

PubMed数据库建设探析*

葛红梅, 徐晶晶, 董鹏, 周琴, 任慧玲
(中国医学科学院医学信息研究所, 北京 100005)

摘要: PubMed数据库是世界公认的最有影响力的医学专业数据库。本文从数据库建设的角度, 分析PubMed数据库的收录标准、数据处理流程、基础数据整合、深度标引等, 以期对国内数据库建设有一定的借鉴意义。

关键词: PubMed; 数据库建设

中图分类号: G25

DOI: 10.3772/j.issn.1673-2286.2015.05.012

PubMed数据库是美国国家医学图书馆(National Library of Medicine, 简称NLM)所属的国家生物技术信息中心(National Center for Biotechnology Information, 简称NCBI)开发的世界上最权威、最具影响力的生物医学信息检索系统, 免费向全世界公众提供生物医学论文的题录检索服务, 并提供期刊或其他相关网络全文链接。该数据库支持智能检索、主题检索、临床查询等方式, 并提供期刊全文链接、相关论文推荐等功能, 平均每天约有300万次网站直接检索, 300万次通过程序调用数据^[1]。

1 概况

PubMed数据库最早可以追溯到1879年创刊的医学印刷本题录型检索刊物医学索引(Index Medicus, 简称IM)。IM是NLM历史上最重要的品牌产品之一, 是医学界最常用的一种医学文献检索工具书, 月刊, 年度出版医学累积索引年刊, 是了解世界医学动态的窗口。1971年NLM创办并推出Medline数据库, 数据处理方式与IM一致, 收录1946年以来的数据, 对所有题录数

据进行主题标引, 通过光盘或联机数据库提供交互式检索, 便于读者使用。PubMed于1996年创办并对外服务, 第二年宣布对全世界免费开放。与Medline相比, PubMed收录范围更广泛, 年代更早, 检索途径增多, 并增加相关论文推荐、评论等功能。PubMed数据库广泛应用后, IM因订户减少于2004年终止出版发行。

PubMed学科范围包括生物医学、健康、生命科学、行为科学、化学科学, 2000年以后生命科学延伸为生物学、环境科学、海洋生物学、植物与动物科学、生物物理学、生物化学等。到2014年底PubMed收录范围包括Medline期刊5642种; PMC收存期刊/手稿, 其中全开放期刊1617种、NIH资助期刊298种、选择性存缴期刊2903种; NCBI书架3120种图书。在文献论文方面, 截止到2014年底, PubMed收录论文数据2400余万条。其中包括Medline数据库论文约2100万, PMC数据库题录约330万, 其他数据是收录期刊中不属于Medline收录范围的论文(如不符合Medline收录的学科天体物理学等)、只被PubMed收录期刊的论文、提前出版的论文、1946前的论文、NCBI书架图书章节等^[3]。

* 本研究得到中央级公益性科研院所基本科研业务费专项课题“医科院图书馆特色数据库建设研究”(编号: 12R0114)和国家科技图书文献中心“NSTL联合目录元数据体系建设研究”(编号: 2014XM004)资助。

2 PubMed数据库建设

2.1 收录标准

为保证数据库的权威和数据质量, PubMed收录的期刊必须经过审核遴选。为保证收录期刊的质量, NLM于1988年成立了文献选择项目技术审核委员会(Literature Selection Technical Review Committee, 简称LSTRC), 专门从事审核期刊、评价其内容、增选新刊, 同时剔除不符合要求的已收录期刊。每年召开三次会议, 每次审核约140种期刊。

LSTRC选择确定收录期刊, 主要从以下几方面进行评议^[2]: ①期刊收录的范围和覆盖面, 是否属于生物医学领域; ②内容质量, 包括论文学术质量可靠性、创新性及其对本领域的贡献; ③编辑质量, 期刊录用稿件的标准, 外送同行专家审稿的程序是否明确详细等; ④出版质量, 指期刊版面设计、印刷质量、图表及其注释等; ⑤读者群, LSTRC把读者分为5类, 即医学研究人员、执业医师、教育人员、医学管理人员和医学生; ⑥论文类型, 如原始研究报告、含有分析和讨论的临床观察、有讨论的病例报告等; ⑦地域范围, 注重期刊的国际化程度, 多为世界性期刊、区域性期刊。

2.2 数据处理流程

PubMed数据库在整个数据处理流程中, 可以分成两个阶段: 基础数据整合和深度标引。基础数据整合阶段, 接收不同来源的属于收录范围内的论文, 按照一定标准存储, 形成PubMed基础数据库, 并对外进行服务, 数据状态标识为[PubMed-as supplied by publisher]、[PubMed-in process]。深度标引指PubMed对数据进行审核, 并邀请学科专家、图书情报专家对

