How comprehensive is the PubMed Central Open Access full-text database? *

Jiangen He[1 [0000−0002−3950−6098]] and Kai Li[1 [0000−0002−7264−365X]]

Department of Information Science, Drexel University, Philadelphia PA, 19104, USA.
jiangen.he@drexel.edu, kl696@drexel.edu

Abstract. The comprehensiveness of database is a prerequisite for the quality of scientific works established on this increasingly significant infrastructure. This is especially so for large-scale text-mining analyses of scientific publications facilitated by open-access full-text scientific databases. Given the lack of research concerning the comprehensiveness of this type of academic resource, we conducted a project to analyze the coverage of materials in the PubMed Central Open Access Subset (PMCOAS), a popular source for open-access scientific publications, in terms of the PubMed database. The preliminary results show that the PMCOAS coverage is in a rapid increase in recent years, despite the vast difference by MeSH descriptor.

Keywords: Database coverage · PubMed Central Open Access· PubMed.

1 Introduction

Database has become a central piece of scientific infrastructure in our contemporary data-driven mode of scientific practice. The increasing volumes of data stored in structured formats gradually became an indispensable source for scientific discoveries in nearly every knowledge domain. However, one question that often shrouds this source is how comprehensive the database is as compared to the reality the database is claimed to represent.

A large number of studies in the field of quantitative studies of science have been devoted to this question since the end of the 20th century: they have compared various parameters, especially the number of documents, references, and journals covered, among databases such as Web of Science, Scopus, and Google Scholar [2, 3, 5, 8, 6, 7]. Besides these metadata-driven databases, another important use case of scholarly databases developed more recently is large-scale text-mining analyses facilitated by open-access full-text publications. PubMed Central Open Access Subset (PMCOAS) is an important source for this purpose. It is a collection of open-access materials in the overall PubMed database. Thanks to its free policy and large size, this subset has been frequently used in studies analyzing the full-text characteristics of scientific corpus (e.g., [1, 4, 9]).

* Jiangen He wishes to thank the support of the National Science Foundation (Award Number: 1633286) and Kai Li wishes to thank the support of the Institute of Museum and Library Services (Award Number: RE-07-15-0060-15).

Similar with other types of scientific research involving databases, the quality of these large-scale text analyses also heavily depends upon the comprehensiveness of the indexed scientific texts. Yet, very few studies have been taken to analyze the coverage of this type of database. In order to bridge this gap, the present study aims to examine the degrees to which PMCOAS covers the PubMed database. Our methodology and some initial results are reported in this poster. The next step of our work is discussed by the end of this proposal.

2 Method

2.1 Data

MEDLINE/PubMed (PubMed) The MEDLINE/PubMed database contains over 26 million journal citations and abstracts for biomedical literature from around the world, which is often cited as the largest database of biomedical publications . Moreover, it is also the superset of the PMCOAS dataset. Because of both reasons, PubMed was selected as the baseline for the coverage of PMCOAS. In this analysis, we used the most recent baseline set of the PubMed data as released on November 28, 2017. [1].

PubMed Central Open Access Subset (PMCOAS) PMCOAS is a full-text document repository covering all open-access literature in the PubMed Central (PMC) database, which itself is a free-access full-text archive of scientific publications that has deep connections to PubMed. Figure 1.(a) shows the relationship among PubMed, PMC, and PMCOAS collections. Even though both PMC and PMCOAS can be seen as subsets of PubMed, a small number of publications in PMC and PMCOAS are not indexed by PubMed. In this project, all publications that are indexed in PMCOAS but not PubMed were removed from our analysis to ensure the comparability between PubMed and PMCOAS publications. We created a list of PMC publications that were indexed in the 2018 PubMed baseline and then retrieved the publication XML files in PMCOAS for computing the coverage.

Medical Subject Headings (MeSH) Most publications in PubMed were indexed by one or more MeSH descriptors . These subject headings describe how each publication is located in different scientific fields. In this study, we mapped the coverage of PMCOAS publication on the MeSH terms to understand the comprehensiveness of PMCOAS in different fields. We used the 2018 version of the MeSH vocabulary [2].

