• 医学信息组织与利用 •

多来源作者数据加工策略与实现——以西太平洋地区医学索引为例*

王 蕾 方 安 范云满 王 茜 王军辉 胡佳慧

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 分析西太平洋地区医学索引（WPRIM）已收录文献作者著录特点及国内外文献检索数据库著录标准，设计 WPRIM 作者数据著录标准，提出一种多来源作者数据加工策略，阐述关键步骤和方法，指出其有助于解决 WPRIM 作者著录数据存在的问题，实现作者数据规范化。

〔关键词〕 一带一路；西太平洋地区医学索引；数据加工；内容整合；质量控制

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2019.02.015

Multi－source Author Data Processing Strategy and Implementation: taking the Western Pacific Region Index Medicus as an example

WANG Lei, FANG An, FAN Yunman, WANG Qian, WANG Junhui, HU Jiahui, Institute of Medical Information/Medical Library, CAMS & PUMC, Beijing 100020, China

〔**Abstract**〕 The paper analyzes the features of the documentation of the author whose paper has been included in the Western Pacific Region Index Medicus (WPRIM) and the documentation standard of document retrieval database home and abroad. It designs the documentation standard of WPRIM author data and brings forward a strategy of multi－source author data processing. It also elaborates the key steps and methods, points out that such a strategy is conducive to solve the problems existing in the documentation of WPRIM author data and therefore achieve standardization of author data.

〔**Keywords**〕 The Belt and Road; Western Pacific Region Index Medicus (WPRIM); data processing; content integration; quality control

1 引言

在世界卫生组织（World Health Organization, WHO）和全球卫生图书馆（Global Health Library, GHL）项目支持下，中国医学科学院医学信息研究所开发并建设了西太平洋地区医学索引平台（Western Pacific Region Index Medicus, WPRIM）。目前已设计并提出西太平洋地区医学索引元数据方案[1]，实现部分“一带一路”沿线国家重要医学期

〔修回日期〕 2018－09－20

〔作者简介〕 王蕾，硕士，助理研究员；通讯作者：方安，副研究馆员。

〔基金项目〕 中国医学科学院医学与健康科技创新工程服务“一带一路”战略先导科研专项“卫生信息服务研究”（项目编号：2017－I2M－B&R－10）。

刊汇聚。截至 2017 年底 WPRIM 收录韩国、日本、中国及“一带一路”沿线的西太平洋地区国家医学领域期刊642 种，涵盖英文、中文、韩文、蒙古语等多语种文章信息，支持 WHO 西太平洋区域成员国出版、医学领域且具有英文题录的期刊文献资源集成，面向全球用户提供便捷的互联网访问，确保本地区医疗和卫生研究的全球可及性[2]。近期 GHL 项目各成员单位着手完善文献数据资源，改进现有索引系统数据的不足。WPRIM 作为 GHL 项目数据的重要来源，存在各国语言特点多样[3]、数据来源多样[4]、各国提交成果质量参差不齐、历史遗留情况复杂等问题，亟待通过多种处理策略解决现存问题。

2 作者数据著录特点

2.1 概述

WPRIM 作者数据来自 PubMed、J - stage、KoreaMed 等文献数据库或者由马来西亚、越南、老挝等国家的志愿者手动提交。受本国语言、数据库著录标准等因素影响，著录情况复杂。从各国语言特点分析，西方语言国家、东方印欧语系国家（如印度、孟加拉、伊朗等）、南岛语系部分国家（如印尼、马来西亚、菲律宾）等个人姓名排序一般为倒序[5]；汉藏语系国家（如中国）、南岛语系部分国家（如印尼、马来西亚、菲律宾）华人、日本、韩国等个人姓名排序一般为顺序[6]。从数据著录特点分析，作者不仅存在语言特点本身造成的数据著录问题，还存在同一作者姓名表述形式不同、大小写不规范、作者间分隔符不统一、包含噪音数据等问题。此外多来源的文献数据在数据收割过程中会存在内容缺失、解析不正确、作者姓名顺序错误的问题，也存在普通作者、机构和团体作者混淆的情况。部分数据存在同一国家志愿者反复提交，产生较多重复数据的问题。由于上述多种原因，未经质量控制的作者数据存在较多问题。