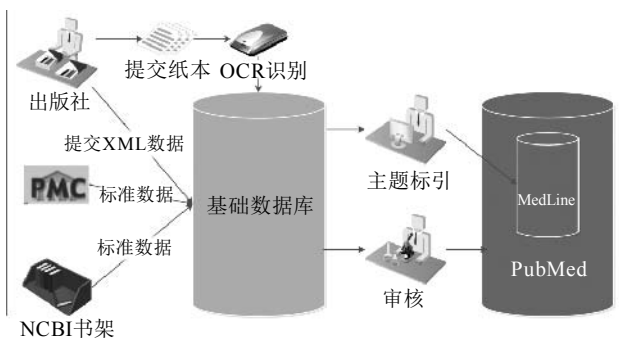


图1 PubMed数据处理流程图

Medline收录的论文进行主题词、特征词、物质等标引, 最终数据状态标识[PubMed-indexed for MEDLINE]、[PubMed], 用于扩大论文的检索途径, 提高论文检索准确率。整个处理流程如图1所示。

2.3 基础数据整合

2.3.1 基础数据来源

1997年之前, NLM一直采用手工键盘双人双录入方式建设Medline数据库, 将纸本文献转换为数字题录数据。1997年初, NLM引入对纸本文献扫描、光学字符识别(Optical character recognition, 简称OCR)加工方式。随着数字出版技术的发展, 从1998年起NLM接受出版商提供的标准化的可扩展标识语言(Extensible Markup Language, XML)数据。随后, PubMed数据库手工键盘录入方式产生的数据比例逐年缩小, 并在2005年不再采用; 出版商提交的数据比例逐年增加, 2013年产生的数据占当年数据总量的95%^[4]。

NLM鼓励出版商提交XML数据的方式, 在提高数据质量的同时, 缩短了数据加工的周期。目前在出版商提交正确格式的XML文件48小时后, PubMed通过审核即可提供检索服务。随着越来越多的期刊采用数字版提前出版模式, 并同时向PubMed提供数据, PubMed对外服务的数据更新非常及时, 甚至会早于印本期刊的发行, 这部分数据增加标识[Epub ahead of print]。当印本期刊数据正式发布后, 该数据将被正式数据替换。

2.3.2 基础数据提交

面对众多的PubMed数据库中收录文献的来源出版商, NLM制定了一整套数据提交指南和标准, 公布在其官网上供出版商下载使用。出版商工作指南的内容包括如何确定其期刊被PubMed收录、如何处理数据、数据如何更新等。通过详细的操作指南, NLM保证了各个数据来源的数据的一致性和数据质量的高标准。

出版商在确定其期刊被PubMed收录后, 便可以通过FTP方式提交文摘元数据给NLM。主要步骤如下^[5]: ①出版商向PubMed管理者(publisher@ncbi.nlm.nih.gov)申请FTP账号, 并提供PubMed数据库收录的期刊

列表。②出版商在获得批准后通过网站下载PubMed DTD (PubMed Document Type Definition, 文档类型定义)。PubMed DTD规定了PubMed数据库各元素的语法规则。为方便出版商理解和使用PubMed DTD, PubMed在网站上提供了XML (Extensible Markup Language, 可扩展标记语言) 格式的数据样例和可编辑的无内容的标准XML文件。出版商按照PubMed要求在本地将待提交的期刊论文元数据转换为PubMed XML文件。③出版商通过FTP工具或ftp-private.ncbi.nlm.nih.gov网站, 登录后提交XML格式的数据文件进行PubMed DTD验证。④接收PubMed收录反馈, 并接收期刊论文在PubMed数据库中唯一标识号 (PubMed Unique Identifiers), 简称PMID。

出版商提交数据注意事项如下: 必须整期论文完整提交, 不允许同期论文分批次提交; 期刊名称、ISSN、年、卷期信息完整必备, 便于论文的汇聚; 论文类型为研究论文、社论、病例报告、编辑邮件, 不能提交的论文类型为书评、广告、声明、勘误、软件等; 提交纯电子期刊时, 如无页码, 必须著录doi (Digital Object Unique Identifier, 数字对象唯一标识符) 或pii (Publisher Item Identifier, 出版物件标识符); 提交每篇论文的出版状态, 出版过程中的时间等。

如果出版商提交的数据尚未被PubMed数据接受处理, 即数据状态为[PubMed - as supplied by publisher]时, 出版商能启用replace标签对提交的数据进行替换修改。如果出版商提交了带删除标记的数据, 则该数据被PubMed接受后, PubMed会对原数据进行删除处理, 并提供每天更新的删除数据PMID列表, 供出版商随时查看数据处理结果。

当数据被PubMed接受后, 不论数据处于[PubMed - in process]、[PubMed - Indexed for MEDLINE]、[PubMed]、[PubMed - OLD MEDLINE]任何状态, 均需要将替换的数据提交给NLM员工进行处理。在PubMed数据库中, 原始数据和更替后数据均保留, 用版本元素作区分。

