www.csdata.org

ISSN 2096-2223
CN 11-6035/N

文献 DOI：
10.11922/csdata.2021.0009.zh
数据 DOI：
10.11922/sciencedb.j00001.00213
文献分类：中医学与中药学

收稿日期：2021-01-20
开放同评：2021-04-22
录用日期：2021-08-21
发表日期：2021-09-30

中英对照中医药术语数据集

梁昊[1,3]，吴佳泽[1]，段伦慧[1]，彭清华[2,3,4,5*]，胡志希[1,3,6]，周小青[7]

1. 湖南中医药大学中医学院，长沙 410208
2. 湖南中医药大学中西医结合学院，长沙 410208
3. 中国中医药信息学会中医诊断信息分会，北京 100700
4. 世界中医药学会联合会中医诊断学专业委员会，北京 100020
5. 国际数字医学会数字中医药分会，长沙 410208
6. 中国中西医结合学会标准化技术专业委员会，北京 100700
7. 世界中医药学会联合会翻译专业委员会，北京 100020

摘要：中英对照中医药术语数据集基于人民卫生出版社（PMPH）制定的《中医英语术语（内部草案）》、世界卫生组织（WHO）制定的《WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region》和世界中医药学会联合会（WFCMS）制定的《International Standard Chinese-English Basic Nomenclature of Chinese Medicine》3 个权威术语标准整合而成，旨在促进中医药术语标准化和中医药国际交流。本数据集通过 Python pandas 包及 OCR 技术将数据进行采集、清洗、整理、合并，最终分为 56 类，共整理数据 16 189 条，经合并为 8975 条。本数据集促进了中医术语的规范化，方便了学术交流和中医的继承发扬，同时有助于中医药信息化建设。

关键词：中医药；术语；中英对照

数据库（集）基本信息简介

数据库（集）名称	中英对照中医药术语数据集
数据作者	梁昊、吴佳泽、段伦慧、彭清华、胡志希、周小青
数据通信作者	彭清华（pqh410007@126.com）
数据时间范围	2007–2020
地理区域	世界各国
数据量	1.45 MB
数据格式	*.csv
数据服务系统网址	http://www.dx.doi.org/10.11922/sciencedb.j00001.00213
基金项目	湖南中医药大学教学改革研究项目（2020-JG006）；湖南省科技创新计划（2020RC2061）。
数据库（集）组成	数据集由 1 部数据表构成：表中有 10 个字段，包括 ID、中文简体、中文繁体、拼音、WHO 英文术语、PMPH 英文术语、WFCMS 英文术语、术语的英文解释、类别编码、类别名称。共计 8975 条。

* 论文通信作者
彭清华：pqh410007@126.com

引　言

为了促进中医药及民族医药的国际化，方便在学术科研、教育教学及经济贸易等领域的沟通交流，中国官方及中医药国际组织一直致力于中医药标准化和规范化。术语规范，尤其是中医药英语术语规范，是中医药标准化进程中最基础、最亟待解决的问题[1]。得益于谢竹藩、帅学忠、李照国等前辈们的不懈努力，多部术语标准先后出版并广泛传播。人民卫生出版社（PMPH）制定的《中医英语术语（内部草案）》、世界卫生组织（WHO）制定的《WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region》和世界中医药学会联合会（WFCMS）制定的《International Standard Chinese-English Basic Nomenclature of Chinese Medicine》是当前知晓度和应用率最高的 3 个术语标准[2-3]。然而，近几年在阅读文献和教学中发现，学生、中医从业者、科研工作者对中医术语标准的知晓度和使用频率不高[4]。中英文的中医/中西医结合类学术期刊投稿指南鲜有要求投稿时注意术语规范或推荐使用已经发布的术语标准。中医学作为一个偏传统的学科，尚缺乏标准化和规范化意识，在术语使用上较为随意。究其原因，当前这些术语大部分为纸质版或电子书形式，不利于查找[5]；另外，3 个标准也有差异，虽各有千秋，但也有一些局限性和片面性[6]。因此，我们基于以上术语标准建设中医药术语中英对照数据集，合并词义相同的术语，研究术语差异和建立术语查询系统，为建立更权威、合理、全面的中医药术语数据库打下基础。

