

开放期刊资源元数据质量管理研究

董微 赵捷

(中国科学技术信息研究所, 北京 100038)

摘要: 开放期刊资源已成为数字图书馆的重要资源, 对开放期刊资源元数据进行质量管理成为开放期刊资源建设的重要内容。在总结开放期刊资源特点的基础上, 分析在开放期刊资源建设过程中影响开放期刊资源元数据质量的因素, 并对开放期刊资源元数据质量管理进行研究与探索, 构建开放期刊资源元数据质量的管理模型, 最后提出开放期刊资源元数据质量管理流程。

关键词: 开放期刊资源; 元数据; 数据质量; 质量评估; 质量管理

中图分类号: G250.7

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2018.03.012

Research on Metadata Quality Management of Open Academic Resource

DONG Wei, ZHAO Jie

(Institute of Scientific and Technical Information of China, Beijing 100038)

Abstract: Open journal resources have become the important resources of digital library, therefore, metadata quality management of open access is essential. In this paper, we summarized the characteristics of open journal resources, analyzed the influencing factors on the quality of the open journal resources metadata during the process of open resources construction, and researched and explored the open resources metadata quality management. At last, we built metadata quality management model of the open resources, and constructed the open journal resources metadata quality management process.

Keywords: open journal resource, metadata, data quality, quality evaluation, quality control

0 引言

伴随着网络信息基础设施的完善以及学术交流模式的转变, 开放获取运动逐步得到学术团体、政府等支持, 资源数量急剧增长。而开放期刊资源作为知识传承、传播的重要载体, 通过全面采集、组织与揭示实现了高效传播与有效利用, 形成“互联网+”时代满足用户需求的新信息资源体系。元数据作为揭示开放资源内容、属

性及特征等信息的一种数据形式, 在资源整合、资源检索、资源发现、资源评估及资源管理等应用中发挥着非常重要的作用^[1-4], 因而开放期刊资源的元数据质量直接影响着开放期刊资源的建设水平与服务质量。目前, 由于开放期刊资源获取渠道多样、著录格式不统一、管理流程不尽合理以及图书馆缺乏不同专业背景的技术人员等原因, 造成了开放期刊资源的数据质量参差不齐, 从而影响了开放期刊资源的有效利用。如何对开

作者简介: 董微 (1987—), 女, 博士, 中国科学技术信息研究所博士后, 研究方向: 数据挖掘、网络爬取 (通讯作者); 赵捷 (1959—), 男, 中国科学技术信息研究所高级工程师, 研究方向: 信息组织与构建、知识服务与知识链接、网络信息系统建设。

基金项目: 科技部科技创新战略研究专项“基础研究领域年度重点创新进展报告”(ZLY201636); NSTL专项基金项目“开放学术资源建设”(2016XM16)。

收稿时间: 2017年8月21日。

放期刊资源进行质量管理、建设高质量的元数据仓储、为用户提供更好的服务是图书馆亟待解决的问题。

开放期刊资源元数据质量管理是通过建立统一的元数据标准,对管理流程进行规划,并运用相应的技术和方法,对人员进行培训和要求,以提高元数据质量,最终向用户提供资源服务。目前,针对开放期刊资源的元数据质量管理尚不成熟,还没有一套能够适合国内实际情况的元数据质量管理方案。因此,本文结合国家科技图书文献中心(NSTL)开放期刊资源建设的实践经验,对期刊元数据质量管理的相关问题进行研究探讨。

1 影响开放期刊资源元数据质量因素分析

开放期刊资源具有分布广泛、来源各异,具有数量巨大、生命周期短暂、存储分散、类型格式复杂、资源组织异构、粒度层级复杂等特点^[5-6],而开放期刊资源元数据采集、加工的单位众多,著录者对元数据元素理解不准确,使得元数据在质量方面存在参差不齐的现象,不仅降低了开放期刊资源服务的质量,而且为开放期刊资源信息的采集、整合和服务带来了新的挑战。

