多源期刊元数据汇聚研究*
——以世界卫生组织西太平洋地区医学索引为例

王蕾　方安　杨雨生　范云满　王茜
（中国医学科学院医学信息研究所，北京 100020）

摘要：基于世界卫生组织西太平洋地区医学索引（WPRIM）开发建设过程中数据资源的现状和期刊元数据汇聚面临的问题，从期刊文献数据源遴选、元数据标签映射、内容著录规范化、非结构化数据转换4个维度设计多源数据汇聚方案。结果表明，面向WPRIM的多源期刊元数据汇聚框架能够较好地解决多源期刊数据汇聚问题，可为类似场景提供方案参考。

关键词：数据汇聚；多源数据；西太平洋地区医学索引；多源异构
中图分类号：G354.49；G255.2　　DOI：10.3772/j.issn.1673-2286.2021.01.007
引文格式：王蕾，方安，杨雨生，等. 多源期刊元数据汇聚研究——以世界卫生组织西太平洋地区医学索引为例[J]. 数字图书馆论坛，2021 (1)：47-53.

如何快速整合分散于各国尤其是发展中国家的专业领域文献[1]，消除各国资源之间的信息孤岛，建立开放服务的资源共享机制[2]，是走向数据融合和知识融合的关键问题，也是当前世界卫生组织西太平洋地区医学索引（WHO Western Pacific Region Index Medicus，WPRIM）面临的主要挑战。本文以中国医学科学院医学信息研究所开发并建设的WPRIM[3]为例，探索各国数据来源复杂、元数据标准不同、数据著录水平参差不齐、结构化程度不一致背景下多国医学领域文献的汇聚策略与方法，总结多源期刊汇聚过程中的优势与不足，以期提供高质量的索引服务，为相关机构开展多国、多源专业文献数据汇聚提供方案参考。

1　WPRIM面临的现状与挑战

1.1　现状

截至2020年底，WPRIM收录了包括中国、日本、韩国、蒙古、菲律宾、马来西亚、新加坡、老挝、越南、斐济、文莱、巴布亚新几内亚等国家出版的西太平洋地区科技期刊论文、灰色文献等生物医学领域文献资源，其中生物医学期刊665种。汇聚对象来源方面，WPRIM收录期刊的数据来源包括期刊编辑部和第三方数据平台两类。来自编辑部的数据一般通过人工录入或上传可扩展标记语言（XML）文件的方式进行数据汇交；来自第三方数据平台（包括韩国KoreaMed、日本科学技术信息集成系统（J-STAGE）、美国PubMed等）的数据由WPRIM平台统一管理采集。汇聚对象元数据标准方面，编辑部提供的结构化数据主要采用JATS[4-7]作为元数据标准；第三方数据平台提供的数据采用KoreaMed标签集、J-STAGE标签集以及JATS等元数据标准。汇聚对象结构化程度方面，WPRIM元数据对象包括结构化[8-10]期刊数据、非结构化期刊数据与半结构化期刊数据。结构化期刊数据一般保存在XML文件中并进行数据传输，如*Acta Pharmaceutica Sinica*期刊汇交XML格式的数据文件至WPRIM数据管理平台。非结构化数据通过TXT或HTML格式的文本文件进行数据交换，如部分编辑部提供方正书版导出的文本形式的数据进

*本研究得到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”（编号：2017-I2M-B&R-10）和中国医学科学院医学与健康科技创新工程“医学科技创新评价与卫生服务体系研究”（编号：2016-I2M-3-018）资助。

行数据汇交。半结构化数据是介于结构化数据与非结构化数据之间的一种数据对象，主要存在于XML文件或接口采集的成果中。

1.2 挑战

1.2.1 同一期刊存在多个数据来源

部分WPRIM收录的期刊存在同一期刊数据来源多样的情况，即同一本期刊被多个数据库收录或存在编辑部和第三方检索平台都能提供题录数据的情况。如*Singapore Medical Journal*（ISSN：0037-5675）同时被PubMed、Web of Science、Embase等数据库收录，同时该刊物的编辑部也能够提供题录数据。如果同时获取不同来源的期刊数据，就会造成数据重复的问题，增加数据管理的复杂度。