[1] ftp://ftp.ncbi.nlm.nih.gov/pubmed/baseline
[2] ftp://nlmpubs.nlm.nih.gov/online/mesh/MESH_FILES/xmlmesh/

2.2 Coverage

We conceptualize PMCOAS coverage as the ratio of publications in PMCOAS over PubMed, which is calculated as follows.

$$coverage(r) = \frac{N_{pmcoas}(r)}{N_{pubmed}(r)} \tag{1}$$

where r is the restriction and N is the number of publications.

However, a special case is the coverage for a specific MeSH descriptor. Because of the hierarchical structure of MeSH thesaurus, a MeSH descriptor could have other broader or narrower descriptors. Given this structure, to count all publications connected to one descriptor, we also included publications under all its narrower descriptors. For example, all publications under 'Eye Diseases [C11]' were added to the total number of publications for the descriptor 'Diseases [C]'. Therefore, we measure the coverage for a MeSH descriptor m as follows.

$$coverage(m) = \frac{N_{pmcoas}(m) + \sum N_{pmcoas}(m')}{N_{pubmed}(m) + \sum N_{pubmed}(m')} \tag{2}$$

where m' is the descriptors directly or indirectly subordinated to m. We only used the major MeSH descriptors of publications in the coverage computation.

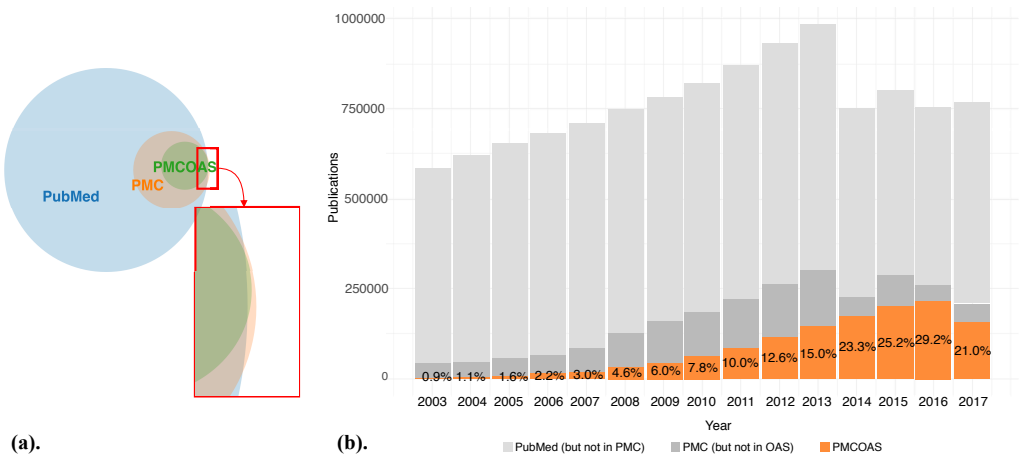

Fig. 1. The coverage from 2003 to 2017.

3 Results

In our total sample, there were 26,481,639 PubMed publications and 1,406,839 PMCOAS publications. The overall coverage of PMCOAS within PubMed is 5.31%.

3.1 Coverage over time

Figure 1.(b) shows the change of PMCOAS coverage from 2003 to 2017. This figure shows an overall upward trend in PMCOAS coverage. One exception to this trend is the year of 2017, which might be able to be explained by the delay of indexing in both PubMed and PMCOAS.

3.2 Coverage over MeSH descriptors

Table 1 shows PMCOAS coverage over the top-level descriptors in the MeSH vocabulary. As shown in the table, there is a vast difference in the coverage over major knowledge categories in the MeSH system. Most categories have the coverage between 2% to 4%. Being an outlier and the smallest category in MeSH, Geographicals have the coverage of nearly 20%.

Table 1. The coverage of top-level descriptors.