结合实际工作,通过分析发现,目前影响开放期刊资源元数据质量的主要因素有元数据、管理流程、人员、技术支持等4个方面。

(1) 元数据。在开放期刊资源建设过程中,元数据的生成主要通过两种方式:一是利用软件通过元数据抽取(metadata extraction)和元数据收集(metadata harvesting)自动生成^[7];二是对采集到的元数据进行补充加工。然而,通过两种方式生成的元数据质量存在诸多问题,比如:数据源所在平台较多、页面揭示多样、更新时间不稳定、元数据厚薄不均、全文格式不统一等,从而造成采集的元数据准确性、完整性、及时性、有效性较差;数据补充加工由不同的著录者进行著录,这是一种分散的、难以控制的行为,若著录者对元数据元素理解不准确,则造成数据著录不一致、格式各异等问题。

(2) 管理流程。开放期刊资源数据量巨大,对于大型服务系统往往是由多成员单位共同承担建设,因此科学性、规范化的管理流程至关重要。通过从整体流程考虑元数据的质量控制,对元数据的质量控制贯穿于元数据的整个生命周期,实现对元数据质量的完全控制,从根本上提高元数据质量。

(3) 人员。开放期刊资源的元数据质量管理是一项复杂的工作,对于工作人员的专业水平要求较高,如对编目流程、著录规则和著录规范的掌握、对数据定期抽样审核以及辅助工具的使用等,因此需要安排专业人员把控各个环节,及时发现其中的错误并进行改正,以防止问题的长期积累。

(4) 技术支持。随着开放期刊资源的飞速增长,描述资源的元数据数量也在剧增,完全依赖于人工的质量控制已经无法实现,因此机器的自动化辅助十分必要。

2 开放期刊资源元数据质量管理模型

本文提出的模型构建基于多成员单位协同建设开放期刊资源,考虑到开放期刊资源不同于普通的数字资源,具有动态更新的特点,即对于已发布的数据由于字段缺失、错误等原因导致再次发布数据等问题。本文在总结NSTL开放期刊资源质量管理工作经验、分析影响开放期刊资源元数据质量因素的基础上,构建了开放期刊资源元数据质量管理模型。

该模型主要由三大组件构成,分别为元数据生命周期、元数据质量维度以及影响要素(图1)。元数据的生命周期说明了开放期刊资源元数据在整个生命周期的管理活动;元数据的质量维度则提供了评估和控制元数据在整个生命周期质量的依据;影响要素是影响元数据管理与质量控制的要素。这三大组件紧密联系,相辅相成。下面对构成元数据质量管理模型的三大组件进行详细介绍。

