MEDINFO 2017: Precision Healthcare through Informatics
A.V. Gundlapalli et al. (Eds.)

doi:10.3233/978-1-61499-830-3-457

Characterising the Scope of Exposome Research: A Generalisable Approach

Philip Kiossoglou[a], Ann Borda[a], Kathleen Gray[a,c], Fernando Martin-Sanchez[a,b], Karin Verspoor[a,c], Guillermo Lopez-Campos[a,d]

[a] Health and Biomedical Informatics Centre, The University of Melbourne, Parkville, Victoria, Australia,
[b] Environmental and Participatory Health Informatics (ENaPHI) Research Group, Division of Health Informatics, Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY, USA,
[c] School of Computing and Information Systems, The University of Melbourne, Parkville, Victoria, Australia,
[d] Centre for Experimental Medicine, Queen's University of Belfast, Belfast, Northern Ireland, United Kingdom

Abstract

Scientific advancement and the development of new research fields bring uncertainties about what the current topics of research emphasis are and thus, what new knowledge might need to be represented. The exposome is an example of one such new field for which these uncertainties exist. The exposome is the analogue to the genome, from an environmental exposure perspective; research on the exposome has gained momentum only since 2011. In this work, we propose a generally applicable methodology that aims to characterise the landscape of a new research area based on linguistic analysis of its associated publications. Using abstracts of 261 exposome research articles, we illustrate a methodology that combines (1) inductive analysis based on word frequency counts, and term analysis to identify the topics, methods and applications of the new field and (2) deductive analysis using the NCBO Ontology Recommender to identify to what extent this new area is covered by current knowledge representation tools. Applying this method to the exposome literature, we uncover both the current focus of exposome research and the ontologies that are most relevant to the domain.

Keywords:

Environmental Exposure; Medical Informatics, Text Mining, Biomedical Ontologies

Introduction

The role of environmental factors in health and disease along with the controversial "nature versus nurture" debate have been ongoing themes in science and research for over a century [1], [2], [3]. With the development of precision medicine, the relevance of both themes and the need to consider both in research have been explicitly acknowledged [4], [5]. This represents an attempt to satisfy the equation "Phenotype = Genome x Environment". The "nature" component of this equation, the genome, has experienced a formidable boost in the last two decades. It has also fostered the development of what are known as "omics" approaches (proteomics, metabolomics, etc.) to characterize molecular phenotypes. A commonality among these approaches has been their reliance on recent technological advances to generate very large data sets. On the other hand, the "environmental" aspect of the equation was typically studied with population or environmental health approaches, until 2005 when C.P. Wild coined the term "Exposome" and further defined it as an exposure-oriented analogue of the genome [6]. The exposome represents the sum total of exposures an individual receives over time, from both internal and external sources. Similarly to what has happened with the development of other "omics" approaches, study of the exposome is benefitting from recent advances in technology, especially with the development and reduction in cost of sensors. These new technologies have opened the door to the development of new strategies that shift research focus from population-based exposure assessment to more individualised approaches [7].

The concept of the exposome was initially established over a decade ago, and has gained significant attention only recently with the establishment of several large research projects [8-11], due in part to its demonstrated relevance to the new paradigm of "precision medicine" [12]. As a consequence of this increased focus, the body of knowledge and the volume of literature associated with the exposome have grown continuously over the last six years (Figure 1). Although still young, this is a rapidly evolving research field.

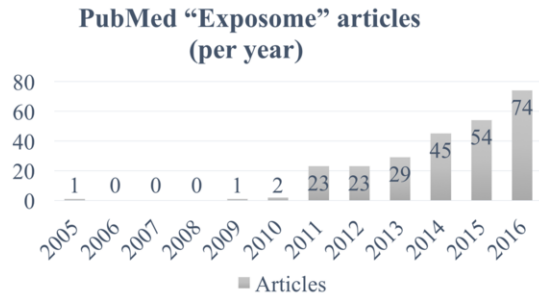

Figure 1 – Number of Articles Indexed in PubMed Containing the Term "exposome" since it was Coined in 2005.

The exposome is posing new challenges for biomedical informatics and requires the adaptation and development of methods and tools [13]. Knowledge representation methods, such as ontologies, have greatly supported "omics" research; examples of successful informatics tools include the Gene Ontology [14] and the Human Phenotype Ontology [15]. These ontologies provide formal vocabularies to describe knowledge and represent the relationships among entities within those vocabularies.

