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

doi:10.3233/SHTI190468

Linked Open Data in the Biomedical Information Area: A Keywords Analysis

Stefano Bonacina

Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Stockholm, Sweden

Abstract

The objective of this paper was to determine the extent of the usage of "Linked Open Data" within biomedical literature. Applying PRISMA statement for literature reviews, forty-six papers were included in the analysis and keywords identified. Keywords have been classified according to MeSH categories, when possible. Twenty-three keywords had a frequency > one, 146 keywords had a frequency equal to one. Two MeSH categories were recurring. Future work includes applying association rules learning to keywords.

Keywords:

Semantic Web, Review Literature as Topic.

Introduction

Based on protocols defined by the World Wide Web Consortium (W3C), Semantic Web allows data sharing among applications and application-application cooperation, even without human intervention [1].

In the context of Semantic Web, the term "Linked data" is about linking and sharing resources of data (e.g., datasets, knowledge bases) generated and exposed by heterogeneous authors (e.g., governmental departments, research institutions, industry) [1]. When linked resources of data, e.g., heterogeneous datasets connected each other, are exposed through open access, the "Linked Open Data" paradigm is fulfilled, according to the "Linking Open Data" community project [2]. Operatively, "Linked Data" resources are represented with the Resource Description Framework (RDF) [3], a family of standardised data formats (RDF is explained briefly in the next section).

In 2016, we published a review paper that gave a view of research relating to "Linked Open Data" (LOD) in the area of health- and clinical-care [4]. Widening the scope, and including publications from the last two years, the objective of this paper is an attempt to determine the extent of the usage of linked open data within biomedical literature. The main questions guiding this review were as follows,

1. To what extent "Linked Open Data" is used within Biomedical Information research?
2. How could it be represented in a synthesised way?

The rest of the paper is organized as follows. The next section is a short background on Semantic Web technology, and RDF data format. Then, the methods used in the research and the obtained results are presented. Finally, some concluding remarks end the paper.

Semantic Web

Semantic Web stack is a set of protocols and computer languages, organised according to a layered structure, where communication is allowed between adjacent layers [5]. Semantic Web stack has been implemented to achieve data sharing among applications. In the stack, three main computer languages are: the Web Ontology Language (OWL) [6], the Resource Description Framework (RDF) Language [3], and the SPARQL Protocol And RDF Query Language - SPARQL language [7], which has a recursive acronym. OWL and RDF are for specifying ontologies, which are collections of data and meta-data exposed to the web and shared among computer applications [1,6]. Similarly to a simple sentence in English (which is composed by a Subject a Predicate, and an Object), a RDF triple - the basic element of this language - is composed by a Subject, a Predicate, and an Object. An example is, "The patient is suffering from diabetes", where "The patient" is the Subject, "is suffering from" is the Predicate, and "diabetes" is the Object. Then, , a set of triples is represented by a graph, where the nodes of the graph are the Subjects and the Objects, and the edges are the Predicates. Triples are stored according to the Extensible Mark-up Language (XML) [3].

Methods

To answer the first research question, a review of the literature has been performed, following the different phases of a systematic review (PRISMA Statement), described in [8]. Then, to answer the second research question, MeSH hierarchy has been used to aggregate concepts represented by the keywords associated to each paper. The following sub-sections describe the applied methods in detail.

Search strategy

The author searched online bibliographic databases for relevant papers. Specifically, PubMed/Medline, Scopus, Web of Science (MEDLINE), Web of Science (All the other Indexes), and IEEE Xplore - Digital Library were enquired, using "Linked Open Data" as main search term. Because Scopus, Web Of Science (All the other Indexes), and IEEE Xplore - Digital Library are not specific to biomedical literature, the searches on these databases were limited to the following subject areas: Biochemistry, Genetics and Molecular Biology, Medicine, Pharmacology, Toxicology and Pharmaceutics, Psychology, Immunology and Microbiology, Neuroscience, and Nursing. In this way, the search was limited to papers in the Biomedical Information area.

Inclusion and exclusion criteria

Inclusion criteria were as follows: 1 - Journal articles in English published from 1st January 2009 to 13th August 2018 (the date of searches); 2 - Papers published on journals related to Linked Open Data within the Biomedical Information Domain; 3 - Abstract and full text pdf are available; and 4 - Papers should describe Linked Open Data solutions within the Biomedical Information Domain.

The exclusion criteria were complementary to the inclusion criteria. In addition, duplicated items were considered and deduplicated. The final selection of papers was performed by screening title, abstract, and through reading the full text, when more details were needed.

Data extraction

To answer the second research question, keywords associated to each paper were extracted. Keywords are descriptors of the content of a publication, and in the present research, they have been preferred to the title, as the titles cannot reflect the contents of a publication [9]. Keywords are defined by the authors, even according to a specified vocabulary (e.g., Medical Subject Headings-MeSH). Publications indexed in Medline database are also described by MeSH terms. Publications in IEEE are indexed according to IEEE descriptors.