Descriptors	PubMed	PMCOAS	Coverage
Anatomy	1,614,418	27,148	1.68%
Organisms	802,799	26,879	3.35%
Diseases	3,124,614	55,284	1.77%
Chemicals and Drugs	3,162,663	61,061	1.93%
Analytical, Diagnostic and Therapeutic ...	5,858,321	174,965	2.99%
Psychiatry and Psychology	2,574,914	77,902	3.03%
Phenomena and Processes	5,376,267	326,172	6.07%
Disciplines and Occupations	1,334,133	30,534	2.29%
Anthropology, Education, Sociology, ...	1,496,982	43,138	2.88%
Technology, Industry, and Agriculture	846,558	25,639	3.03%
Humanities	362,203	8,089	2.23%
Information Science	940,935	59,649	6.43%
Named Groups	584,740	14,418	2.47%
Health Care	4,448,055	156,763	3.52%
Geographicals	3,208	607	18.92%

4 Search System

In the Results section, we only displayed a small piece of results based on MeSH vocabulary. In order for researchers to explore more detailed results, we developed a search system, where users could search the MeSH terms they are interested in and gain a better understanding of how PMCOAS could fulfill their needs. The beta version of the system can be accessed via http://jiangenhe.com/pmc_coverage.

5 Conclusion and Future Work

In this project, we aim to analyze the comprehensiveness of PMCOAS database, a popular source for text-mining analysis focusing on medical publications, as compared to the PubMed database. Our initial results suggest that PMCOAS is able to cover 5.31% of all publications in the PubMed database. Despite the overall increase of the coverage during the past decade, there is a large gap among different scientific fields, as represented by the MeSH subject heading. Moreover, we also designed an online system for users to explore our data, to gain deeper insights into their own research interests.

In the next step of this project, we will investigate the reference coverage of PMCOAS, i.e., the ratio of publications with in-text citation context in PMCOAS over PubMed publications, which is a critical factor affecting the effectiveness of citation content analysis. Another exciting direction is to improve our design for the search system so that researchers can gain better knowledge about the validity of PMCOAS as a potential source for their full-text research.

References

1. Agarwal, S., Yu, H.: Figure summarizer browser extensions for PubMed Central. Bioinformatics **27**(12), 1723–1724 (2011)
2. Bauer, K., Bakkalbasi, N.: An Examination of Citation Counts in a New Scholarly Communication Environment. D-Lib Magazine **11**(9) (Sep 2005), http://www.dlib.org/dlib/september05/bauer/09bauer.html
3. Bergman, E.M.L.: Finding citations to social work literature: The relative benefits of using Web of Science, Scopus, or Google Scholar. The journal of academic librarianship **38**(6), 370–379 (2012)
4. Flórez-Vargas, O., Brass, A., Karystianis, G., Bramhall, M., Stevens, R., Cruickshank, S., Nenadic, G.: Bias in the reporting of sex and age in biomedical research on mouse models. Elife **5** (2016)
5. Kulkarni, A.V., Aziz, B., Shams, I., Busse, J.W.: Comparisons of citations in Web of Science, Scopus, and Google Scholar for articles published in general medical journals. Jama **302**(10), 1092–1096 (2009)
6. Meho, L.I., Rogers, Y.: Citation counting, citation ranking, and h-index of human-computer interaction researchers: A comparison of scopus and web of science. Journal of the American Society for Information Science and Technology **59**(11), 1711–1726. https://doi.org/10.1002/asi.20874
7. Mingers, J., Lipitakis, E.A.E.C.G.: Counting the citations: a comparison of web of science and google scholar in the field of business and management. Scientometrics **85**(2), 613–625 (Nov 2010). https://doi.org/10.1007/s11192-010-0270-0
8. Trapp, J.: Web of Science, Scopus, and Google Scholar citation rates: a case study of medical physics and biomedical engineering: what gets cited and what doesn't? Australasian Physical & Engineering Sciences in Medicine **39**(4), 817–823 (Dec 2016). https://doi.org/10.1007/s13246-016-0478-2
9. Verspoor, K., Cohen, K.B., Hunter, L.: The textual characteristics of traditional and Open Access scientific journals are similar. BMC bioinformatics **10**(1), 183 (2009)