Exposome research is a relatively new field with an increasing corpus of literature but it is not yet well characterised in terms of formal knowledge representation, hampering our ability to develop suitable informatics tools. For this reason, in this work we propose a methodology to characterise this new research landscape, by identifying the terms and concepts that better

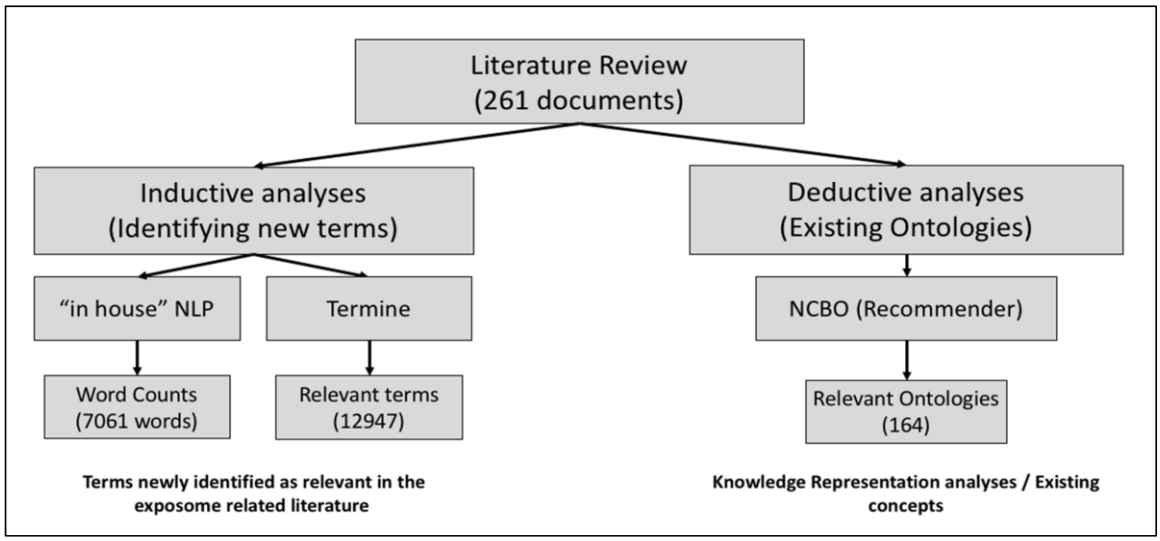

Figure 2 – Methodology for Characterising Exposome Research through Inductive and Deductive Analysis of the Literature.

describe exposome related knowledge. We believe that this methodology is applicable not only to the exposome but also is transferable to any new or developing research field.

Methods

Our characterization of the exposome research landscape is based on a literature review to discover the extent of relevant articles, followed by two different and complementary approaches to content analysis of their abstracts (Figure 2):

- An inductive approach based on text mining to identify the most relevant words and terms;
- A deductive approach based on identifying the most relevant ontologies, i.e. those that offer the best coverage of terms.

The literature selection was performed by searching for papers published between January 2005 and April 2016 with the term "exposom*" in the title, abstract or keyword fields. Searches were undertaken in the CABI, CINAHL, EMBASE, PubMed and Web of Science databases. Articles were excluded if they were written in a language other than English, if they were an e-book, or if they were related to research in non-human subjects. 22 items were excluded under these criteria, resulting in a total of 261 documents that were finally used in these analyses. Abstracts of the selected documents were then retrieved and stored in a database to form a corpus for analysis.

The inductive stage of the analysis was performed applying two different strategies. In the first strategy, "technical terms" were identified in the corpus using Termine [16], a tool that approaches novel multi-word term identification using statistical methods. The second strategy consisted of calculating word frequency counts in the corpus applying a natural language processing (NLP) protocol developed "in house" using Python's Natural Language Tool Kit (NLTK) [17].

Using Termine, abstracts were concatenated into a single file in order to provide enough text for effective processing in one pass. Prior to this concatenation, a preprocessing step on the individual abstracts was carried out to improve the final outcomes and the submission process, as follows: Copyright notices were removed from the text; they were converted to lowercase and they were lemmatized to reduce the number of duplicated terms (for example "exposure" and "exposures" were grouped together); and finally sentences were tokenized (using the NLTK method) and replaced with new lines.