2.1 开放期刊资源元数据生命周期

元数据的生命周期反映了元数据的管理活

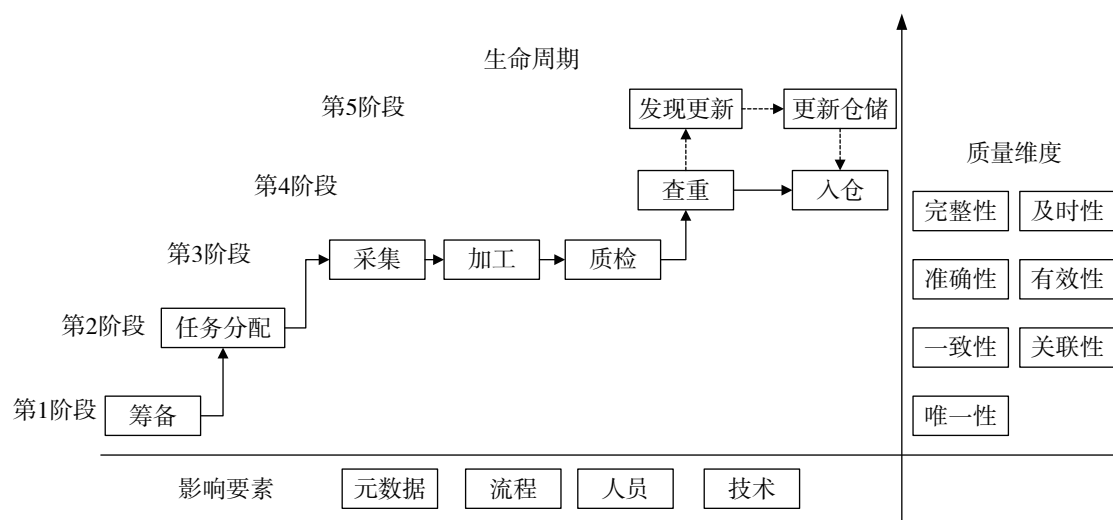


图1 元数据质量管理模型

动，元数据质量控制应该从元数据创建的全局考虑，贯穿于元数据整个生命周期，实现元数据质量的全程控制，从根本上提高元数据质量。在各元数据的整个生命周期里，都要依据下面即将叙述的数据质量维度实施质量控制。本文考虑了多成员单位协同建设开放期刊资源，并且采用了“总一分一总”的管理模式，将元数据的生命周期划分为以下5个阶段。

（1）筹备阶段。在元数据采集的筹备阶段，需要构建资源的遴选标准、制定统一的元数据标准、元数据标准的执行细则以及元数据的质量标准。资源的遴选标准首先将质量较高的期刊资源进行遴选；元数据标准一般由元素名称、标识符、定义/描述、注释、著录规则以及示例组成；元数据标准的执行细则是对元数据标准的进一步描述与统一规定；元数据的质量标准对每个环节提出质量要求。

（2）任务分配阶段。由于考虑到多成员单位协同建设开放期刊资源，需要在任务分配阶段将不同的任务分给不同的成员单位，避免重复采集。

（3）采集、加工、质检阶段。各成员单位在接到采集任务后将完成元数据在该阶段的生命周期，包括采集、加工、质检的工作。

采集：将不同来源、不同渠道的元数据按照统一的元数据标准与规范进行采集。

加工：由于不同平台来源的数据页面揭示多样、更新时间不稳定、元数据厚薄不均、全文格式不统一，导致完全依赖于工具进行采集的数据质量较差，在加工阶段，需要著录人员进行补充加工。

质检：在元数据质检阶段，结合机器自动化质检的同时，需要人工进行抽样检查，汇总入仓之前需要进一步进行机器质检与人工抽检，分析可能存在的数据问题，并利用辅助修正机器质检功能。

（4）查重和入仓阶段。将不同成员单位提交的元数据进行查重，构建统一的元数据仓储。对于完全重复的数据则直接剔除；对于相同卷期下的数据，相似度大于阈值则机器进行冲突解决，合并为一条数据，并对元数据进行字段补齐，增加元数据厚度；若是不同数据，则对数据均进行入仓处理。

（5）更新数据与更新仓储阶段。对于不同卷期下的数据，相似度大于阈值则需要人工进行核查，对开放期刊资源动态更新数据，并及时更新数据仓储的数据。

2.2 开放期刊资源元数据质量维度

元数据的质量维度提供了评估和控制元数据在整个生命周期质量的依据。国外专门讨论元数据质量的研究大约始于2004年。Moen等^[8]在

GILS项目评估中提出了一套元数据质量评估的指标和方法。该模型设计的出发点是GILS系统的评估而非仅仅针对元数据质量,评估范围广、内容多,使得该评估模型难以在更广的范围内推广^[9]。Stvilia等^[10]在Moen等^[8]提出模型的基础上提出了评估信息质量的模型。该模型的使用需要再根据元数据创建、使用环境进一步明确各指标的涵义、确定量化方法。Bruce和Hillman^[11]对Stvilia等^[10]提出的指标维度加以浓缩,归纳为7个指标。该模型的使用更为广泛。

本文综合考虑开放期刊资源的特性、元数据规范、管理流程以及图书馆实际业务需求等因素,制定了以下7个方面的开放期刊资源元数据质量维度,该元数据质量维度贯穿于上述整个管理流程。