1.2.2 不同数据源的元数据标签不一致

WPRIM来源数据有多种元数据标准，存在作者、语种、时间等元数据与WPRIM元数据标准命名或含义不一致的情况。元数据项命名包括同名和不同名两种情况，如JATS中的名字标签（NAME）的姓名类型（NAME-STYLE）为西文的姓标签（SURNAME），与WPRIM的姓标签（LASTNAME）不同名。元数据标签含义包括同义、近义、不同义3种情况，如J-STAGE标签集中作者（AUTHORS）与WPRIM元数据中姓（LASTNAME）、名（FIRSTNAME）标签名称近义。

1.2.3 不同数据源著录标准不同

WPRIM收录期刊的各个数据源著录标准不一致，作者、刊名、语种、时间、卷期元数据项存在全称与简称、语种等著录形式的差异。以刊名为例，*Journal of Breast Disease*在KoreaMed数据源中著录为简称J Breast Dis，而非期刊全称。以语种为例，*Annals of the Academy of Medicine, Singapore*期刊文献的语种在PubMed数据源中著录为eng，而WPRIM元数据著录标准要求著录为English。如果只开展元数据标签项的融合，则会出现内容不一致的情况，导致数据质量下降。

1.2.4 非结构化数据人工加工成本高

为解决WPRIM数据缺失的问题，需要对非结构化历史数据进行补充。由各国数据管理人员、编辑部编辑等通过逐条录入的形式向WPRIM系统汇交非结构化数据。这种数据汇交模式不仅增加了数据管理人员和编辑的工作量，还会出现更新速率慢、易出错的情况，不符合数据管理未来可持续发展的趋势。

2 WPRIM的元数据汇聚策略

2.1 多源异构元数据汇聚的相关实践

为解决数据来源多、形式多（如同型/质异源、异质异构和多种语言[11]；结构化、非结构化和半结构化[12-13]）、内容杂（如系统异构、语法异构、结构异构和语义异构[14]）的现状，学术界从质量评估、元数据映射、领域本体等角度进行多源数据汇聚路径的探索。林鑫等[15]、周艳会等[16]、Bruce等[17]从元数据、数字字典、用户要求、数据应用等角度进行数据质量评估，设置数据质量控制规范规则，提升集成对象的数据质量。Moghaddasi等[18]、于倩倩等[19]等通过元数据标签映射等方式，从内容标准化和元数据映射两个维度实现多源数据汇聚。刘盼雨等[20]依据数据流向通过多源异构数据转换、清洗、元数据管理等手段构建涵盖“生产-存储-计算-应用”的多源异构数据服务平台。侯鑫鑫等[21]提出数据获取、数据整合、关联关系建立、入库及调用的异构大数据整合方案技术路线。曲建升等[14]、崔佳[22]以需求为导向，选择领域知识本体，并根据知识本体开展数据标准化，实现异构数据的汇聚。

2.2 WPRIM的元数据汇聚思路

面向提供西太平洋地区出版的生物医学领域文献、促进欠发达地区生物医学科技文献传播、提供及时准确数据服务的基本需求，破解现有数据加工人工成本高、历史数据不完整的难题，参考于倩倩等学者提出的基于元数据映射的多源异构数据汇聚策略，从系统需求与内容特征视角，补充数据源遴选制度、内容著录规范化与非结构化数据转换环节，增加J-STAGE等元数据标准的映射方法，形成如图1所示的WPRIM期刊元数据汇聚思路。