Word frequency analysis was performed in the following fashion. Abstracts were concatenated first, then converted to lowercase, and punctuation characters were removed. A frequency distribution was then produced using the nltk.probability.FreqDist procedure. This result was stored and then manually curated to limit the terms considered potentially relevant (for example, discarding closed-class function words such as 'the', 'of' and 'by').

For the deductive analysis, we used the repository of biomedical ontologies found at the National Center for Biomedical Ontology (NCBO). This resource currently catalogues more than 500 ontologies. The NCBO Ontology Recommender tool API [18] was applied in order to identify what ontologies covered this new field. An automatic process was used to pass each abstract separately to the API and record the ontologies recommended in each response.

Results

Inductive Analyses

Word count analysis identified 7061 unique words with a range of counts between 1 and 2844. The count distribution shows a Zipfian-like distribution, where only 12.7% (897) of the words have a word count greater or equal than 10. We further filtered these results applying a minimal word-count threshold of 10 followed by manual filtering of the remaining words. This filtering step, designed to remove frequent general English words and other words irrelevant for this field based on our own linguistic knowledge, resulted in a final list of 427 relevant words with word frequencies between 10 and 614. Of these 427 words, the top 25 showed a word frequency >100 (Table 1).

Table 1 – Top 25 Words in Exposome Abstracts (word counts >100) identified using word frequency analysis.

Word	Count	Word	Count
Exposure	614	Biomarker	151
Environmental	405	Analysis	136
Study	386	Effect	135
Disease	339	Cancer	133
Exposome	289	Method	128
Health	252	Genetic	118
Human	239	Molecular	115
Risk	215	Biological	114
Data	200	Individual	112
Factor	190	Development	104
Research	172	Assessment	101
Approach	166	Interaction	101
Chemical	165		

Termine identifies complex and technical terms comprising two or more words. These results are then ranked according to their frequency and an internally calculated score (c-value) that is derived from statistical and linguistic information [16]. This analysis resulted in the identification of 12,947 terms with c-values ranging between -1 and 105.8 and frequencies between 1 and 107. Of those terms identified, 95% occurred at a very low frequency (<= 2), and only 1% had a frequency of ten or greater. Table 2 shows the top 25 terms identified.

Table 2 – Top 25 Multi-Word Terms in Exposome Abstracts, as Identified by Termine.

Term	c-value	Frequency
environmental exposure	105.8	107
environmental factor	50.9	52
risk factor	43.4	45
risk assessment	37.8	39
chronic disease	33.9	35
public health	33.9	35
association study	30.3	32
exposure science	28	29
gene expression	26.9	28
exposure assessment	23.9	25
environmental health	22.7	24
gene environment interaction	20.7	22
mass spectrometry	19.9	21
epidemiological study	19.8	21
disease risk	19.8	21
genome wide association study	19.0	13
human health	18. 9	20
metabolic profiling	17.8	19
human exposome	17	18
air pollution	15.9	17
complex disease	15	16
human genome	14.8	16
human disease	14.6	16
omic technology	14	15
cohort study	13.7	15

Deductive Analysis

The NCBO Ontology Recommender tool uses four separate scores for each ontology based on internal calculations measuring the "coverage", "detail of knowledge", "specialization" and "acceptance" of the proposed ontology. These four categories are weighted to produce a default ranking. Using the NCBO default settings, the weights employed are Coverage: 0.55, Detail of Knowledge: 0.15, Specialization: 0.15 and Acceptance: 0.15. These analyses returned a list limited to the twenty five (25) most highly ranked ontologies for a supplied abstract.

A total of 164 different ontologies were recommended, across the 261 abstracts. To better interpret these results we calculated a measure, the ontology recommendation frequency percentage (ORFP), to represent how often an ontology was recommended for this dataset. Applying an arbitrary threshold of ORFP > 50, only 17 (~10%) of the 164 suggested ontologies were recommended for more than 50% of abstracts (Figure 3 and Table 3).

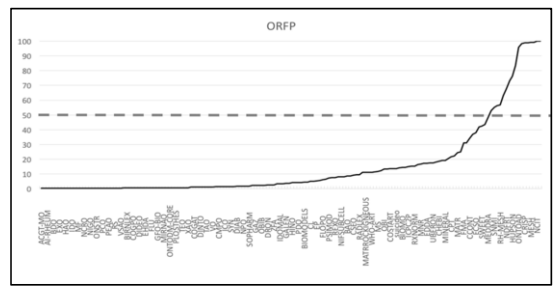

Figure 3 – Recommendation Frequency for all 164 Ontologies Recommended by the NCBO Ontology Recommender. The Dotted Line Shows the 50% ontology Recommendation Frequency Percentage (ORFP) Threshold. Not all Ontologies are Named on the X-Axis.

