 1.

To identify the number of topics we used the R library "LDATuning" determining the presence of 25 topics, based on a balance among different criteria.

Table 1 – List of Ontologies Recommended with ORF >50 in 2016 and in 2018.

Shared 2016 & 2018		Found in 2018	Found in 2016
NCIT	EFO	IOBC	SNMI
SNOMEDCT	CRISP	CHEAR	RH-MESH
MESH	HL7	GO-EXT	
LOINC	NDFRT	MEDDRA	
HUPSON	EDAM	NBO	
ONTOAD	GO		
NIFSTD	RCD		
MEDLINEPLUS			

Discussion

Deductive analyses using the knowledge representation tools led to the identification of 171 recommended ontologies for the 2018 dataset of documents, in comparison with 165 ontologies identified in 2016. The differences between the 2016 and 2018 recommendations are slightly reduced after applying the ORF >50 threshold.

However, there are five differences in the ontologies that fulfill the requirement of being recommended for at least 50% of the documents that appear as significant in the 2018 analyses. Among these five newly identified ontologies it is important to note that three of them, NBO (Neurobehaviour Ontology), MEDDRA (Medical Dictionary for Regulatory Activities Terminology) and GO-EXT (Gene Ontology Extension), were present in the 2016 results but did not reach the significance threshold of ORF >50 at that time. In the most recent analysis, these three have jumped to be recommended for around 60% of the documents. This change was more relevant for NBO which experienced a nearly 50% increase in its recommendations (from 41% to 62% of the documents). In terms of the recommendation ranking, MEDDRA did not change its average recommendation position (a corrected average of 20.5) whereas in the case of GO-EXT it improved its ranking in three positions, from an average corrected rank of 20.7 to 18.7. Similarly NBO's rank improved from 22.3 to 19.5.

These changes in the recommended ontologies can be explained by the update of existing technologies and by an expansion of the exposome research in areas already covered by the ontologies that led them to be deemed as significant.

The other two newly recommended ontologies CHEAR (Children's Health Exposure Analysis Resource) [7] and IOBC (Interlinking Ontology for Biological Concepts) are recently developed and cover important aspects related to the exposome. CHEAR has been recommended for 64% of the documents. This resource was developed for the analysis of environmental exposures in an "exposome context" and therefore it could be considered as an "authentic" exposome ontology. IOBC is a recently developed ontology containing both biological concepts and earth and environmental sciences concepts.

Conclusions

Exposomics research and literature have doubled over the last two years. This increment reflects the increasing relevance of the field. Analysis of the recommended ontologies that could potentially be used to annotate or model exposome research still shows that broad and general tools such as National Cancer Institute Thesaurus (NCIT), Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT), Medical Subject Headings (MESH), and Logical Observation Identifiers Names and Codes (LOINC) would cover aspects of all the documents. However, in the last couple of years, CHEAR and IOBC have emerged as two new and promising tools to describe the knowledge around the exposome more formally.

Acknowledgements

C. Hawthorne is funded by a Northern Ireland Department for the Economy (DfE) postgraduate studentship Award.

References

[1] C.P. Wild, Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology, *Cancer Epidemiol Biomarkers Prev* **14** (2005), 1847-1850.

[2] National Research Council (U.S.). Committee on A Framework for Developing a New Taxonomy of Disease., Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease, *National Academies Press*, Washington, D.C., 2011.

[3] F. Martin Sanchez, K. Gray, R. Bellazzi, and G. Lopez-Campos, Exposome informatics: considerations for the design of future biomedical research information systems, *J Am Med Inform Assoc* **21** (2014), 386-390.

[4] S. Carbon, H. Dietze, S.E. Lewis, C. J. Mungall, M.C. Munoz-Torres, and S. Basu, The Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium, *Nucleic Acids Res* **45** (2017), D331-D338.

[5] P. Kiossoglou, A. Borda, K. Gray, F. Martin-Sanchez, K. Verspoor, and G. Lopez-Campos, Characterising the scope of exposome research: A generalisable approach, MEDINFO 2017: Precision Healthcare through Informatics*, Stud Health Technol Inform* **245** (2017), 457-461.

[6] C. Jonquet, M.A. Musen, and N.H. Shah, Building a biomedical ontology recommender web service, J Biomed Semantics 1 Suppl 1 (2010), S1.

[7] D.M. Balshaw, G.W. Collman, K.A. Gray, and C.L. Thompson, The Children's Health Exposure Analysis Resource: enabling research into the environmental influences on children's health outcomes, Curr Opin Pediatr 29 (2017), 385-389.

Address for correspondence

Dr. Guillermo Lopez Campos.
Email: G.LopezCampos@qub.ac.uk