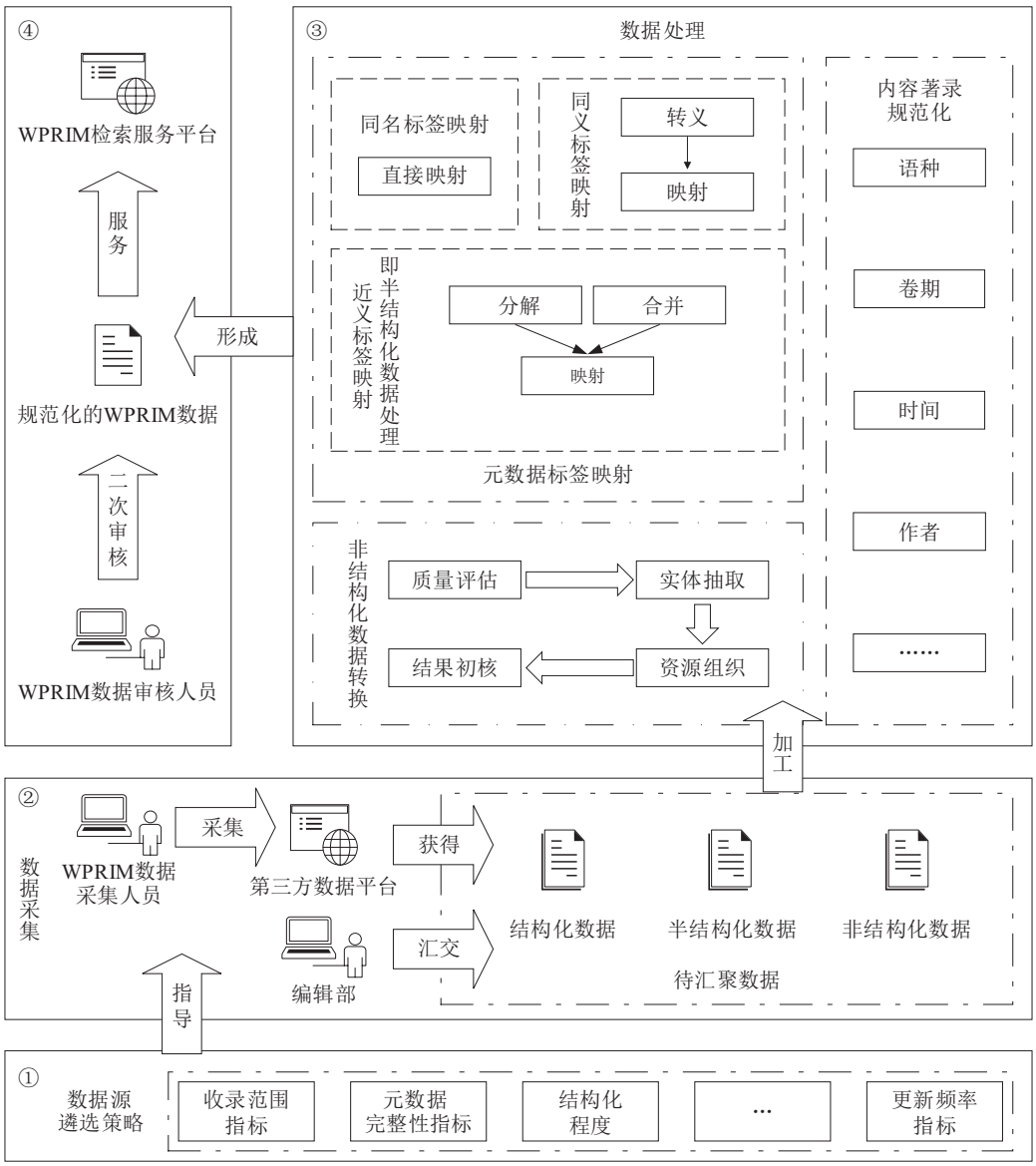

图1　面向WPRIM的期刊元数据汇聚思路

第一，面向WPRIM的期刊元数据汇聚通过设置不同场景下的指标及其权重确定数据源遴选策略指导数据采集及汇交（如①所示）。第二，数据采集人员和编辑部等分别通过采集第三方数据与提交文档的形式，提供结构化、半结构化与非结构化的待汇聚数据（如②所示）。第三，对待汇聚数据进行元数据标签映射、内容著录规范化的数据处理与非结构化数据转换，汇聚并形成规范化的WPRIM数据（如③所示）。第四，对规范化的WPRIM数据开展二次审核（如审核作者姓名是否为全拼），审核合格的数据通过WPRIM检索服务平台对外提供服务（如④所示）。

2.3　WPRIM元数据的汇聚实施方案

2.3.1　数据源遴选

遴选数据源指标和权重设置方面，WPRIM面向不同需求的服务场景设置6个一级指标、19个二级指标进行数据源评价。其中，一级指标包括收录范围、元数据完整性、结构化程度、期刊变更信息准确度、是否具有全文或全文链接、更新频率6个指标（见表1）。通过专家咨询法并结合系统需求场景的变化设置数据源指标权重。WPRIM的基本需求是占用较少的人力资源保证定期、批量更新期刊文献资源。在这一基本需求下，重