1　数据采集和处理方法

1.1　原始数据来源

所有数据来源于人民卫生出版社（PMPH）制定的《中医英语术语（内部草案）》、世界卫生组织（WHO）制定的《WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region》[7]和世界中医药学会联合会（WFCMS）制定的《International Standard Chinese-English Basic Nomenclature of Chinese Medicine》[8]。获得所有中医术语的字段，并进行合并。

1.2　数据采集和处理方法

原始数据为 WHO、PMPH、WFCMS 3 个标准的书籍或电子文档。把原始数据通过 OCR 和 PDF 转化工具整理成规范的数据表（dataframe）格式，命名为 WHO.csv、PMPH.csv、WFCMS.csv。将每个数据表每条记录均以术语的中文简体名称作为唯一字段方便进行数据合并，使用 Python 的 pandas 包对数据进行合并和清洗。最终合并的数据表字段为：ID、中文简体、中文繁体、拼音、WHO 英文术语、PMPH 英文术语、WFCMS 英文术语、术语的英文解释（基于 WHO 标准）、类别编码、类别名称。共整理数据 16 189 条，其中 WHO 术语 3262 条，PMPH 术语 6848 条，WFCMS 术语 6079 条（图 1）。最终合并为 8975 条。

1.3　数据规范化处理

为了便于进行归类，我们基于《中华人民共和国国家标准 GB/T 13745-2009 学科分类与代码》[9]进行了更进一步分类（表 1）。部分分类下没有条目，是为了以后填充术语而暂时保留。文档编码为 UTF-8，针对生僻字或数据合并后可能出现的乱码，根据原始数据进行修正。每条数据的繁体中文和拼音均使用计算机自动生成，为了避免多音字错误，对一些常见多音字进行了拼音修正。对于某个

标准中没有的英文术语，保持该字段为空。只有 WHO 标准提供了术语的英文解释，对于 WHO 中没有的术语条目，术语的英文解释字段为空。所有方名、药名均为实体词首字母大写，所有简写均为大写字母，所有穴位名均为大写字母；其余英文术语均为小写。数据集采集和处理由吴佳泽完成（7 年编程经验，在 GitHub 拥有 10 项开源项目，荣获 Arctic Code Vault Contributor，https://github.com/BillEliot）。

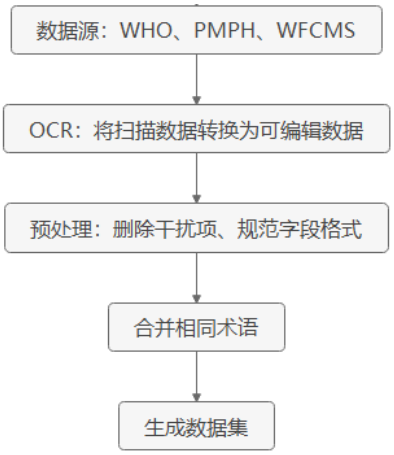

图 1　数据采集和处理方法流程

表 1　术语分类表

分类代码	学科分类名称	Category	数量
360.1011	中医基础理论(包括经络学等)	Basic theory	0
360.1011a	学科总论	General	98
360.1011b	阴阳五行	Yin yang and five phases	130
360.1011c	气血津液精神	Qi blood fluid essence spirit	79
360.1011d	藏象	Visceral manifestation	271
360.1011e	形体官窍	Body Constituents and Orifices of Sense Organ	198
360.1011f	经络	Meridian and collateral	64
360.1011g	病因	Cause of disease	240
360.1011h	病机	Mechanism of disease	575
360.1011i	治则治法	Rules and methods of treatment	765
360.1011j	治疗手段	Approaches	37
360.1014	中医诊断学	Traditional Chinese diagnostics	2
360.1014a	诊法总论	General of diagnostic method	11
360.1014b	望诊	Inspection	232
360.1014c	闻诊	Listening and smelling	57
360.1014d	问诊	Inquiry	300