To further refine these results, we aggregated all 164 recommended ontologies and calculated a new combined ranking for the 17 ontologies above the threshold. This was performed by means of computing a "corrected rank" for each ontology, taken as the average rank it received for each individual abstract. If an ontology was not recommended for a given abstract, it was given an arbitrary rank of 26.

Table 3 – 17 ontologies recommended by the NCBO Ontology Recommender with ORFP > 50, showing their average and corrected ranks.

Ontology	ORFP	Average rank	Corrected rank
NCIT	100.00	1.16	1.16
SNOMEDCT	100.00	3.62	3.62
MESH	99.23	3.20	3.38
NIFSTD	99.23	7.37	7.51
CRISP	98.85	5.89	6.12
RCD	98.85	6.25	6.48
LOINC	98.47	5.63	5.94
ONTOAD	95.79	9.32	10.03
EFO	83.52	12.73	14.92
HUPSON	76.25	14.53	17.25
HL7	73.18	12.64	16.22
NDFRT	67.82	14.67	18.31
EDAM	63.22	14.85	18.95
RH-MESH	56.70	12.46	18.32
GO	56.32	14.37	19.45
SNMI	55.17	15.60	20.26
MEDLINEPLUS	52.87	13.51	19.40

To check for consistency between our inductive and deductive results, we added a deductive step where we used the list of multi-word terms filtered from Termine as input to the NCBO Ontology Recommender tool. This resulted in 25 recommended ontologies, all of which also appeared in the list of 164

ontologies derived from the deductive analysis of complete abstracts. These 25 ontologies matched 15 out of the 17 ontologies with an ORFP >50, demonstrating the overall representativeness of the inductive terms as compared to the broader literature. The two missing ontologies were EDAM and RH-MESH (Figure 4).

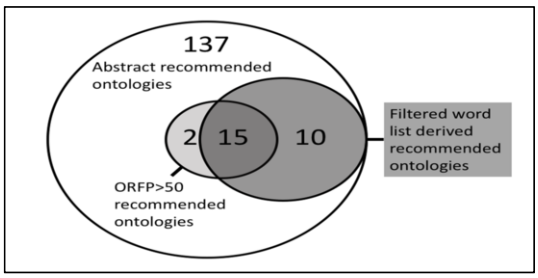

Figure 4 – Consistency Between Two Sets of Ontologies Recommended for Exposome Abstracts: 25 Recommended for Filtered Multi-Word List from Inductive Analysis, and 17 Recommended from Deductive Analysis of Whole Abstracts.

Discussion

The proposed methodology is based on two different and complementary approaches, an inductive strategy aiming to unveil unknown aspects associated with the terms and words (eventually concepts) that might characterise exposome research, and a deductive method that attempts to identify what existing elements in the biomedical knowledge representation space can be used to annotate exposome related published research.

The two inductive analyses were complementary to each other. The word frequency method resulted in the identification of individual words such as "environmental", "exposure" or "biomarker" whereas the use of Termine focused on the identification of more complex terms and concepts such as "environmental health", "environmental exposure" or "metabolic profiling".

The word frequency approach was applied to identify the most frequent words in the literature and gave insights into individual elements that provide an overall perspective on the field. Although Table 1 shows only the top 25 results, this analysis enabled the identification of other relevant words in this context that show different areas where research has focused such as "epigenetic", "metabolomic", "maternal" or "lead" (a metal) and even more specialised words such as "adductomics".

The second inductive approach, using Termine, allowed us to identify more complex terms and concepts that elude the simple word count approach. The reason why Termine analysis identified a large number of terms, greater than the individually identified words, is because the results include terms of different length (word *n-grams*, or multi-word terms) that are sequences of individual words (i.e. "environmental factor", "environmental factor environome" and "environmental factor such investigation"). The results in Table 2 clearly show the complementarity to those identified by word count. The analysis of terms allowed us to identify methodologies (such as "cohort studies" or "genome wide association studies") and techniques ("mass spectrometry" or "metabolic profiling") that are relevant in exposome research. They thus identified terms which correspond to important concepts in this emerging field.

