PKG2023S4 (1781-Dec. 2023) Features

We built PKG with bio-entities extracted from PubMed titles and abstracts, author name disambiguation results of PubMed authors, and the integrated multi-source information. This dataset is freely available on https://pubmedkg.github.io, which contains both the PubMed raw data and PKG dataset to facilitate the application of PKG.

The new version PKG, PKG2023S4 (1781-Dec. 2023), updated the previous PKG version with PubMed 2024 baseline files, extracted bio-entities, author disambiguation results, and extended author information. In addition, PKG2.0 established links between Paper, Patent, and Clinical trial studies from multiple perspectives.

Table 1 summarizes the descriptions of main tables.

Table 1. Main table details.

Table	**# of Lines**	**# of Distinct paper, trial, and patent identify**	**# of Distinct AND_IDs**	**Short description**
Articles	36,554,965	36,551,113	-	Table containing all articles from PubMed.
Patents	1,344,469	1,344,469	-	Table containing biomedical patents from USPTO.
Clinical trial studies	480,795	480,795	-	Table containing all clinical trial studies from ClinicalTrials.gov.
Authors	26,217,594	-	26,217,594	Table containing all authors in PKG2.0.
AuthorShips	160,848,959	35,809,481	26,217,595	Table containing PubMed authorships.
Bio-entities	464,643,559	31,631,391	-	Table containing all types of extracted bio-entities from papers by BERN2.
Affiliations	96,296,020	25,341,875	19,652,501	Table containing affiliations and their extracted fine-grained items.
Researcher_Employment	2,354,570	-	649,499	Table containing employment history from ORCID.
Researcher_Education	1,849,409	-	551,208	Table containing educational background from ORCID.
NIH_Projects	53,424,596	2,501,621	191,969	Table containing projects from NIH ExPORTER and mapping relation between PI_ID, PMID, and AND_ID.
Journals	29,215,730	29,212,322	-	Table containing journal information from SciMago for the year in which the paper was published.

Table 2 describes the date coverage and the version information of data sources.

Table 2. Date coverage and version information of data sources

Data Source	Start Year	End Year	Version Information
PubMed 2024 baseline files	1781	2023	The PubMed 2024 baseline files were released in December 2023. It also includes some papers published after 2023 and majority of them are preprints.
Author-ity 2021 dataset	1781	2018	The dataset was generated based on PubMed 2019 baseline files, and was released on Apr 22, 2021.
Semantic Scholar dataset	1786	2023	The dataset was downloaded via api in Feb. 2024.
NIH ExPORTER dataset	1985	2023	The articles marked with projects span from 1981 to 2023, and project details cover from 1985 to 2023. The dataset was downloaded in Feb. 2024.
Employment History Data from ORCID	1913	2023	The dataset was released on October 22, 2023. ORCID publishes the data once per year.
Educational Background Data from ORCID	1913	2023	The dataset was released on October 22, 2023. ORCID publishes the data once per year.
MapAffil 2021 dataset	1975	2018	The dataset is based on a snapshot of PubMed taken in the first week of October, 2016, and was released on April 5, 2021.
USPTO patents	1976	2023	The dataset was download from patentview in Jan. 2024.
ClinicalTrial.gov	1999	2023	The dataset was download from AACT in Jan. 2024.
Patent-cite-Article	1976	2023	Citations from patents to articles(run by M. Marx and A. Fuegi). The dataset was downloaded in Jan. 2024.
SCImago	1999	2022	The SCImago Journal & Country Rank is a publicly available portal that includes the journals and country scientific indicators developed from the information contained in the Scopus® database (Elsevier B.V.). These indicators can be used to assess and analyze scientific domains. Journals can be compared or analyzed separately. We just focus on the journals scientific indicators.

Please cite the authors in any work or product based on this material:

Xu Jian, Kim Sunkyu, Song Min, Jeong Minbyul, Kim Donghyeon, Kang Jaewoo, Rousseau Justin F., Li Xin, Xu Weijia, Torvik Vetle I., Bu Yi, Chen Chongyan, Ebeid Islam akef, Li Daifeng & Ding Ying. Building a PubMed knowledge graph. Scientific Data 7, 205 (2020). https://doi.org/10.1038/s41597-020-0543-2