Keywords describing the papers selected from the previous phase of the research have been collected in a MS Excel spreadsheet and then processed according to the following steps. First, keywords on the full text pdf have been considered; then, if they were not included, the journal web page describing the paper has been considered, and the keywords available there collected. After that, if the keywords were not available on the web site, MeSH terms have been considered for the publications available on Medline. With this process all the papers had a set of words associated to them.

The phrase “Linked open data” and combinations of those words have been excluded, as the main query in the search phase was that one (i.e., to retrieve papers having LOD as main topic), duplications have been removed, and acronyms have been used to replace their explanations (e.g., RDF instead of “Resource Description Framework”). After that, the entire list has been alphabetized and the frequency of each keyword counted, and stored in the spreadsheet. Finally, a MeSH term and a MeSH category have been connected with each keyword, when possible.

Results

The search phase provided 181 potential articles, and following the four-phase flow diagram [8], 46 have been included in the analysis. Due to space limitation, the references of the 46 included papers are not included in the present manuscript; however, they are available upon request. The analysis of the 46 papers gave 265 keywords, including duplications. After processing the keywords according to the methods explained in the previous section, 225 keywords had been obtained, including duplicates. 169 distinct keywords were extracted; 23 had a frequency from 2 to 15 (maximum frequency), while 146 had a frequency equal to one. The most recurrent keyword was “RDF” (15 times), followed by “Semantic web” (9 times), “Electronic Health Records” appeared 4 times, and “Unified Medical Language System” – UMLS appeared 3 times. The main MeSH categories matched by the keywords were "Information Science Category" (45 keywords) and "Analytical, Diagnostic and Therapeutic Techniques and Equipment Category" (12). Unfortunately, it was not possible to associate 80 keywords to a MeSH term (and category).

Conclusions

The objective of this paper was to determine the extent of the usage of linked open data within biomedical literature. Applying the PRISMA statement for literature reviews, forty-six papers have been included in the analysis and their keywords identified. Twenty-three keywords had a frequency greater than one (and the maximum frequency was 15), 146 keywords had a frequency equal to one. "Information Science Category" and "Analytical, Diagnostic and Therapeutic Techniques and Equipment Category" of MeSH were the main MeSH categories keywords belong to. Future work includes applying association rules learning (shopping basket analysis) to the keywords.

Acknowledgements

This research has been supported by the Health Informatics Centre (HIC), Institutionen för Lärande, Informatik, Management och Etik (LIME), Karolinska Institutet, Stockholm, Sweden.

References

[1] T. Berners-Lee, J. Hendler, O. Lassila, The Semantic Web, *Scientific American* 284 (2001), 34-43.
[2] C. Bizer, T. Heath, T. Berners-Lee, Linked data, the story so far, *International Journal on Semantic Web and Information Systems (IJSWIS)* 5 (2009), 1-22.
[3] G. Klyne, J.J. Carroll, editors, Resource Description Framework (RDF): Concepts and Abstract Syntax. [Online, 2003], W3C Proposed Recommendation 15 December 2003. Available at http://www.w3.org/TR/rdf-concepts/ Last Access: 22 November 2018.
[4] S. Bonacina, Linked Open Data in Health and Clinical Care. A Review of the Literature, *European Journal for Biomedical Informatics – EJBI* 12 (2016), en2-en11..
[5] T. Berners-Lee, WWW past & future. [Online, 2003]. Available at http://www.w3.org/2003/Talks/0922-rsoc-tbl/. Last Access: 22 November 2018.
[6] W3C OWL Working Group, OWL 2 Web Ontology Language Document Overview (Second Edition). 11 December 2012. W3C Recommendation. URL: http://www.w3.org/TR/ owl2-overview/ Last Access: 22 November 2018.
[7] E. Prud'hommeaux, A. Seaborne, editors, SPARQL Query Language for RDF. W3C Recommendation 15 January 2008. [Online, 2008]. Available from: http://www.w3.org/TR/rdf-sparql-query/ Last Access: 22 November 2018.
[8] A. Liberati, D.G. Altman, J. Tetzlaff, C. Mulrow, P.C. Gøtzsche, J.P.A. Ioannidis, et al, The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration, *BMJ* 339 (2009), b2700.
[9] R.G. McGee, J.C. Craig, What is being published? A word cloud of titles from the Journal of Paediatrics and Child Health, *J Paediatr Child Health*, 48 (2012), 452.

Address for correspondence

Stefano Bonacina, M.Sc. | Ph.D. | Senior Researcher in Health Informatics | Health Informatics Centre | Department of learning, informatics, management and ethics | Karolinska Institutet | Tomtebodavägen 18a, 171 77 Stockholm, Sweden.
Phone: +46 8 524 83346 Email: stefano.bonacina@ki.se
Web site: http://ki.se/HIC