The approach based on deductive analysis using the ontology recommender allowed us to contextualise current formal knowledge representation of the exposome. Based on this analysis, only a small portion of the overall recommended ontologies are broadly useful in this field of research at this point. More importantly, none of the ontologies that were frequently recommended had major associations with the science of environment or exposures, suggesting that there is a need to bridge interdisciplinary research gaps to build knowledge in this field.

Limitations

In spite of having successfully applied this methodology to the characterisation of exposome research with complementary and consistent results, this study has several limitations in its current form. One of these is connected with the use of abstracts rather than full text content in the analysis, since abstracts highlight the key aspects of the full-text content of the document but are not intended to be exhaustive summaries. However, since many full text publications are not available for analysis, this limitation is balanced with more exhaustive coverage of relevant literature using abstracts.

Another limitation lies in the use and interpretation of the results from the proposed automatic analyses with regard to the ambiguity of terms; the tools do not have the capacity to disambiguate different meanings. For example, in the inductive analysis the word "lead" would be equally counted regardless of whether it is acting as a noun (a metal) or as a verb (to precede). In the deductive analysis there is a similar problem with different ontologies recognising the same term but without contextualising it. In the aforementioned example "lead" is defined by NCIT as "Be in charge of; a position of leadership" whereas in CRISP it is defined as a "Soft grayish blue metal with poisonous salts; symbol, Pb, atomic number 82".

General Application of the Approach

Scientific advancement and the development of new research areas bring uncertainties about how new knowledge can be or should be represented. The methodology proposed here enables objective assessment of such questions and thus facilitates analysis of the evolution of new fields of biomedical science, including those driven by new technologies.

For example in this study of exposome research, the deductive analysis showed the relevance of toxicogenomics in exposome research in the form of a number of genomic and bioinformatics related ontologies (GO, EDAM, EFO and HUPSON).

Our suggested method combining induction and deduction is generalisable and suitable to be applied in corpora of literature derived from other biomedical fields. The inductive stage is "blind" to the origin of the corpus used, and independent of any existing knowledge resources. Therefore it can equally applied in any area with an identifiable body of literature, allowing the most relevant words and terms that charaterise the new field to surface and thus providing insight into the current methods, applications and general interests in the field – as well as the gaps. On the other hand, the deductive stage relates the corpus to existing knowledge encoded in the form of ontologies, and maps the places where aspects of current knowledge overlap with with research in the new field of interest, or where there is no intersection.

Future Work

In future work, we plan to analyse full texts of articles from the literature instead of just abstracts. In addition, more complex natural language processing strategies will be applied in order to reduce the disambiguation problem. These results will eventually be combined with the results generated from the deductive approach to improve the interpretation of the latter.

This will support eventual comparison between the results from the analysis of exposome research literature and the results of

analysis of representative bodies of literature from related fields (for instance, other omics research). This may determine which features of exposome research are particularly novel and thus may help to direct further research.

Conclusion

We have developed a methodology that combines inductive and deductive approaches for the characterisation of the landscape of terms or concepts and availability of formal knowledge representation tools in a relatively new area of biomedical research. When applying this methodology to the study of exposome research, results from the inductive and deductive approaches are consistent and complementary, allowing us to identify different sets of terms, concepts and ontologies that describe the current status of human exposome-related knowledge. This is a major first step towards enabling development of informatics tools to support systematic comprehensive integration of exposome data into precision medicine.

The use of automated tools and methods like the ones described in this work adds to the growing body of work to address the limitations of the current processes in systematic reviews of this kind [19; 20]. Our methods could enable the generation of an overall perspective of research scope, and characterisation of the most relevant topics under investigation, in new disciplines. The approach further facilitates the assessment of whether new formal knowledge representation tools are required for the new discipline.

References

[1] F. Galton, *Hereditary genius : an inquiry into its laws and consequences*, Macmillan and co., London, 1869.

[2] B.S. Burks, Foster Parent-Foster Child Comparisons as Evidence upon the Nature-Nurture Problem, *Proc Natl Acad Sci U S A* **13** (1927), 846-848.

[3] B.T. Eiduson, S. Eiduson, and E. Geller, Biochemistry, genetics, and the nature-nurture problem, *Am J Psychiatry* **119** (1962), 342-350.

[4] F.J. Martin-Sanchez and G.H. Lopez-Campos, The New Role of Biomedical Informatics in the Age of Digital Medicine, *Methods Inf Med* **55** (2016), 392-402.

