# PKG2021S4 (1781-Dec. 2021) 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 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, PKG2021S4 (1781-Dec. 2021), updated the previous PKG version with PubMed 2022 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 | 33,405,863 | 33,405,863 | - | Table containing PubMed articles' bibliographic information. |
| A02_AuthorList | 140,752,944 | 32,727,256 | 18,337,630 | Table containing PubMed authors and AND_IDs. |
| B08_ORCID_Education | 1,466,087 | - | 430,514 | Table containing educational background from ORCID. |
| B09_ORCID_Employment | 1,736,503 | - | 497,777 | Table containing employment history from ORCID. |
| C03_Affiliation_Merge | 72,746,190 | 22,356,026 | 11,398,491 | Table containing affiliations and their extracted fine-grained items. |
| C04_ReferenceList | 674,014,525 | 24,975,782 | - | Table containing reference relations between PMID and reference PMID. |
| C05_NIH_PubMed | 49,436,604 | 2,339,383 | 173,737 | Table containing projects from NIH ExPORTER and mapping relation between PI_ID, PMID, and AND_ID. |
| C06_BERN2_Main | 388,307,224 | 28,007,619 | - | 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 2022 baseline files | 1781 | Apr, 2022 | The PubMed 2022 baseline files were released in Dec. 2021. |
| Bio_entity dataset | 1781 | Apr, 2022 | Bio entities are extracted by the BERN2, a state-of-the-art bio-entity extraction algorithm. |
| 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, 2021 | The dataset was released on Jan 1, 2022. |
| NIH ExPORTER dataset | 1985 | Dec, 2021 | The articles marked with projects span from 1981 to Dec. 2021, and project details cover from 1985 to Dec. 2021. The dataset was downloaded in Feb. 2022. |
| Employment History Data from ORCID | 1913 | Oct, 2021 | The dataset was released on October, 2021. ORCID publishes the data once per year. |
| Educational Background Data from ORCID | 1913 | Oct, 2021 | The dataset was released on October, 2021. 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, 2021 | 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. |
| ReferenceList | - | 2021 | The C04_ReferenceList contains 674014525 citations from 24975782 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 268400936), it increased by 405613589. |

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