2.2 同一作者姓名表述形式不同

同一作者姓名表述形式存在著录顺序不一致、姓氏与名字之间分隔符不同、全拼中双名中间的连字符不同等问题。不同国别来源期刊的著录标准不同，故同一作者姓名存在著录顺序不一致的情况。一部分数据存在姓氏与名字的分隔符不一致，甚至存在姓氏与名字未分隔的情况，见表 1。同一作者姓名也存在全拼和简写两种形式。如作者“王承书”存在全拼“Wang Chengshu”与简写“Wang CS”两种著录形式。同一全拼作者还存在双名中间的连字符不一致的情况，部分采用横线、空格作为连接符，也有数据没有使用横线作为连接符，如“Wang Cheng - Shu”、“Wang Cheng Shu”、“Wang Chengshu”。

表 1 著录不规范数据样例

错误原因	样例
姓前名后	2 zhang jin - wen1 wang di
姓后名前	ai xia wang
姓与名之间用逗号分隔	ahn, hwang ran
姓与名之间用空格分隔	ai xia wang
姓氏与名字未分隔	zhupingzeng

2.3 英文著录大小写不规范

常见 WPRIM 作者数据采用每个单词首字母大写的形式，如“Chong - xing Zhou”。作者数据还存在姓氏全部大写、全部字母大写、全部字母小写的情况，如“Wenzhi DU”、“QIN MENG”、“chen ximing”。

2.4 多作者间分隔符不统一

一般情况下 WPRIM 多个作者之间采用分号进行分隔，如“CHEN Yan; ZOU Tian - ning”。部分数据使用空格、数字来区分不同作者，如“Ye Ling Qian Guan - Xiang Ge Sheng - Fang”。

2.5 噪音数据

主要由非法字符、非作者信息组成。非法字符如“\”、“.”、“No Authors Listed”、“Et Al.”、“No author”、“Checking”、“Reviewing”等。非作者信息常见的有团体作者（如 Extracurricular Re-

search Team、Group）、机构或地址（如 Suzhou Medical College、Shangqiu Central Hospital、100061、Zhengzhou University）、作者头衔（如 Director、Tutor、Ph D、MD、Lord）、邮箱、通信作者描述（如 Correspondence：Xu Guoming）等。

3 作者著录标准研究与设计

3.1 概述

全球医学索引分为地区索引、Medline 以及 SciELO3 大部分。WPRIM 作为地区索引的主要组成部分，其作者著录标准重点参考 Medline、SciELO 数据库的元数据项设置与著录规则，对标国内外重要文献检索数据库，提出 WPRIM 作者数据著录标准。

3.2 国内外重要数据库著录特点

国内外数据库之间的元数据标准、数据著录特点具有一定差异，见表2。作者分类方面，国际标准认为作者一般分为个人和团体作者两类[7]。国外数据库的个人作者元数据通常由一组作者信息组成，包含姓氏、名字、序号、简写、全称等内容，多个作者之间采用多条记录进行表示。部分国内数据库的个人作者元数据项设置作者一项，不划分姓氏、名字、简写和全称，多个作者之间使用分号进行分隔。作者名著录顺序方面，作者姓氏与名字前后顺序不固定。母语为英语国家的期刊，作者姓名一般采用姓氏在后、名字在前的著录规则。中国期刊的西文文献，作者著录一般符合国标 GB7713－87[8] 要求，一般采用姓氏在前、名字在后的著录规则。

表2 国内外数据库作者著录特点对比

数据库	组成内容	样例
Medline[9-11]	由 Author、author identifier、full author 组成。Full author 中包含 last name、fore name 等内容 FAU 由姓氏＋逗号＋名字组成	FAU：Anderson, Sarah－Jane AU：Anderson SJ AUID－ ORCID：http：//orcid. org/0000－0002－5945－2285
SciELO[12]	分为作者（author）和团体作者（corporate author）。单个作者由姓氏（surname）和名字（firstname）组成	［fname］A. C. ［/fname］ ［surname］Almeida ［/surname］
Web of science[13]	由 display_ name、full_ name、wos_ standard、first_ name、last_ name 组成。 （其中 display_ name 名字是由 last_ name＋“,”first_ name 的形式组成）	display_ name：Adu－Gyamfi, Y. O. full_ name：Adu－Gyamfi, Y. O. wos_ standard：Adu－Gyamfi, YO first_ name：Y. O. last_ name：Adu－Gyamfi
NSTL[14]	组成（3 部分）：Author sequence、Author name、Author name alternative 构成：姓氏＋名	中文作者：姓在前，名在后 西文作者：能够判断出姓和名时，姓在前、名在后，姓和名之间以逗号分隔；无法判断姓和名时，按原文著录顺序姓和名首字母大写，其余字母小写
SinoMed[15]	构成：姓＋“空格”＋名。多个作者之间用分号空格进行分隔	Yu Jing; Zhang Ju; Xing Hong; Zhao Xiao