点考量数据收录范围、元数据完整性等要素。因此，收录范围、元数据完整性相关的二级指标在基本需求场景下的所占权重较高。遇突发情况时，用户的主要需求是快速获得第一手的科技论文资源。面对这类特殊需求，则以数据更新速率指标为最高权重来遴选数据来源。如新型冠状病毒疫情爆发初期，WRPIM平台与期刊编辑部合作，在不考虑数据是否结构化的基础上，提供人力支持，辅助编辑部优先汇交新型冠状病毒主题文献资源。同时，WPRIM监测国内外医学检索平台（如PubMed、KoreaMed、SinoMed、万方医学网等）、出版商（如中华医学会出版社等）的新型冠状病毒文献专题，及时发现优先出版的期刊文献资源，不严格限制文献资源来源唯一性。

表1 数据源遴选指标及权重

一级指标	二级指标	基本需求下的遴选权重（α）	特殊需求下的遴选权重（α）
收录范围	收录的最早出版年是否与期刊出版同步	0.150	0.025
	最新更新年度是否与期刊出版同步	0.150	0.025
元数据完整性	英文题名、英文摘要	0.133	0.075
	第一作者全名（英文表达形式）	0.067	0.075
	第一作者机构（英文表达形式）	0.067	0.075
	出版年、卷、期、页码	0.133	0.075
结构化程度	XML或JSON文件	0.075	0.017
	作者姓名是否分为姓、名、中间名	0.038	0.017
	每个关键词单独存储	0.038	0.017
期刊变更信息准确度	数据中刊名变更时间是否与实际一致	0.033	0.017
	数据中ISSN变更时间是否与实际一致	0.033	0.017
	是否有期刊变更信息的说明文档	0.033	0.017
是否具有全文或全文链接	DOI	0.007	0.067
	URL	0.007	0.067
	PDF全文	0.007	0.067
更新频率	按日更新	-	0.250
	按月更新	0.015	0.100
	按季度更新	0.010	-
	按年更新	0.005	-

数据源遴选实现上，当一本期刊有多个数据来源时，WPRIM通过计算各个数据源的分数值（Score），并取最大Score值对应的数据来源作为该本期刊的数据源。计算方法见公式（1）。

$$Score_{source}=\sum \alpha_i \times s_i \ (i \in I) \qquad (1)$$

其中，$Score_{source}$表示期刊某一数据源的分数值，等于指标及其权重乘积的和；I表示评价期刊数据源的全部指标，i表示I中的一个指标，S_i表示i指标的分值，a_i表示i指标在评价中所占权重。当满足指标要求时，S_i设为100，反之则设为0，若二级指标包含多个三级指标时，S_i平均分配至三级指标。

2.3.2 元数据标签映射

WPRIM参考全球医学索引、美国PubMed、韩国KoreaMed等文献检索系统元数据标准，提出并建立了WPRIM元数据方案，规定采用包括论文题名在内的12个元数据项描述文献资源。WPRIM汇聚的结构化、半结构化数据的元数据标签与WPRIM元数据标签存在同名、同义及近义3种情况。这3类标签的映射方法具体包括以下内容。

（1）同名标签的元数据映射。同名标签的元数据映射必须确定标签项的含义是否一致。如中文期刊数据中题名标签指中文题目，WPRIM的题名标签指文献的英文标题，两者含义是不同的。又如，JATS的期标签

（ISSUE）与WPRIM的期标签（ISSUE）的含义是相同的。在保证含义一致的情况下，采用直接映射取值的方式，实现同名元数据项取值。

（2）同义标签的元数据映射。通过对数据源元数据标签含义的调研，确定同义标签的对应关系，构建同义标签的转义工具，将非WPRIM元数据标准的数据标签转换为WPRIM元数据标准的元数据标签，实现同义标签数据的映射。

（3）近义标签的元数据映射。近义标签的元数据映射（半结构化数据处理）是将与WPRIM元数据标签近义的、内容半结构化的数据进行分解或重组，提取处理后的元数据值，并映射至目标元数据的相近标签。以J-STAGE的作者映射为WPRIM作者为例，J-STAGE元数据虽然部分利用XML结构化的形式存储数据，但作者（authors）包含非结构化的作者姓、名。这类数据被称为半结构化数据。通过分解，该半结构化数据被分解形成姓（LASTNAME）和名（FIRSTNAME）两部分，取值分别为Masahiro和Hamashima，并赋值于相应元数据项。

