PKG2022S4 (1781-Dec. 2022) 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 dataset.

The new version PKG, PKG2022S4 (1781-Dec. 2022), updated the previous PKG version with PubMed 2023 baseline files, and extracted bio-entities, author disambiguation results, extended author information and citations integrated from iCite, OpenCitations and PubMed.

Table 1 summarizes the descriptions of main tables.

Table 1. Main table details.

Table	# of Lines	# of Distinct PMIDs	# of Distinct AND_IDs	Short description
A01_Articles	34,960,699	34,957,126	-	Table containing PubMed articles' bibliographic information.
A02_AuthorList	150,607,698	34,246,887	22,408,066	Table containing PubMed authors and AND_IDs.
B08_ORCID_Education	144,541,025	-	537,723	Table containing educational background from ORCID.
B09_ORCID_Employment	653,225	-	626,439	Table containing employment history from ORCID.
C03_Affiliation_Merge	84,376,512	23,843,591	-	Table containing affiliations and their extracted fine-grained items.
C04_ReferenceList	748,980,377	27,056,926	-	Table containing reference relations between PMID and reference PMID.
C05_NIH_PubMed	50,091,901	2,410,180	166,681	Table containing projects from NIH ExPORTER and mapping relation between PI_ID, PMID, and AND_ID.
C06_BERN2_Main	440,184,212	30,178,652	-	Table containing all types of extracted bio-entities by BERN2.

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 2023 baseline files	1781	2022	The PubMed 2023 baseline files were released in Dec. 2022.
Bio_entity dataset	1781	2021	Bio entities are extracted by the BERN2, a state-of-the-art bio-entity extraction algorithm.
Author-ity 2018 dataset	1865	2018	The dataset is based on a snapshot of PubMed taken in December, 2018, and was released on Apr. 22, 2021.

Semantic Scholar dataset	1786	2023	The dataset was obtained in May 2023 through the Semantic Scholar API.
NIH ExPORTER dataset	1985	2022	The articles marked with projects span from 1981 to Dec. 2022, and project details cover from 1985 to Dec. 2022. The dataset was downloaded in May. 2023.
Employment History Data from ORCID	1913	2022	The dataset was released on Jan, 2023. ORCID publishes the data once per year.
Educational Background Data from ORCID	1913	2022	The dataset was released on Jan, 2023. ORCID publishes the data once per year.
MapAffil 2018 dataset	1975	2018	The dataset is based on a snapshot of PubMed taken in December, 2018, and was released on May. 07, 2021.
Affiliation Parser Library	-	-	Fast and simple parser for MEDLINE and PubMed Open-Access affiliation string, which was published on Sep. 02, 2021. We apply it to parse multiple fields from the affiliation string, including department, institution, zip code, location, and country.
ReferenceList	-	2022	The C04_ReferenceList contains 748,980,377 citations from 27,056,926 articles. The sources of data integration include PubMed's own citation data, NIH Open Citation Collection(run by iCite) and the OpenCitations Index of Crossref open DOI-to-DOI citations(run by OpenCitations). Compared with PubMed's own citation data (the amount of data is 282,142,987), it increased by 466,837,390.

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