3.3 WPRIM 作者数据著录标准

在上述调研分析的基础上 WPRIM 制定规范化的作者著录标准，见表3。元数据设计上，由于 WPRIM 作者以中国、日本、韩国文献数据为主，作者名一般由姓氏、名字两部分组成。巴布亚新几内亚、斐济等国家的文献内容，作者名一般由姓氏、中间名和名字 3 部分组成。故 WPRIM 作者数据全名包含姓氏、中间名、名字 3 部分。构成顺序上 WPRIM 主要面向西太平洋国家的全部用户进行服务，故借鉴 Medline 和 Web of Science 的作者著录顺序，规定其为名、中间名和姓氏。多作者分隔策略

上借鉴 SinoMed 数据库，采用分号进行分隔，便于数据清晰展示。拼写要求上借鉴 NSTL、Web of Science、J－stage、KoreaMed 多种数据库的拼写特点，规定作者名、中间名按首字母大写、其他字母小写规则著录，并要求姓氏按全部字母大写规则著录。

表 3 WPRIM 数据著录标准

著录项	著录标准	必选项
全名	firstname + 空格 + middlename + 空格 + lastname	是
名	首字母大写，其他字母小写。	是
中间名	首字母大写，其他字母小写。	是
姓	全部大写	是
团体作者	每个单词首字母大写	否

4 多来源内容整合策略及实现

4.1 加工策略

作者数据规范策略实现技术路线，见图 1，分为数据检查、数据拆分、二次检查、数据修正和数据重构 5 个步骤。WPRIM 不同来源的文献数据在 5 个步骤中根据来源数据特点进行不同的加工处理。

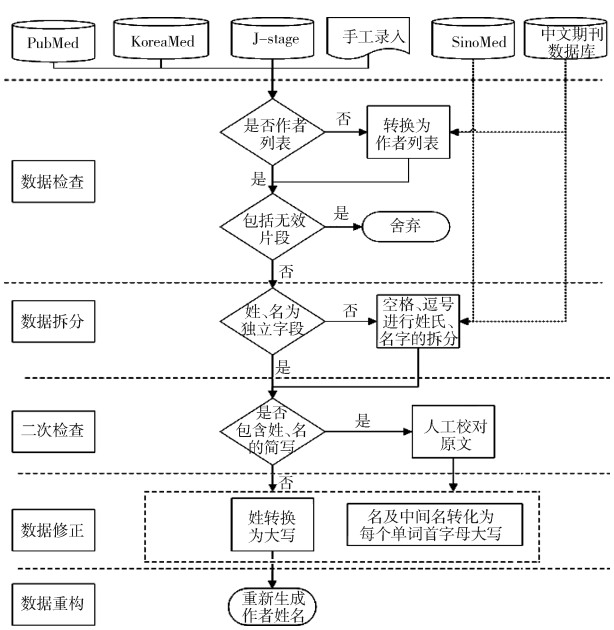

图 1 多来源数据加工策略

4.2 关键步骤和方法

4.2.1 数据检查 是对各种来源中的作者字段进行检查，即检查是否是作者列表和是否包含无效片段。是否是作者列表根据数据来源判断作者字段是否由多个作者字段形成的作者列表。根据前期调查，PubMed、KoreaMed、J－stage、手工录入的数据是按照作者列表的形式提交的，SinoMed 和中文期刊数据库文章是一个字段存放多个作者，多个作者之间以分号或其他分隔符进行分割。检查作者列表对 SinoMed 和中文期刊数据库的数据按照分隔符分割成作者列表。是否包含无效片段检查作者列表中的数据是否包含噪音数据。针对噪音数据，先通过团体作者和一般作者特征词进行筛选与判断，若包含则提取团体作者信息、修正个人作者信息。噪音数据不包含团体作者特征词时，经人工审核，将无效数据舍弃并反馈给数据提供方。

4.2.2 数据拆分 是利用界定条件与界定方法确定文献中作者姓、名的著录顺序，依据该顺序并结合姓和名之间的分隔符号进行数据拆分，实现每个作者的名（First Name）和姓（Last Name）的分离。（1）界定条件。依据 WPRIM 作者著录特点总结与提炼后形成的单一作者著录顺序判断条件。假设 X 与 Y 表示连续、无空格、无下划线的连续英文字符串，常见作者著录类型、附加判断条件、样例、界定结果，见表 4。通常利用条件 1 至 7 就可以界定作者姓名的著录顺序。中国、韩国等国家存在作者复姓的情况，故利用条件 8 至 11 进行姓名著录顺序的界定。中国作者数据利用除“n、g”以外的同一个辅音字母两次以上的方法界定姓和名的著录顺序有较好的界定效果。其他国家作者数据则通过常见复姓语料进行分析与处理，见表 5。当作者著录特点满足多个界定条件时，多组界定条件组合进行著录顺序的判定，形成多个界定结果。若多个界定结果一致，则认为界定条件的判断结果准确；若不一致，则认为该作者著录顺序界定结果不宜作为界定方法中的判断依据，界定结果判断流程，见图 2。（2）界定方法。作者著录顺序界定方法是优先以期、篇顺序进行自动判断，并辅以复杂数据的人工审核，确定某一篇文献的作者著录顺序。以期刊的一期数据为期界定单位，根据第一作者自动判断该期全部作者的著录顺序。出现“姓＋名”著录形式则界定本期全部文献作者著录顺序为“姓＋名”的表述形式；出现“名＋姓”著录形式则界定本期全

