World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:10, 2018

Mining Scientific Literature to Discover Potential Research Data Sources: An Exploratory Study in the Field of Haemato-Oncology

Authors : A. Anastasiou, K. S. Tingay

Abstract : Background: Discovering suitable datasets is an important part of health research, particularly for projects working with clinical data from patients organized in cohorts (cohort data), but with the proliferation of so many national and international initiatives, it is becoming increasingly difficult for research teams to locate real world datasets that are most relevant to their project objectives. We present a method for identifying healthcare institutes in the European Union (EU) which may hold haemato-oncology (HO) data. A key enabler of this research was the bibInsight platform, a scientometric data management and analysis system developed by the authors at Swansea University. Method: A PubMed search was conducted using HO clinical terms taken from previous work. The resulting XML file was processed using the bibInsight platform, linking affiliations to the Global Research Identifier Database (GRID). GRID is an international, standardized list of institutions, including the city and country in which the institution exists, as well as a category of the main business type, e.g., Academic, Healthcare, Government, Company. Countries were limited to the 28 current EU members, and institute type to 'Healthcare'. An article was considered valid if at least one author was affiliated with an EU-based healthcare institute. Results: The PubMed search produced 21,310 articles, consisting of 9,885 distinct affiliations with correspondence in GRID. Of these articles, 760 were from EU countries, and 390 of these were healthcare institutes. One affiliation was excluded as being a veterinary hospital. Two EU countries did not have any publications in our analysis dataset. The results were analysed by country and by individual healthcare institute. Networks both within the EU and internationally show institutional collaborations, which may suggest a willingness to share data for research purposes. Geographical mapping can ensure that data has broad population coverage. Collaborations with industry or government may exclude healthcare institutes that may have embargos or additional costs associated with data access. Conclusions: Data reuse is becoming increasingly important both for ensuring the validity of results, and economy of available resources. The ability to identify potential, specific data sources from over twenty thousand articles in less than an hour could assist in improving knowledge of, and access to, data sources. As our method has not yet specified if these healthcare institutes are holding data, or merely publishing on that topic, future work will involve text mining of data-specific concordant terms to identify numbers of participants, demographics, study methodologies, and sub-topics of interest.

Keywords : data reuse, data discovery, data linkage, journal articles, text mining
Conference Title : ICCSIB 2018 : International Conference on Cybermetrics, Scientometrics, Informetrics and Bibliometrics
Conference Location : Barcelona, Spain
Conference Dates : October 29-30, 2018