[5] D.P. Jones, Y. Park, and T.R. Ziegler, Nutritional metabolomics: progress in addressing complexity in diet and health, *Annu Rev Nutr* **32** (2012), 183-202.

[6] 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.

[7] M.J. Nieuwenhuijsen, D. Donaire-Gonzalez, M. Foraster, D. Martinez, and A. Cisneros, Using personal sensors to assess the exposome and acute health effects, *Int J Environ Res Public Health* **11** (2014), 7805-7819.

[8] C. Potera, The HELIX Project: tracking the exposome in real time, *Environ Health Perspect* **122** (2014), A169.

[9] M. Vrijheid, R. Slama, O. Robinson, L. Chatzi, M. Coen, P. van den Hazel, C. Thomsen, J. Wright, T.J. Athersuch, N. Avellana, X. Basagana, C. Brochot, L. Bucchini, M. Bustamante, A. Carracedo, M. Casas, X. Estivill, L. Fairley, D. van Gent, J.R. Gonzalez, B. Granum, R. Grazuleviciene, K.B. Gutzkow, J. Julvez, H.C. Keun, M. Kogevinas, R.R. McEachan, H.M. Meltzer, E. Sabido, P.E. Schwarze, V. Siroux, J. Sunyer, E.J. Want, F. Zeman, and M.J. Nieuwenhuijsen, The human early-life exposome (HELIX): project rationale and design, *Environ Health Perspect* **122** (2014), 535-544.

[10] HERCULES Exposome Research Center, http://emoryhercules.com/

[11] P. Vineis, M. Chadeau-Hyam, H. Gmuender, J. Gulliver, Z. Herceg, J. Kleinjans, M. Kogevinas, S. Kyrtopoulos, M. Nieuwenhuijsen, D.H. Phillips, N. Probst-Hensch, A. Scalbert, R. Vermeulen, C.P. Wild; EXPOsOMICS Consortium, The exposome in practice: Design of the EXPOsOMICS project, *Int J Hyg Environ Health* **16** (2016), 30130-4.

[12] 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.

[13] 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.

[14] C. The Gene Ontology, Expansion of the Gene Ontology knowledgebase and resources, *Nucleic Acids Res* (2016).

[15] S. Kohler, N.A. Vasilevsky, M. Engelstad, E. Foster, J. McMurry, S. Ayme, G. Baynam, S.M. Bello, C.F. Boerkoel, K.M. Boycott, M. Brudno, O.J. Buske, P.F. Chinnery, V. Cipriani, L.E. Connell, H.J. Dawkins, L.E. DeMare, A.D. Devereau, B.B. de Vries, H.V. Firth, K. Freson, D. Greene, A. Hamosh, I. Helbig, C. Hum, J.A. Jahn, R. James, R. Krause, S.J. Laulederkind, H. Lochmuller, G.J. Lyon, S. Ogishima, A. Olry, W.H. Ouwehand, N. Pontikos, A. Rath, F. Schaefer, R.H. Scott, M. Segal, P.I. Sergouniotis, R. Sever, C.L. Smith, V. Straub, R. Thompson, C. Turner, E. Turro, M.W. Veltman, T. Vulliamy, J. Yu, J. von Ziegenweidt, A. Zankl, S. Zuchner, T. Zemojtel, J.O. Jacobsen, T. Groza, D. Smedley, C.J. Mungall, M. Haendel, and P.N. Robinson, The Human Phenotype Ontology in 2017, *Nucleic Acids Res* (2016).

[16] K. Frantzi, Ananiadou, S. and Mima, H., Automatic recognition of multi-word terms: the C-value/NC-value method, *International Journal on Digital Libraries* **3** (2000), 15.

[17] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*, O'Reilly, Beijing ; Cambridge Mass., 2009.

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

[19] A. O'Mara-Eves, J. Thomas, J. McNaught, M. Miwa, and S. Ananiadou, Using text mining for study identification in systematic reviews: a systematic review of current approaches, *Syst Rev* **4** (2015), 5.

[20] M. Verdonck, F. Gailly, S. de Cesare, and G. Poels, Ontology-driven conceptual modeling: A systematic literature mapping and review, *Applied Ontology* **10** (2015), 197-227.

Address for correspondence

Guillermo Lopez-Campos

Email : guillermo.lopez@unimelb.edu.au