2.3.3 内容著录规范化

通过设置规范策略，在不破坏数据本身含义的前提下，对不同表达形式的内容进行分析与修正，统一数据内容形式，保证数据著录规范。异构内容依照WPRIM数据著录标准进行汇聚，对不满足著录标准的内容进行修正。常见修正内容包括语种、时间、卷期、作者的表达形式（见表2）。

表2 元数据标准及内容异构数据处理示例

对 象	元数据标准	修正前	修正后
语种	采用世界卫生组织语种表达形式，使用语种全拼	EN	English
		中文	Chinese
时间	采用公历日期标准格式，年、月、日采用数字表示，不足两位的数字使用0进行补全	2020-1-1	2020-01-01
		2019-July-20	2019-07-20
卷期	除增刊卷期外，采用数字表示卷期信息，不使用0进行补全	Volume：01	Volume：1
		Issue：01	Issue：1
作者[23]	作者姓名由“名+空格+中间名+空格+姓”组成，其中名首字母大写其他字母小写、姓全部大写	Gu（姓）Hongsheng（名）	Hongsheng（名）GU（姓）

2.3.4 非结构化数据转换

非结构化数据转换包括质量评估、实体抽取、资源组织与结果审核4个环节，实现非结构化数据转换为结构化数据，用于资源汇聚（见图2）。

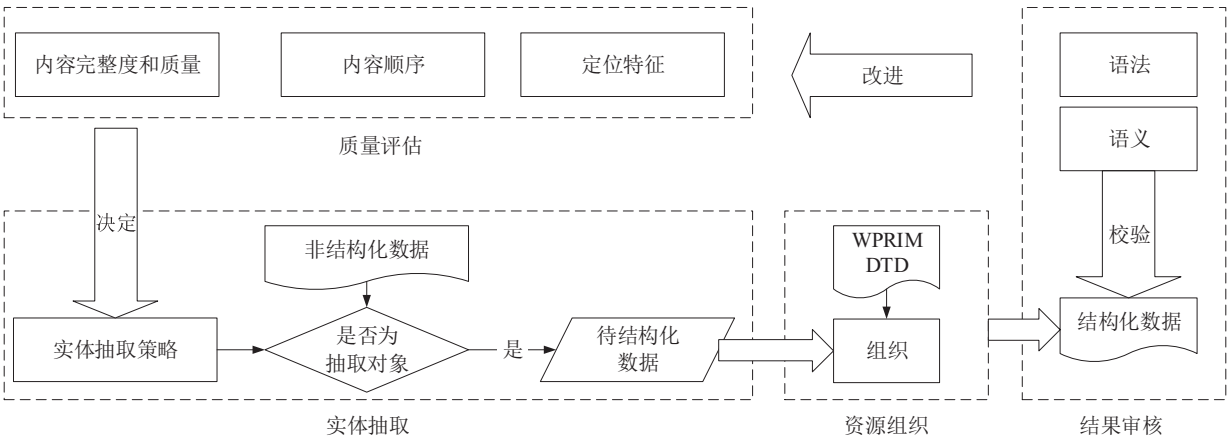

图2 非结构化数据转换流程

（1）质量评估。采用随机抽样分析的方法进行质量评估。即抽样一期或两期的非结构化文档进行内容、顺序、特征3个维度的质量评估。首先，内容层面对内容完整度和质量进行评估。内容完整度上，要求著录内容基本完整，包括但不限于英文题名、英文作者信息（姓名全拼、机构）、摘要、页码的数据内容。质量上，要求同一元数据位于同一行，如文献标题不出现中间换行。其次，要求非结构化文档内容顺序具有一致性。即同一本期刊题名、作者、关键词、摘要出现顺序保持一致。最后，非结构化文档应具备定位特征。即非结构化文

档存在明确识别出一篇文章的开始或结束的定位标志。“中图分类号：……”的内容是一篇文章开始的特征；“DOI：……”的内容是一篇文章的结束特征。这两个特征之间的内容符合英文标题、作者及机构、摘要、关键词的著录顺序，组成了WPRIM所需的数据项。