分类代码	学科分类名称	Category	数量
360.1014e	切诊	Palpation	131
360.1014f	辨证总论	General of pattern identification	7
360.1014g	八纲辨证	Eight-principle pattern identification	97
360.1014h	病因辨证	Disease cause pattern identification	88
360.1014i	气血辨证	Qi-blood pattern identification	47
360.1014j	津液辨证	Fluid-humor pattern identification	30
360.1014k	脏腑辨证	Visceral pattern identification	218
360.1014l	六经辨证	Six-meridian pattern identification	61
360.1014m	卫气营血辨证	Defense, qi, nutrient and blood pattern identification	30
360.1014n	三焦辨证	Triple energizer pattern identification	9
360.1014o	其他辨证	Other pattern identification	13
360.1017	中医内科学	Chinese internal medicine	437
360.1021	中医外科学	Surgery of Chinese medicine	192
360.1024	中医骨伤科学	Chinese orthopedics and traumatology	249
360.1027	中医妇科学	Chinese gynecology	264
360.1031	中医儿科学	Chinese pediatrics	172
360.1034	中医眼科学	Chinese ophthalmology	150
360.1037	中医耳鼻喉科学	Chinese otorhinolaryngology	128
360.1041	中医口腔科学	Chinese stomatology	0
360.1044	中医老年病学	Chinese geriatrics	0
360.1047	针灸学(包括针刺镇痛与麻醉等)	Acupuncture and moxibustion	1
360.1047a	针法	Acupuncture	302
360.1047b	灸法	Moxibustion	64
360.1047c	拔罐	Cupping	25
360.1047d	穴位	Acupoint	606
360.1051	按摩推拿学	Tuina	29
360.1054	中医养生康复学(包括气功研究等)	Chinese health preservation and rehabilitation	98
360.1057	中医护理学	Chinese nursing	0
360.1061	中医食疗学	Diet therapy of Chinese medicine	0

分类代码	学科分类名称	Category	数量
360.1064	方剂学	Formula study	0
360.1064a	方剂总论	General of formula study	204
360.1064b	方剂名称	Formula name	580
360.1067	中医文献学	Chinese medical literature	153
360.1099	中医学其他学科	Other subjects of Chinese medicine	0
360.30	中西医结合医学	Integration of Chinese and Western medicine	0
360.40	中药学	Chinese pharmacy	2
360.40a	中药总论	General of Chinese pharmacy	295
360.40b	中药名称	Herbal names	1234

2 数据样本描述

2.1 数据结构

本数据集包含 1 张数据表。表中有 10 个字段，包括 ID、中文简体、中文繁体、拼音、WHO 英文术语、PMPH 英文术语、WFCMS 英文术语、术语的英文解释、类别编码、类别名称。每个类别的数据量如表 1。

2.2 数据样本展示

以中医术语“关格”为例，表 2 全面展示了该术语的中英文术语名称和英文解释。归类以类别编码和类别名称表示，可根据表 1 归类对应。

表 2　中医药术语中英对照数据集样本展示

序号	说明	数据示例
1	编号	67
2	术语名称（简体）	关格
3	术语名称（繁体）	關格
4	拼音	guān gé
5	WHO 术语	block and repulsion (disease)
6	PMPH 术语	anuria and vomiting
7	WFCMS 术语	anuria and vomiting
8	英文解释	a diseased state characterized by urinary stoppage and vomiting
9	类别编码	424.1017
10	类别名称	中医内科学