2.3.3 元数据规范

PubMed数据库的元数据可以分为四类^[6]: 第一类是期刊与卷期信息, 元数据包括出版商、期刊名称、ISSN、出版年、卷、期、出版时间; 第二类是论文信息, 元数据包括论文各类唯一标识符如DOI或PII等、论文题

名、文摘、关键词、起止页、语种、论文类型、出版时间、出版历史、非英文题名、非英文关键词、论文间关系、版权信息等; 第三类是论文作者信息, 元数据包括个人作者名称、作者的姓名全称、作者的姓、作者中间名字、作者后缀、作者地址、作者唯一标识符, 团体作者名称、团体作者地址、贡献者相关信息等; 第四类是PubMed数据库的专有信息, 如PMID、数据版本、PubMed数据库收录时间等。

PubMed数据库除对单篇论文信息详细著录外, 还记录论文的出版过程中的状态, 如数字版提前出版、印本出版; 论文处理的时间, 如论文投稿时间、修改时间、被接受时间、发表时间等。

PubMed数据库作者相关字段很详细全面。其中作者类型有个人作者、团体作者、贡献者; 作者名字的字段有姓、名、全称、头衔、后缀 (如“Jr”, “Sr”, “II”, “IV”) 等; 地址包括机构名称、机构所在地址、个人作者Email; 作者标识ID目前有ORCID (Connecting Research and Researchers ID, 开放研究者与贡献者身份)、ISNI (International Standard Name Identifier, 国际标准名称识别码)、VIAF (Virtual International Authority File, 虚拟国际规范文档) 几大国际作者规范库的ID号。

2.3.4 著录标准

PubMed元数据的设置及著录要求, 也随着数字出版的发展、论文形式的变化而不断完善^[7]。最典型的元数据是作者元数据。最初只设有作者名称、作者姓、作者名、作者地址四个元数据。2002年增加元数据作者全称和作者标识ID。作者标识ID有则必备, 信息来自出版商采用的ORCID、ISNI、VIAF等国际组织的作者规范ID号, 例如orcid 0000-0001-5027-4446; 2013年增加作者关键词、作者顺序等, 便于区分作者的贡献; 2014年增加所有作者的地址, 便于区分同名作者, 之前仅著录第一作者地址。

在20世纪50年代, 每篇论文的作者平均为1.7人, 而随着科研的深入开展, 科研合作项目越来越多, 导致代表科研成果的学术论文合作者人数不断增加, 到2010年, 每篇论文的作者平均数已超过5人^[8]。PubMed数据中对作者个数的著录标准也多次调整, 1983年以前多人合作者现象不多时, 是不限制的。后来随着合作者的论文数量和同论文合作者数量均不断增加, PubMed数据

库1984-1995年限定作者不超过10个, 1996-1999限定作者不超过25个, 其他作者用“et al.”代替。2000年以后, 随着对科研工作者个人科研评价的重视, 该数据库再次调整为对作者个数不限定。

2.4 深度标引

PubMed/Medline数据库能够在生物医学领域众多数据库中成为工作者查找文献的首选信息源, 除收录文献学术水平高外, 很大程度上依赖于其智能的检索、丰富的文献过滤、检索结果分面、相似度排序等功能。而这些功能的实现是建立在PubMed对文献深度标引规范的基础之上。

学术论文中, 针对同一事物不同作者有不同的表述习惯, 这些被称为自然语言的词汇多存在同义、多义现象, 同时有词汇量巨大、词间关系不明晰、具有模糊性和不确定性等问题。PubMed数据库采用NLM研制、逐年更新的受控医学主题词表(Medical Subject Headings, 简称Mesh)对文献进行标引, 用主题词、副主题词来标识论文讲述的内容, 使论文描述准确、词间关系丰富。

Mesh词表包括主要叙词、副主题词和款目词。主要叙词根据不同功能分为主题词、特征词、出版类型主题词和地理主题词。主要主题词指描述主题或内容特性的受控词汇, 2014年底总量达到27,149个。特征词是文献中涉及研究对象的具有区分功能的主题词, 如“动物”、“人类”; “男(雄)性”、“女(雌)性”; “儿童”、“成年人”等各种年龄组、年代组。这些主要叙词、特征词和出版类型主题词、地理主题词配合使用, 能更好的描述标注揭示文献主题, 最终利于用户的限定检索, 缩小检索范围。

3 PubMed数据库建设特色

3.1 制定严格收录标准, 保证高质量数据内容

PubMed数据库从历史沿革看, 除收录的学科范围有所扩展外, 其遵循的收录标准一直延续原IM的收录标准, 对期刊的学术内容、编辑质量、论文类型等均有明确的规定, 收录期刊增长不多, 确保PubMed收录文献的质量。PubMed并没有随论文数据提交的便利而

