PKG2020S4 (1781-Dec. 2020) Features

We built PKG with bio-entities extracted from PubMed abstracts, author name disambiguation results of PubMed authors, and the integrated multi-source information. This dataset is freely available on http://er.tacc.utexas.edu/datasets/ped (folder for CSV files). It contains seven comma-separated value (CSV) files named "Author_List," "Bio_entities_Main," "Bio_entities_Mutation," "Affiliations," "Researcher_Employment," "Researcher_Education," and "NIH_Projects". PubMed raw data are not included into CSV files set because the amount of PubMed raw data is too large and they are not generated or altered by our methods. PubMed raw data can be freely downloaded from PubMed website.

We also provide the download link http://er.tacc.utexas.edu/datasets/ped (folder for MySQL export files), which contains both the PubMed raw data and PKG dataset to facilitate the application of PKG dataset.

The new version PKG, PKG2020S4 (1781-Dec. 2020), updated the previous PKG version with PubMed 2021 baseline files, PubMed daily updates files (up to Jan. 4th 2021), and extracted bio-entities, author disambiguation results, extended author information, Scimago that containing journal information, and WOS citations which contains reference relations between PMID and reference PMID and extracted from WOS.

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	31,928,777	31,926,861	-	Table containing PubMed articles' bibliographic information.
A02_AuthorList	131,446,038	31,270,411	18,519,492	Table containing PubMed authors and AND_IDs.
B10_BERN_Main	295,921,671	20,136,150	-	Table containing all types of extracted bio-entities by BioBERT.
B12_BERN_Mutation	1,415,427	320,025	-	Table containing additional items of mutations from Bio-entities Main file.
C03_Affiliation_Merge	62,015,712	20,941,553	9,502,394	Table containing affiliations and their extracted fine-grained items.
B08_ORCID_Education	934,507	-	448,110	Table containing educational background from ORCID.
B09_ORCID_Employment	1,194,697	-	531,916	Table containing employment history from ORCID.
C05_NIH_PubMed	22,946,601	1,886,856	116,530	Table containing projects from NIH ExPORTER and mapping relation between PI_ID, PMID, and AND_ID.
B14_Scimago	539,714	-	-	Table containing journal information from Scimago. Connections can setup by field B14_Scimago.eISSN and

				A01_Articles. Journal ISSN
C04_ReferenceList	633,401,975	23,856,949	-	Table containing reference relations between PMID and reference PMID. It extracted from Web of Sciences.

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 2021 baseline files and daily update files (up to Jan. 4, 2021)	1781	Jan. 4, 2021	The PubMed 2021 baseline files were released in December 2020. The Daily Update Files are update records after baseline files published. The latest date of daily update files is Jan. 4, 2021.
Bio_entity dataset	1781	Jan. 4, 2021	Bio entities are extracted by the BERN, a state-of-the-art bio-entity extraction algorithm. The extracted results covers up to Jan. 4, 2021.
Author-ity dataset	1865	2008	The dataset was generated based on PubMed 2009 baseline files. It also includes AND results of 93,228 papers published after 2008, and majority of them are preprints.
Semantic Scholar dataset	1786	Dec, 2020	The dataset was released on Dec 1, 2020.
NIH ExPORTER dataset	1985	Dec, 2020	The articles marked with projects span from 1981 to Dec. 2020, and project details cover from 1985 to Dec. 2020. The dataset was downloaded in Jan. 2021.
Employment History Data from ORCID	1913	Oct, 2020	The dataset was released on October, 2020. ORCID publishes the data once per year.
Educational Background Data from ORCID	1913	Oct, 2020	The dataset was released on October, 2020. ORCID publishes the data once per year.
MapAffil 2016 dataset	1975	2017	The dataset is based on a snapshot of PubMed taken in the first week of October, 2016, and was released on April 5, 2018.
Affiliation Parser Library	1786	Dec, 2020	Fast and simple parser for MEDLINE and PubMed Open-Access affiliation string, which was published on March 15, 2018. We apply it to parse multiple fields from the affiliation string, including department, institution, zip code, location, and country.
SCImago	1999	2018	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.
ReferenceList	-	2020	The C04_ReferenceList contains 633401975 citations from 23856949 articles. The sources of data integration include PubMed's own citation data, NIH's opencitation collection,

			opencitations(run by David Shotton and Silvio Peroni) and the citation data from WOS. Compared with PubMed's own citation data (the amount of data is 223261597), it increased by 410140378. Compared with the previous version of PKG, the WOS citation data (the amount of data is 447596685), it increased by 185805290.

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