3 数据质量控制和评估

在通过 Python 完成数据合并后，我们依靠人工核对的方式对数据进行修正。由 2 人首先对数据对应性问题进行核查，保证无串行、错位等现象；然后对照源数据对数据转化中出现的乱码分别进行修复；重点核对生僻字和多音字条目的拼音。对于源数据中本身就是乱码，无法进行核实的，暂时保留，待以后通过其他途径核查条目确认后再进行修改。对于名称不同，但意思相同的术语，暂不合并，全部视为不同记录，予以保留。同时，以 Vue.js + Django 为基础框架搭建了在线检索网站（https:// medai.vip）。在网站中检索术语时，若使用者发现错误的条目，可以直接点报错（图 2），我们在系统后台定期进行修正。质控人员为梁昊（本科毕业于湖南中医药大学医学英语专业，从事中医英译工作 10 年）和周小青（曾任世界中医药学会联合会翻译专业委员会副会长，长期从事中医英译工作）。

图 2　中医术语中英对照查询系统术语报错演示

4 数据使用方法和建议

本数据集以 csv 文件为存储格式，使用者可以使用主流的数据管理及统计软件来对数据进行修改和查看，尤其方便利用 Python 和 R 语言对术语进行文本分析和处理。同时，基于本数据集搭建了术语检索系统（https://www.medai.vip），可以在该网站上随时检索术语。任何组织和个人可以以非商业目的使用本数据集，如搭建自己的术语库或术语检索系统。

5 数据价值

国内目前未见相似中英对照中医术语数据集。随着中医在全球的发展，国家对发展中医药的支持，国外对中医的了解需求日益增加。但由于种种原因的限制，不能及时查阅到中医术语对应的英文，导致国内外中医爱好者、学习者在学习交流的过程中，存在交流障碍，限制了中医对外发展及中医的对外交流。本数据集的公开，方便了中医从业者查询术语，促进了中医术语的规范化应用，有利于学术交流和中医的继承发扬。同时，标准化的术语也方便了中医药信息化建设，尤其是在 HIS、电子病历系统、医学数据分析系统中，让数据更加整洁，避免产生垃圾数据，减少不必要的数据清洗工作。与此同时，中医药领域开源的数据集稀少，不利于科学研究的开展和共享，本数据集的发布也是中医药开源模式的一次尝试，希望同行能够分享更多数据集，促进中医药的开放与发展。

数据作者分工职责

梁昊（1986—），男，河北保定人，博士，讲师、主治医师，研究方向为医学人工智能。主要承担工作：数据源搜集、数据集设计、论文写作。

吴佳泽（1999—），男，河北保定人，本科生在读，研究方向为医学数据挖掘。主要承担工作：数据清洗、合并，术语系统开发。

段伦慧（2000—），女，湖南常德人，本科生在读，研究方向为医学数据挖掘。主要承担工作：数据核查，论文写作。

彭清华（1965—），男，湖南望城人，博士，教授，研究方向为数字中医药。主要承担工作：组织数据集构建，修改论文。

胡志希（1962—），男，湖南娄底人，博士，教授，研究方向为中医药标准化。主要承担工作：修改论文。

周小青（1957—），男，湖南浏阳人，博士，教授，研究方向为数字中医药。主要承担工作：修改论文。

参考文献

[1] 贾静, 赵海磊. 中医术语英译标准的研究现状[J]. 临床医药文献电子杂志, 2016, 3(54): 10859–10860. DOI:10.16281/j.cnki.jocml.2016.54.138.

[2] 付甜甜, 都立澜, 刘艾娟. 基于 WHO 版和世中联版两大国际标准的中医病机术语英译对比研究[J]. 中国中医基础医学杂志, 2016, 22(2): 252–254.

[3] 李珊珊. 浅析世中联与 WHO 中医名词术语英译国际标准[J]. 2016(4): 82–83. DOI:10.3969/j.issn.1009-5349.2016.04.038.

[4] 徐丽, 张喆, 闵玲, 等. 中医术语英译标准的回顾与前景[J]. 西部中医药, 2021, 34(3): 158–162. DOI:10.12174/j.issn.2096-9600.2021.03.40.

[5] 董燕,朱玲,于彤,崔蒙,李海燕.中医临床术语研究现状与系统构建方法探讨[J].国际中医中药杂志,2014,36(11):965–968.