(1) 完整性。根据元数据标准和规范要求,详尽地描述资源信息。如果元数据不完整,将会影响其他质量的评价,最终影响整个元数据质量管理的流程。

(2) 准确性。元数据对资源内容客观、准确反应的程度。

(3) 一致性。保障元数据所有字段所描述的资源是同一资源,避免不同字段指向不同资源。

(4) 唯一性。保障元数据仓储中的数据都是唯一的,没有重复,以更好地为用户提供服务。

(5) 及时性。针对资源易逝性的特点,及时获取开放资源信息,对资源进行保障。

(6) 有效性。开放期刊资源具有动态性,需要动态地监测元数据是否仍然有效,是否被更新、删除等。

(7) 关联性。揭示开放资源与相关资源、相关系统及数据库的关系,如印本转为开放获取的资源与其对应的印刷型资源的关联。

2.3 影响要素

元数据质量的影响要素是影响元数据管理与质量控制的关键因素,主要有以下4个方面。

(1) 元数据。元数据是质量控制与管理的研究对象,对于多来源、多类型的元数据进行质检,把控质量,为用户提供高品质服务。

(2) 流程。对元数据管理的流程进行合理规划,保障元数据的管理活动能够科学开展。

(3) 人员。开放期刊资源元数据质量管理需要在不同的管理环节进行把控,如元数据标准的制定、元数据的录入、元数据的分析以及元数据的冲突处理等,因此需要多种不同专业背景和专业技能的技术人员团结协作,共同承担质量管理的重任。

(4) 技术。由于元数据质量管理的相关工作需要机器辅助完成,因此需要选用相关的系统、技术、软件、工具等,如数据格式转换工具、元数据格式校验工具、数据查重系统、数据仓储等。

3 开放期刊资源元数据质量管理流程

在NSTL实际开放期刊资源建设工作中,对质量管理流程采用“总一分一总”的管理模式。根据上述模型中所述的元数据的生命周期构建元数据质量管理流程,在各阶段,利用质量维度对元数据质量进行控制。由于开放资源数量巨大,因此在质量检查上采用以机器为主、人工为辅的抽检方式,制定了对开放期刊的元数据质量管理的流程,如图2所示。

(1) 第一总阶段,根据遴选标准遴选出质量高的品种,并根据制定的查重规则对品种进行去重操作,保证各成员单位承担的采集任务并无重复。

(2) 第二分阶段,各成员单位在采集数据时应构建本地仓储数据库,避免数据重复采集;根据元数据标准、执行细则,结合上述的质量维度,制定自动化的格式、内容质检工具,并根据工具的质检结果,人工进行质量抽检。若数据质量合格方可提交至总仓储。

(3) 第三总阶段,将各成员单位提交至总仓储的数据再次查重,在此阶段可以发现第一阶段品种重复分配的可能,并及时调整第一阶段任务;质检人员再次利用工具进行格式、内容的质量检查,在此阶段可以发现由于网络或人工疏忽等原因造成的文件丢失等问题。再次的人工质量抽检可以有效地发现采集数据质量的问题,并能

