

PKG2021S4 (1781-Dec. 2021) Database

Description

The database recovery process was successfully tested under ubuntu 18.04.3, using a mysql 5.7.38. Please do the following steps to restore the PKG2021S4 to your MySQL database.

1. Download all the compressed files and their MD5 files;
2. For files ending with “part*”, you should combine them to generate a single file like:

```
cat A02_AuthorList.sql.gz.part* > A02_AuthorList.sql.gz
```

3. Verify each compressed file with their corresponding MD5 file. For example, we can use the following command to verify if the file “A01_Articles.sql.gz” download without any damage:

```
md5sum -c A01_Articles.sql.gz.md5sum
```

4. Create a new Database in your MySQL server, and make sure the new database “Charset” is set to: utf8mb4, and “Order rule” is set to: utf8mb4_bin.

5. Next, you can inject every table into the target database using the command like:

```
gunzip < A01_Articles.sql.gz | mysql -uusername -ppassword destinationDatabaseName
```

The above command will import the table A01_Articles.sql.gz into destination database.

Similarly,

The tables beginning with A are the original data tables of PubMed. The tables with the beginning of C are synthesized by extracting key information for subsequent statistics and calculations.

The original document of each field description of PubMed can be found at:

https://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html

1 A01_Articles

Specific information for each article.

Column Name	Description
id	Automatically incremented identifier
PMID	Unique ID assigned by PubMed to identify PubMed articles.
PMID_Version	PMID version number, which is the record corresponding to all PMID versions in this table. That is, if the PMID version number is different, there are different records in this table, and the id field of each record is different. In other related article tables (such as C01_Article_simple), only the record with the highest version number is kept.
MedlineCitation_Owner	Organization responsible for creating and verifying citations, including: NLM, NASA, PIP, KIE, HSR, HMD, SIS, NOTNLM

MedlineCitation_Status	It is the stage of the article. There are seven possible values: Completed In-Process PubMed-not-MEDLINE In-Data-Review Publisher MEDLINE OLDMEDLINE
Journal_JournalIssue_PubDate_Year	Publication year of the current article.
Journal_JournalIssue_PubDate_MedlineDate	Publication year of the current article. If <i>Journal_JournalIssue_PubDate_Year</i> is null, the publication year of the current article is recorded in this field

2 A02_AuthorList

Specific information for each author.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
Au_Order	Author order of the current authors in the author list of current articles.
LastName	Last name of the current author.
ForeName	Current author's name excluding the last name and suffix
LastNameForeName	MD5 code generated by the LastName and the first letter of ForeName
AuthorNum	Co-author number of the current article
Vetle_aid	Unique author ID allocated by Vetle.
S2ID	Unique author ID allocated by Semantic Scholar
AID	Unique author ID (The final author disambiguation result. S2id is the main source, supplemented by other disambiguation results)

3 A03_KeywordList

Article keyword information: keyword information in this table is provided by the data producer.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
Keyword	Keywords of the current article

4 A04_Abtract

The abstract of each article.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.

AbstractText	The abstract of the current Article
--------------	-------------------------------------

5 A05_GrantList

Grants details of each article.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
GrantID	Funding identifiers, including research grant numbers or contract numbers (or both) that are financially supported by the US Public Health Service or any agency of the National Institutes of Health (NIH)
Project_Number	The NIH project number corresponding to the funded article. Each article may be funded by multiple projects. Multiple project numbers are separated by ",".

6 A06_MeshheadingList

Mesh Heading details of each article.

Mesh Heading refers to the NLM control vocabulary and medical subject heading (MeSH®), which is used to characterize the content of the articles represented by MEDLINE citations.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
DescriptorName	Descriptors extracted from each article
DescriptorName_MajorTopicYN	If the MeSH descriptors assigned to the article is the Key word of the article, the value is Y, otherwise it is N.
DescriptorName_UI	Identify MeSH's unique encoding for each descriptor and qualifier
QualifierName	Qualifiers, including numbers and words

7 A07_SupplMeshList

Supplementary conceptual terms and protocol terms for each article.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
SupplMeshName	Supplementary conceptual term
SupplMeshName_Type	The type of supplementary conceptual term
SupplMeshName_UI	MeSH unique identifier for supplementary protocols and

	diseases
--	----------

8 A08_ChemicalList

The chemical substances and registry number covered in each article.

Registry Number refers to a code assigned by Chemical Abstracts Service to a specific chemical substance.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
RegistryNumber	Unique Identifier to a specific chemical substance assigned by Chemical Abstracts Service
NameOfSubstance	The name of the specific chemical substance

9 A09_CommentsCorrectionsList

Reference information for each article, including the source, type, and PMID of the reference.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
RefSource	Reference sources
RefType	Reference type
RefPMID	The PMID of the reference
RefNote	Corrections to records with incorrect references

10 A10_DatabankList

The search number of the molecular sequence database that appears in the PubMed article. The search number can find the information of the corresponding chemical molecule from the established molecular sequence database, avoiding the use of lengthy molecular formulas and graphics in the article.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
DataBankName	Name of the molecular sequence database

11 A11_PersonalNameSubjectList

Individuals' names appear in <PersonalNameSubject> for citations that contain a biographical note or obituary, or are entirely about the life or work of an individual or individuals. Data is entered in

the same format as A02_AuthorList.