部文献作者著录顺序为“名 + 姓”的表述形式；若出现一期数据存在两种表述形式，则判断该期数据无法判断整期数据的著录顺序。以篇为界定单位，根据任意作者自动判断该篇文献全部作者的著录顺序。出现“姓 + 名”著录形式则界定本篇文献全部作者著录顺序为“姓 + 名”的表述形式；出现“名 + 姓”著录形式则界定本篇文献全部作者著录顺序为“名 + 姓”的表述形式；若出现一篇文献两种表述形式，则无法判断整篇数据的著录顺序。无法自动判断作者著录顺序的文章需要进行人工界定。

表 4 界定条件示例

序号	著录类型	附加条件	样例	界定结果
1	X X - X	无	Yang Zhao - hua	姓 + 名
2	X, X - X	无	Yu, Sheng - hui	姓 + 名
3	X - X X	无	Jian - min Wang	名 + 姓
4	X - X, X	无	Bai - shen, PAN	名 + 姓
5	X X X	无	ZHU Lu yun	姓 + 名
6	X Y 或 X, Y	X > 7 且 Y < 7		名 + 姓
7	X Y 或 X, Y	X < 7 且 Y > 7		姓 + 名
8	X Y 或 X, Y	X 中出现除“n、g”以外的同一个辅音字母两次以上	Zhanzhu Huang Zhanzhu, Huang	名 + 姓
9	X Y 或 X, Y	Y 中出现除“n、g”以外的同一个辅音字母两次以上	Chen Zhenzhou Chen, Zhenzhou	姓 + 名
10	X Y 或 X, Y	X 中出现复姓表中的表述形式	Ouyang Jing Ouyang, Jing	姓 + 名
11	X Y 或 X, Y	Y 中出现复姓表中的表述形式	Jing Ouyang Jing, Ouyang	名 + 姓

表 5 常见复姓

常见复姓	常见复姓
Ouyang	Namgung
Linghu	Guyang
Huangpu	Sanggwan
Shangguan	Seonu
Situ	Seomun
Zhuge	Sado
Sima	Dongmun
Yuwen	Gongson
Huyan	Hwangbo
Duanmu	Yeongho

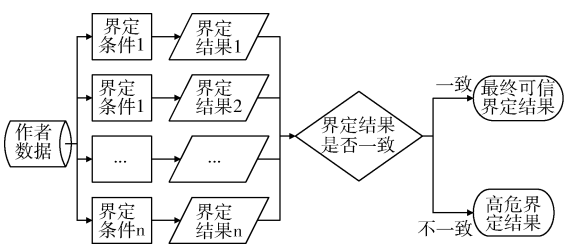

图 2 界定结果判断流程

4.2.3 二次检查 是检查经过拆分得到的姓、名是否正确、是否包含简写及无效信息。首先利用网络资源[16]、构建常见姓氏语料，见表 6。再对数据进行筛选，若名包含常见姓氏，则作为高危数据进行人工审核及干预。若姓包含常见名，也要进行人工审核及干预。是否包含简写信息以姓或名字段值过短、具有“.”符号或两个连续大写辅音字母作为一个词（如 JK）等条件，认定字段项包含简写。简写数据需人工核实原文，补充著录作者姓、名的全拼。是否包含无效信息通过无效信息语料（如逗号等）提取数据进行审核与修正。

表 6 常见姓氏

常见姓氏	常见姓氏
Li	Zhao
Kim	Qian
Lim	Sun
Zhang	Zhou
Park	Woo
Satō	Suzuki
Takahashi	Tanaka
Villanueva	Lin
Sy	Sarmiento
Nguyễn	Lê

4.2.4 数据修正 是对姓、名的著录样式进行规范化。针对中国、日本、韩国的作者将姓氏字母转换为大写，其他字母转换为小写。名及中间名转换为每个单词首字母大写、其他字母小写。数据修正样例，见表 7。