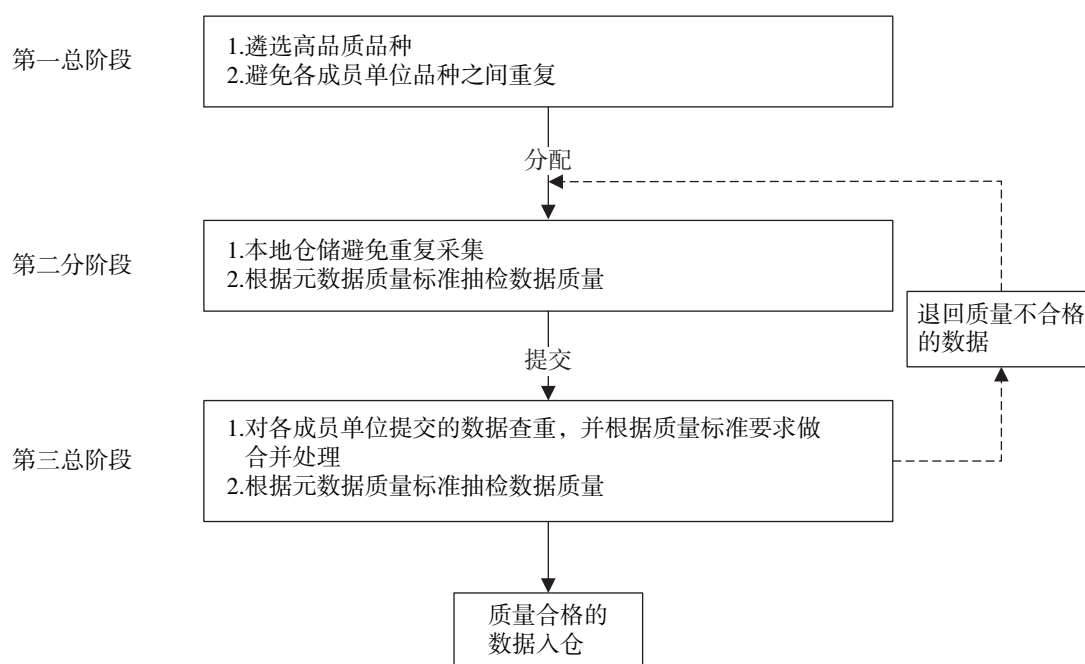


图2 元数据质量控制流程

及时调整辅助工具, 提高工作效率。

4 结语

本文提出的开放期刊元数据质量管理模型对于承担大型资源建设具有非常重要的作用, 采用了“总—分—总”的管理模式, 多单位可以协同工作, 共同建设开放期刊资源。根据开放期刊的特性, 制定了元数据质量的维度, 为元数据的采集标准提供了参考, 提高开放期刊资源元数据质量, 为用户提供更好的服务。然而, 开放资源具有易逝性, 如何探测资源的有效性及可访问性需要进一步的探讨和研究。

参考文献

- [1] 冯秀珍, 陈旒. 电子资源元数据的自动识别研究[J]. 情报杂志, 2010, 29(4): 130-133. DOI: 10.3969/j.issn.1002-1965.2010.04.029.
- [2] 李凌杰. 特色数据库建设中的元数据质量控制研究[J]. 图书情报工作, 2010, 54(5): 43-46.
- [3] 程颖. 数字资源元数据质量管理的研究与探索[J]. 图书馆, 2015(7): 66-69. DOI: 10.3969/j.issn.1002-1558.2015.07.018.
- [4] 杨代庆, 丁遭劲, 周锐. 国家开放信息资源建设策略研究[J]. 数字图书馆论坛, 2015(8): 16-21. DOI: 10.3772/j.issn.1673-2286.2015.08.003
- [5] 王思丽, 马建玲, 王楠, 等. 开放知识资源的元数据自动采集策略研究[J]. 图书馆学研究, 2013(12): 47-51.
- [6] 袁小一, 苏智星. 浅谈特色数据库元数据的建立[J]. 晋图学刊, 2005(5): 26-28. DOI: 10.3969/j.issn.1004-1680.2005.05.008.
- [7] 林爱群. 机构知识库元数据的自动生成与评估研究[J]. 图书馆学研究, 2009(7): 21-23. DOI: 10.3969/j.issn.1001-0424.2009.07.006.
- [8] MOEN W E, STEWART E L, MCCLURE C R. Assessing metadata quality: findings and methodological considerations from an evaluation of the U.S. government information locator service (GILS)[C]//research and technology advances in digital libraries: IEEE International Forum. IEEE, 1998: 246-255. DOI: 10.1109/ADL.1998.670425.
- [9] 黄莺, 李建阳. 元数据质量评估方法及模型研究[J]. 图书馆学研究, 2013(12): 52-56. DOI: 10.15941/j.cnki.issn1001-0424.2013.12.007.
- [10] STVILIA B, GASSER L, TWIDALE M B, et al. A framework for information quality assessment[J]. Journal of the American Society for Information Science & Technology, 2014, 58(12): 1720-1733. DOI: 10.1002/asi.20652.
- [11] BRUCE T R, HILLMANN D I. The continuum of metadata quality: defining, expressing, exploiting[A]. Metadata in Practice, 2004: 238-256.