大量增扩期刊。医学期刊2000年4332种到2013年5649种, 年增长量约2%, 低于同期同行评议学术期刊年增长量3.5%^[9]。全世界医学领域的教研工作者如此青睐PubMed数据库, 与其检索结果均为优质论文有关。

3.2 借助数字出版, 提高自动化处理程度

PubMed积极与各出版商联系, 积极发挥数字出版优势, 直接采用出版商提供的数据, 降低OCR加工数据的比例。通过大量采用原始数据, 提高自动化处理水平, 既降低了数据处理的人工成本, 也提高了数据质量, 避免了识别、录入等产生的错误。

3.3 元数据规范和深度标引

PubMed一直在调整完善其元数据, 随着数字出版的发展以及使用者的反馈, 不断增加论文描述元素。NLM邀请若干领域专家、图情专家对收录的论文在机标的基础上审核其主题标引、特征词等。PubMed还开放其数据接口, 鼓励全世界人员对其数据进一步挖掘利用, 并将研究结果反馈到数据中。

3.4 高度开放共享

PubMed数据库元数据中设有多个与外部资源相联的元素, 如记录DOI、PII等唯一标识符, 将该论文与多个数据库中的全文进行了链接。通过记录PMC ID号、NCBI基因序号、化学物质号等信息, 与其他类型资源保持链接。在PubMed数据库建设过程中, 鼓励出版商参与到数据库建设中, 并及时反馈出版商该论文PMID, 既保证了数据处理高效率, 缩短加工周期, 又使更多的用户通过出版商网站链接使用PubMed数据库。PubMed同时提供FTP下载、API接口等, 供合作方便捷地利用该数据库。

4 结语

作为最受医教研工作者青睐的专业数据库, PubMed数据库建立了一套完整的数据收录、数据处理标准和步骤。本文从数据库建设的角度, 分析PubMed数据库的收录原则、数据处理流程, 介绍数据来源、数据提交、元数据规范、深度标引等, 希望对国内数据库

建设有一定借鉴意义。

参考文献

- [1] How Many PubMed Searches are Performed?[EB/OL]. [2015-03-15].
http://www.nlm.nih.gov/services/pubmed_searches.html.
- [2] MEDLINE Journal Selection[EB/OL].[2015-03-15]. <http://www.nlm.nih.gov/pubs/factsheets/jsel.html>.
- [3] MEDLINE, PubMed, and PMC (PubMed Central): How are they different?[EB/OL]. [2015-03-15]. http://www.nlm.nih.gov/pubs/factsheets/dif_med_pub.html.
- [4] MEDLINE Data Entry by Fiscal Year[EB/OL].[2015-03-15]. http://www.nlm.nih.gov/bsd/stats/data_entry.html.
- [5] XML Help for PubMed Data Providers[EB/OL].[2015-03-15]. <http://www.ncbi.nlm.nih.gov/books/NBK3828/>.
- [6] PubMed Help[EB/OL].[2015-03-15] <http://www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T44>.
- [7] MEDLINE/PubMed Data Element (Field) Descriptions[EB/OL]. [2015-03-15].<http://www.nlm.nih.gov/bsd/mms/medlineelements.html>.
- [8] Number of Authors per MEDLINE/PubMed Citation[EB/OL]. [2015-03-15]. <http://www.nlm.nih.gov/bsd/authors1.html>.
- [9] Mark Ware, Michael Mabe. The stm report: An overview of scientific and scholarly journal publishing (Third Edition)[R]. International Association of Scientific, Technical and Medical Publishers, 2012:22.

作者简介

葛红梅, 女, 1979年生, 研究方向: 图书馆信息组织、图书馆元数据建设。

徐晶晶, 女, 1989年生, 中国医学科学院医学信息研究所助理馆员。

董鹏, 男, 1986年生, 中国医学科学院医学信息研究所助理馆员。

周琴, 女, 1983年生, 中国医学科学院医学信息研究所馆员。

任慧玲, 女, 1971年生, 中国医学科学院医学信息研究所研究馆员, 研究方向: 信息资源建设, 通讯作者, E-mail: ren.huiling@imicams.ac.cn。

An Analysis of Database Construction on PubMed

GE HongMei, XU JingJing, DONG Peng, ZHOU Qin, REN HuiLing

(Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100005, China)

Abstract: PubMed database is recognized as the world's most influential medical database. This paper focuses on the analysis of NLM's PubMed database from the perspective of database construction, concerning its inclusion criteria, data processing procedures, data sources, data submission, metadata standards, and deep indexing, with the purpose of using its experience for the reference of domestic database construction.

Keywords: PubMed; Database Construction

(收稿日期: 2015-04-03; 编辑: 雷雪)