12 A12_InvestigatorList

Each article corresponds to the NASA-funded principal investigator (PI) information, and they participated in the discussion and research of the article (but not necessarily the author).

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
AffiliationInfo_Affiliation	The affiliation which the researchers belong to
AffiliationInfo_Identifier	Unique identifier of the affiliation

13 A13_AffiliationList

Extracted affiliation information.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
Au_Order	Author order of the current authors in the author list of current articles.
Affiliation	Affiliation string.
Affiliation_order	For authors belonging to multiple affiliations, use numbers to identify different affiliations (in no particular order)

14 A14_ReferenceList

Reference information (we use the information in this table to generate the author's self-cited record during the disambiguation of strong features).

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
RefRank	Reference ranking
RefCitation	Reference sources (journals, year of publication, etc.)
RefArticleId	The PMID of the reference
RefIdType	Reference source database (e.g., PubMed)

15 B01_Descriptor

Specific information of Descriptor Name of each article (used for the classification of articles).

Column Name	Description
-------------	-------------

DescriptorUI	The unique identifier of the descriptor
DescriptorName	The unique term of the descriptor
DescriptorClass	The corresponding classification of the descriptor (such as subject index, document type)
DateCreated	The date of the recording
DateEstablished	The date available for retrieval
Annotation	Text information corresponding to the descriptors
HistoryNote	History of text messages designed to help online searchers
NLMClassificationNumber	The NLM classification numbers assigned to the index terms
OnlineNote	Text messages designed to help online searchers
PublicMeSHNote	Change history of information
ConsiderAlso	Cross-reference, point to similar descriptors
PreviousIndex	Different usage in Descriptors and SCRs: In Descriptors: Free-text field referring to Descriptors or Descriptor/Qualifier combinations which were used to index the concept in the MEDLINE databases before the Descriptor was created.
TreeNumberList	One or more sets of elements in the descriptor or qualifier record

16 B02_ProjectList_NIH

Project information published on the NIH official website(<https://reporter.nih.gov/exporter/projects>).

Column Name	Description
APPLICATION_ID	A unique identifier of the project record in the ExPORTER database.
FULL_PROJECT_NUMBER	Commonly referred to as a grant number, intramural project, or contract number. For grants, this unique identification number is composed of the type code, activity code, Institute/Center code, serial number, support year, and (optional) a suffix code to designate amended applications and supplements.
SUBPROJECT_ID	A unique numeric designation assigned to subprojects of a “parent” multi-project research grant.
FUNDING_ICs	The NIH Institute or Center(s) providing funding for a project are designated by their acronyms (see Institute/Center acronyms). Each funding IC is followed by a colon (:) and the amount of funding provided for the fiscal year by that IC. Multiple ICs are separated by semicolons (;). Project funding information is available only for NIH, CDC, FDA, and ACF projects .
FY	The fiscal year appropriation from which project funds were obligated.
PHR	Submitted as part of a grant application, this statement articulates a project's potential to improve public health.
PI_IDS	A unique identifier for each of the project Principal Investigators. Each PI in the RePORTER database has a unique identifier that is constant from project to project and year to year, but changes may be observed for investigators that have had multiple accounts in the past,

	particularly for those associated with contracts or sub-projects.
PI_NAMES	The name(s) of the Principal Investigator(s) designated by the organization to direct the research project.
PROJECT_START	The start date of a project. For subprojects of a multi-project grant, this is the start date of the parent award.
PROJECT_END	The current end date of the project, including any future years for which commitments have been made. For subprojects of a multi-project grant, this is the end date of the parent award. Upon competitive renewal of a grant, the project end date is extended by the length of the renewal award.
PROJECT_TERMS	Prior to fiscal year 2008, these were thesaurus terms assigned by NIH CRISP indexers. For projects funded in fiscal year 2008 and later, these are concepts that are mined from the project's title, abstract, and specific aims using an automated text mining tool.
PROJECT_TITLE	Title of the funded grant, contract, or intramural (sub)project.
TOTAL_COST	Total project funding from all NIH Institute and Centers for a given fiscal year. Costs are available only for: NIH, CDC, FDA, and ACF grant awards (only the parent record of multi-project grants). NIH intramural projects (activity codes beginning with "Z") in FY 2007 and later fiscal years. NIH contracts (activity codes beginning with "N"). For multi-project grants, Total_Cost includes funding for all of the constituent subprojects. This field will be blank on subproject records; the total cost of each subproject is found in Total_Cost_Sub_Project.
TOTAL_COST_SUB_PROJECT	Applies to subproject records only. Total funding for a subproject from all NIH Institute and Centers for a given fiscal year. Costs are available only for NIH awards.