[6] 许吉, 施毅, 袁敏, 等. 中医术语国家标准比较研究[J]. 时珍国医国药, 2015, 26(9): 2294–2295. DOI:10.3969/j.issn.1008-0805.2015.09.096.

[7] World Health Organization. WHO international standard terminologies on traditional medicine in the western pacific region[M]. Geneva: WHO, 2007.

[8] Zhenji. International standard Chinese-English basic nomenclature of Chinese medicine[M]. Beijing: Peoples Medical Publishing House, 2008.

[9] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. 学科分类与代码: GB/T 13745—2009[S]. 北京: 中国标准出版社, 2009.

论文引用格式

梁昊, 吴佳泽, 段伦慧, 等. 中英对照中医药术语数据集[J/OL]. 中国科学数据, 2021, 6(3). (2021-08-25). DOI: 10.11922/csdata.2021.0009.zh.

数据引用格式

梁昊, 吴佳泽, 段伦慧, 等. 中英对照中医药术语数据集[DB/OL]. Science Data Bank, 2021. (2021-04-22). DOI: 10.11922/sciencedb.j00001.00213.

A dataset of Chinese-English terminology of Traditional Chinese Medicine

LIANG Hao[1,3], WU Jiaze[1], DUAN Lunhui[1], PENG Qinghua[2,3,4,5], HU Zhixi[1,3,6], ZHOU Xiaoqing[7]

1. School of Chinese medicine, Hunan University of Chinese Medicine, Changsha 410208, P.R. China
2. School of Integrated Chinese and Western medicine, Hunan University of Chinese Medicine, Changsha 410208, P.R. China
3. Diagnostic Information Branch of Chinese Medicine Information Society, Beijing 100700, P.R. China
4. TCM Diagnostic Committee of the World Federation of Chinese Medicine Societies, Beijing 100020, P.R. China
5. Digital Chinese Medicine Branch of International Society of Digital Medicine, Changsha 410208, P.R. China
6. Standardization Technology Committee of Chinese Association of Integrative Medicine, Beijing 100700, P.R. China
7. Translation Committee of World Federation of Chinese Medicine Societies, Beijing 100020, P.R. China

*Email: pqh410007@126.com

Abstract: The dataset is based on an integration of the *English Terminology of Traditional Chinese Medicine (Internal Draft)* compiled by the People's Health Publishing House (PMPH), the *WHO International Standard Terminologies on Traditional Medicine in the Western Pacific Region* formulated by the World Health Organization (WHO) and the *International Standard Chinese-English Basic Nomenclature of Chinese Medicine* produced by the World Federation of Chinese Medicine Associations (WFCMS). It is aimed to promote the standardization of Traditional Chinese Medicine (TCM) terms and international communication of TCM. We adopted Python pandas package and OCR technology to collect and sort 16,189 items, which were merged into 8,975 items, 56 categories. The dataset can promote the standardization of TCM terms, facilitate academic communication, inheritance and development of TCM, and contribute to the informatization construction of TCM.

Keywords: Traditional Chinese Medicine; terminology; Chinese-English

Dataset Profile

Title	A dataset of Chinese-English terminology of Traditional Chinese Medicine
Data corresponding author	PENG Qinghua (pqh410007@126.com)
Data authors	LIANG Hao, WU Jiaze, DUAN Lunhui, PENG Qinghua, HU Zhixi, ZHOU Xiaoqing

Time range	2007–2020
Geographical scope	Worldwide
Data volume	1.45 MB
Data format	*.csv
Data service system	<http://www.dx.doi.org/10.11922/sciencedb.j00001.00213>
Source of funding	Teaching Reform Research Project of Hunan University of Chinese Medicine (2020-JG006); Science and Technology Innovation Program of Hunan Province (2020RC2061).
Dataset composition	The data set consists of one data table of 8,975 items. There are ten fields in the table, including ID, simplified Chinese, traditional Chinese, Pinyin, WHO English terms, PMPH English terms, WFCMS English terms, English explanation, category code, category name.