表 7 数据修正样例

作者名	姓修正前	姓修正后	名修正前	名修正后
Zhi – gang Fu	Fu	FU	Zhi – gang	Zhi – gang
Huan SONG	SONG	SONG	Huan	Huan
MINA PARK	Park	PARK	MINA	Mina

4.2.5 数据重构 主要是将修正结果构建成服务数据，并补充来源数据、修正结果、服务数据 3 者的对应关系。修正结果构建成服务数据是将修正后的姓和名结果进行重新组合，形成“名”+“空格”+“姓”或“名”+“空格”+“中间名”+“空格”+“姓”著录形式的服务数据。

5 结语

通过分析一带一路沿线国家作者表述方式及 WPRIM 收录期刊作者著录特点，结合国内外知名文献检索系统的作者字段项著录规则，提出 WPRIM 作者数据著录标准，实现期刊作者整合与规范加工方法。WPRIM 已完成 60 余万篇文献数据的作者数据规范，实现作者著录格式的统一，满足 GHL 对作者数据的质量要求。规范后的 WPRIM 数据已被其他文献检索平台（如 GOOGLE SCHOLAR[17]）收录。与此同时 WPRIM 作者数据质量控制方法面临数据质量控制的新挑战，亟待解决作者数据质量控制实时化、人工处理率高的主要问题，积累和扩展数据质量控制相关的语料资源，完善多种来源数据的处理机制，获得更好的作者数据质量控制效果。

参考文献

1 王军辉，钱庆，方安，等. 西太平洋地区医学索引元数据方案的设计与应用［J］. 医学信息学杂志，2011，32（4）：68 – 72.

2 西太平洋地区医学索引. 西太平洋地区医学索引系统介绍［2018 – 01 – 23］.［EB/OL］. http：//wprim. who-cc. org. cn/index. jsp.

3 丁波涛. “一带一路”沿线国家信息资源整合模式——基于国际组织和跨国企业经验的研究［J］. 情报杂志，2017，36（9）：160 – 164.

4 王军辉，钱庆，方安，等. WHO 西太平洋地区医学索引建设进展与问题［J］. 中华医学图书情报杂志，2014，23（2）：75 – 79.

5 李金花. 外国个人责任者在 CNMARC 中的规范标目和著录［J］. 河南图书馆学刊，2009，29（4）：114 – 115.

6 管蔚华. 外国人姓名特点及其在 CNMARC 个人名称字段的规范标目［J］. 大学图书馆学报，2001（4）：80 – 82.

7 ISDB（International Standard Bibliographic Description）［EB/OL］.［2018 – 05 – 10］ http：//www. ifla. org/isbd – rg.

8 GB 7713 – 87 科学技术报告、学位论文和学术论文的编写格式［M］. 北京：中国标准出版社，1987.

9 葛红梅，徐晶晶，董鹏，等. PubMed 数据库建设探析［J］. 数字图书馆论坛，2015（5）：64 – 68.

10 NIH. Glossary & Acronym List［EB/OL］.［2017 – 12 – 15］. http：//grants. nih. gov/grants/glossary. htm#c.

11 NIH. MEDLINE/PubMed Data Element（Field）Descriptions［EB/OL］.［2018 – 01 – 15］. https：//www. nlm. nih. gov/bsd/mms/medlineelements. html#au.

12 SciELO. SciELO DTD［EB/OL］.［2018 – 05 – 08］. http：//www. scielo. org/php/level. php? lang = en&component = 42&item = 4.

13 Web Of Science. Web Of Science DTD［EB/OL］.［2017 – 12 – 15］. http：//ipscience – help. thomsonreuters. com/wosWebServicesExpanded/WebservicesExpandedOverv-iew-Group/Introduction/sampleResponse. html.

14 NSTL. NSTL 文献资源加工规范［EB/OL］.［2018 – 01 – 14］. http：//spec. nstl. gov. cn/specification/index. php? title = NSTL 文献资源加工规范.

15 中国生物医学文献数据库. 中国生物医学文献数据库西文生物医学文献数据库介绍［EB/OL］.［2018 – 01 – 10］. http：//www. sinomed. ac. cn/en/.

16 Wiki. List of most common surnames in Asia［EB/OL］.［2018 – 01 – 10］. https：//en. wikipedia. org/wiki/List_of_ most_ common_ surnames_ in_ Asia.

17 Google Scholar. Inclusion Guidelines for Webmasters［EB/OL］.［2018 – 01 – 10］. https：//scholar. google. com/intl/en/scholar/inclusion. html#indexing.