（2）实体抽取。依据质量评估的结果（特征、顺序）定制实体抽取策略，依次或分批提取英文题名、英文作者、摘要、页码等内容。

（3）资源组织。根据WPRIM元数据标准，对已抽取的实体信息进行结构化组织，形成符合WPRIM元数据标准的XML格式数据文件。

（4）结果校验。利用XML文件中指定的文档类型定义（DTD）对成果进行完整性和合理性自动校验。完整性校验判断是否缺失必备字段项，并对缺失必备项的数据进行剔除。合理性校验对数据内容是否合理开展语法与语义两个维度的审核。语法方面，开展诸如判断页码是否存在非数字字符、作者是否包含数字的语法检查。语义方面，开展诸如作者姓名拼写是否符合西方语言国家、南岛语系国家（如印度尼西亚、马来西亚、菲律宾）、汉藏语系国家（如中国）的内容检查。

3 结论与展望

针对WPRIM数据资源的同一期刊存在多个数据来源、不同数据源的元数据标签不一致、不同数据源著录标准不同、非结构化数据人工加工成本高的现状，从数据源遴选、元数据标签映射、内容著录规范化、非结构化数据转换4个维度设计多源数据汇聚方案，实现WPRIM收录期刊元数据的汇聚。WPRIM平台文献总量已由2016年的60余万篇增长至2020年的80余万篇，回溯非结构化期刊资源2万多篇，规范作者、卷、期、时间数据60余万篇，汇聚与规范成果已被全球医学索引、谷歌学术等文献检索平台收录。2020年，WPRIM平台月均文章点击量达到198 912次，较2018年月均文章点击量增长46%。

国内已开展或建成一系列“一带一路”、中国–东盟等跨国别的数据库，也面临各国数据资源来源、结构化水平和著录质量差异的挑战。结合世界卫生组织西太平洋地区医学索引的建设实践，未来多源数据汇聚可以参考以下5个方面加以改进。

（1）需求驱动汇聚数据资源的遴选。立足用户对文献资源的需求，梳理不同数据源的优势与不足，动态调整获取途径，通过不断完善数据资源遴选标准，快速汇聚成果并提供用户使用。

（2）关注元数据标准及其著录规范。元数据标签映射能实现资源汇聚，但仍存在一定不足。通过著录规范化的视角，一方面能够提高汇聚成果的质量；另一方面也能够减少重复数据的出现，降低数据归一与去重的工作量。

（3）开展精细化、互补化的多源数据融合。WPRIM数据是通过数据遴选制度确定唯一数据来源，从而降低数据去重工作量，加快数据更新效率。但在提高效率的同时，部分字段项内容缺失、预出版数据与正式出版数据重复的问题显现。WPRIM及其他相似索引平台应补充多源篇级论文精准匹配和字段及内容融合的研究，实现多源数据精细化、互补化的融合。

（4）拓展索引服务深度与广度。一方面，索引服务要深挖资源包含的知识内容，开展文献标引研究，深化数据内容，服务智能检索；另一方面，聚焦新媒体的资源传播场景，开展如社交媒体、视频等场景下的文献传播方法研究。

（5）构建数据汇聚的可持续发展机制。一方面，跨国别的资源汇聚平台需要开展国际合作交流，组织深入的数据管理培训，提升编辑或数据管理人员的计算机水平，指导其开展汇聚前的数据结构化，降低汇聚平台的数据复杂度；另一方面，引入自然语言处理、机器学习等不断出现的先进技术，实现精准匹配、文献标引等维度的数据深度融合。

参考文献

[1] 曾建勋. 开放融合环境下NSTL资源建设的发展思考 [J]. 大学图书馆学报，2020，38（6）：63-70.

[2] 赵志耘. 构建国家科研论文和科技信息高端交流平台 [J]. 数字图书馆论坛，2020（11）：1.

[3] 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.

[4] NCBI. Journal Article Tag Suite [EB/OL]. [2020-12-10]. https://jats.nlm.nih.gov/about.html.

[5] NCBI. Journal Publishing Tag Set Standard versions [EB/OL]. [2020-12-10]. https://jats.nlm.nih.gov/publishing/versions.html.

[6] NCBI. Journal Archiving and Interchange Tag Set [EB/OL]. [2020-12-10]. https://jats.nlm.nih.gov/archiving/.