17 B08_ORCID_Employment

Employment information of scientific personnel from ORCID dataset.

Column Name	Description
ORCID	Unique researcher ID that distinguishes the researcher from others allocated by ORCID.
Department	The department which the researcher belongs to.
BeginYear	The beginning year of the researcher's employment.
Organization	The institution which the researcher belongs to.
City	The city where the researcher works.
Region	The region where the researcher works.
Country	The country where the researcher works.
Identifier	The identifier of an organization.
IdSource	The provider of an organizations' identifier.
EndYear	The end year of the researcher's employment.
AID	Unique author ID (The final author disambiguation result. S2id as the

	main source, supplemented by other disambiguation results)
--	--

18 B09_ORCID_Education

Education information of scientific personnel from ORCID dataset.

Column Name	Description
ORCID	Unique researcher ID that distinguishes the researcher from others allocated by ORCID.
BeginYear	The beginning year of the researcher's education.
Organization	The organization the researcher has been educated.
City	The city that the author belongs to.
Region	The region that the author belongs to.
Country	The country that the author belongs to.
Identifier	The identifier of an organization.
IdSource	The provider of an organizations' identifier.
EndYear	The end year of the researcher's education.
Role	The degree that the researcher received.

19 B11_BERN2_EntityType

Entity Type dictionary.

Column Name	Description
Id	Unique ID assigned to identify entity types. Corresponding to the field Type in table C06_BERN2_Main.
Type	Entity type name.

20 C01_Articles_simple

Simple table of A01_Articles.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
PubYear	Year the article was published.
AuthorNum	The number of authors of this article.
CitedCount	The number of times the article was cited (calculated from the C04_ReferenceList)

22 C02_Authorlist_simple

Simple table of A02_AuthorList.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
PubYear	Year the article was published.
AuthorNum	The number of authors of this article.
Au_Order	Ranking of the current author's signature in the paper.
AID	Unique identifier for the author in our PKG.
BeginYear	The year this author published the first paper.
CurrentAge	Author's academic age: CurrentAge=PubYear-BeginYear+1

22 C03_Affiliation_merge

Merging Vetle_Map and A13_AffiliationList information, including the parsed organization information, such as Zip code, Location, Country, etc.

Column Name	Description
Id	Unique ID assigned to each record.
PMID	Unique ID assigned by PubMed to identify PubMed articles.
Au_Order	Author order of an article
AffiliationOrder	Affiliation order of one author, in case one author has multiple affiliations.
Affiliation	Affiliation text.
Department	Department parsed from Affiliation.
Institution	Institution parsed from Affiliation.
Email	Email parsed from Affiliation.
Zipcode	Zipcode parsed from Affiliation.
Location	Location parsed from Affiliation.
Country	Country parsed from Affiliation.
AID	Author disambiguation ID.
City	City parsed from Affiliation.
State	State parsed from Affiliation.
Vetle_Country	Country abbreviation.
Type	Affiliation type, such as COM, EDU, ORG, et.al.
Lat	Latitude of the Affiliation.
Lon	Longitude of the Affiliation.
Fips	Fips code.

23 C04_ReferenceList

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.

Column Name	Description
Id	Unique ID assigned to each record.
PMID	Unique ID assigned by PubMed to identify PubMed articles.
RefArticleId	PMIDs of references.

24 C05_NIH_PubMed

According to PMID and the author's name (full name of the last name and initials), match the author AID and NIH Principal Investigator (PI) number PIID to generate correspondence table C05, including PIID, AID, CORE_PROJECT_NUMBER, PIID, Application_ID, etc.

Column Name	Description
Id	Unique ID assigned to each record.
PI_ID	A unique identifier for each of the project Principal Investigators. Each PI in the RePORTER database has a unique identifier that is constant from project to project and year to year, but changes may be observed for investigators that have had multiple accounts in the past, particularly for those associated with contracts or sub-projects.
PMID	Unique ID assigned by PubMed to identify PubMed articles.
ProjectNumber	Project number.
SubProjectNumber	Sub project number.
AID	Author disambiguation ID.
PI_Name	The full name of PI.
Application_ID	A unique identifier of the project record in the ExPORTER database.

25 C06_BERN2_MAIN

Entity information set extracted from document titles and abstracts using BERN2.

Column Name	Description
PMID	Unique ID assigned by PubMed to identify PubMed articles.
StartPosition	Start position of mention in an abstract.
EndPosition	End position of mention in an abstract.
Mention	Entity mentioned in an abstract.
Entityid(7 columns)	Normalized entity identifier, include mesh, mim, CL, cellosaurus, NCBITaxon, NCBIGene, CHEBI.

Type	Enumerated type of entity; values include species, disease, gene, drug, mutation, cell_line, cell_type, DNA, RNA.
is_neural_normalized	For diseases and chemicals, BERN2 use hybrid NEN models, which are a combination of both rule-based and neural network-based models. An entity that is not normalized by the rule-based model is then normalized by a neural network-based model.