[7] NCBI. Article Authoring Tag Set [EB/OL]. [2020-12-10]. https://jats.nlm.nih.gov/articleauthoring/.

[8] 刘冰，游苏宁. 我国科技期刊应尽快实现基于结构化排版的生产流程再造 [J]. 编辑学报，2010，22（3）：262-266.

[9] 姚伟欣，马建华. 新学术环境下科技期刊数字出版平台的技术发展趋势 [J]. 中国科技期刊研究，2013，24（6）：1039-1043.

[10] 苏磊，李明敏，蔡斐. 科技期刊采用XML结构化排版的优势与应用实践分析 [J]. 科技与出版，2017（10）：108-111.

[11] 化柏林. 多源信息融合方法研究 [J]. 情报理论与实践，2013，36（11）：16-19.

[12] 郭春霞. 大数据环境下高校图书馆非结构化数据融合分析 [J]. 图书馆学研究，2015（5）：30-34.

[13] 涂子沛. 大数据及其成因 [J]. 科学与社会，2014，4（1）：14-26.

[14] 曲建升，刘红煦. 知识发现中异构信息标准化处理研究——以资源环境领域文献为例 [J]. 图书情报工作，2016，60（6）：84-90.

[15] 林鑫，李想，李静. 资源发现系统中基于多源数据融合的文献元数据质量提升 [J/OL]. 情报理论与实践，2021：1-8 [2020-12-10]. http://kns.cnki.net/kcms/detail/11.1762.g3.20201203.1624.004.html.

[16] 周艳会，曾荣仁. 基于元数据的数据质量管理研究 [J]. 信息技术与信息化，2020（7）：26-29.

[17] BRUCE T R，HILLMANN D I. The continuum of metadata quality：defining，expressing，exploiting [C] //HILLMANN D I，WEATBROOKS E L. Metadata in Practice. Chicago：American Library Association，2004：238-256.

[18] MOGHADDASI J，WU K. Multifunctional transceiver for future radar sensing and radio communicating data-fusion platform [J]. IEEE Access，2016，4：818-838.

[19] 于倩倩，张建勇. NSTL集成利用第三方来源元数据的实践与探索 [J]. 现代图书情报技术，2016（1）：97-102.

[20] 刘盼雨，王昊天，郑栋毅，等. 多源异构文化大数据融合平台设计 [J/OL]. 华中科技大学学报（自然科学版），2021：1-8 [2020-12-10]. https://doi.org/10.13245/j.hust.210216.

[21] 侯鑫鑫，朱文佳，朱莉，等. 多源异构学术成果大数据的整合与揭示 [J/OL]. 情报理论与实践，2021：1-11 [2020-12-10]. http://kns.cnki.net/kcms/detail/11.1762.G3.20201204.1105.002.html.

[22] 崔佳. 基于领域本体的多元异构数据融合关键技术研究 [D]. 青岛：中国石油大学（华东），2018.

[23] 王蕾，方安，范云满，等. 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 [J]. 医学信息学杂志，2019，40（2）：75-80.

作者简介

王蕾，女，1989年生，硕士，助理研究员，研究方向：信息技术、大数据处理。
方安，男，1976年生，博士，研究馆员，研究方向：医学知识组织与数字图书馆。
杨雨生，男，1994年生，助理馆员，研究方向：信息技术应用。
范云满，男，1980年生，硕士，助理研究员，研究方向：医学数据自然语言处理、云计算环境下大数据分析算法。
王茜，女，1981年生，博士，副研究馆员，通信作者，研究方向：信息技术应用，E-mail：wang.qian@imicams.ac.cn。

Research on Multi-Source Journal Metadata Fusion:
Taking WHO Western Pacific Region Index Medicus as An Example

WANG Lei　FANG An　YANG YuSheng　FAN YunMan　WANG Qian
(Institute of Medical Information, CAMS & PUMC, Beijing 100020, China)

Abstract: Analyzing status of source data and problems on multi-source journal metadata fusion in WHO Western Pacific Region Index Medicus. This paper designs a multi-source data fusion scheme from source selection, metadata label mapping, content standardization, and unstructured data transformation. The result shows that the path can solve WPRIM multi-source data fusion and provide a reference for similar situations as well.

Keywords: Data Fusion; Multi-Source Data; WPRIM; Multi-Metadata and Heterogeneous

（收稿日期：2020-12-18）