

多来源元数据集成中的组织管理框架研究^{*}

丁道劲, 马袁燕, 李勃慧

(中国科学技术信息研究所, 北京 100038)

摘要: 国家图书馆、CALIS及NSTL均在开展资源发现系统构建, 特别是多来源元数据集成的相关实践, 然而元数据资源的获取渠道相对单一, 尚未形成标准化、模块化的工作流程。通过对比澳大利亚、日本等国家图书馆的元数据集成实践, 提出一个面向多来源元数据集成的组织管理框架, 旨在形成统一的元数据标准、固化元数据获取流程、明晰相关利益主体间的权益关系。

关键词: 元数据集成; 发现服务; 组织管理

中图分类号: G250

DOI: 10.3772/j.issn.1673-2286.2017.12.009

元数据是关于数据的数据。按照馆藏文献的加工揭示粒度, 除描述母体的书目级元数据外, 还涵盖描述单篇论文题名、作者等信息的篇章级元数据。由于书目数据是图书馆OPAC系统建设的数据基础, 所以早期以印本进行母体信息揭示的编目工作尤为重要, 成为对机构内部及机构间文献资源共享的重要前提条件。但伴随着数字出版和数据库的发展, 文献资源揭示粒度已从书目层级逐步深化到篇章级乃至文章内部的图表、实体、公式等层级, 语义出版在出版界方兴未艾, 如Primo、Summon、EDS这类文献资源发现系统以亿计的篇章级元数据集成显现出强大的规模效应, 使图书馆原本以书目元数据资源集成为优势的地位受到挑战。特别是在当前许多高校图书馆以电子资源为主体的文献资源建设时期, 图书馆单纯以编目方式进行的书目数据集成工作受到挑战, 亟须向多来源元数据集成方向转变, 在自主加工元数据的同时, 不断拓展网络采集、第三方提供等其他来源渠道。面对庞杂的元数据资源来源渠道, 固化元数据资源获取流程、明晰元数据集成方与提供方的权益关系等有效的组织管理工作, 将在较大程度上提升元数据集成的工作效率, 为图书馆更大规模地集成、更精准地发现资源奠定坚实的数据基础。

1 国内多来源元数据集成研究进展

多来源元数据集成的目的在于改进异构资源的整合与发现效率。伴随文献资源的数字出版程度不断加深, 图书馆在向数字资源建设模式转变的同时也在不断通过各种努力对数字资源进行整合。如建立期刊导航和数据库导航、提供全文链接服务、建设联邦检索系统等, 但都存在一定局限性^[1]。2009年1月, 全球第一个网络级资源发现系统——Summon发布, 该系统通过整合海量的异构资源元数据, 形成统一的元数据索引, 为用户提供快速有效的资源发现与传递服务^[2], 由此基于资源发现系统的文献资源整合模式得到迅速发展。2011年以来, 国内对资源发现系统的研究逐渐增多, 国内图书馆对资源发现系统的实践程度主要集中于商业性资源发现系统的引进和应用, 因此相关研究多为资源发现服务系统构建原理分析^[3]、商业性资源发现系统的选型与评估^[4]、资源发现系统在图书馆服务中的应用^[5]等。同时, 当前商业资源发现系统中的元数据资源处于集成状态而非经过有效地知识组织与关联, 所能提供的发现服务层次处于“资源”而非“知识”层次。彭佳等以上海交通大学图书馆“思源探索”系统为例, 指出需要在元数据仓储的基础上, 进一步构建资源组织和

^{*} 本研究得到国家社会科学基金重点项目“基于知识组织的图书馆资源发现服务体系研究”（编号：17ATQ002）资助。

深度聚合的整体框架体系,从语义层面上实现信息资源的组织和聚合^[6];曾建勋等在国家科技信息发现服务体系构建中,重点强调本体、叙词表等知识组织工具在文献资源语义知识组织中的应用,促进资源发现系统功能从单一资源级发现服务向引文分析、科研实体识别等知识服务方向拓展^[7]。

多来源元数据集成是资源发现服务系统的数据基础,因此资源发现服务系统构建和实践情况,可在一定程度上反映多来源元数据集成的进展。在我国,除超星、万方数据库等中文商业性资源发现系统外,国家图书馆等国家级文献资源服务机构也正在强化资源发现服务系统建设,构建多来源元数据资源建设渠道体系。其中,国家图书馆“文津搜索”系统通过收割、转换、清洗、装载不同来源的元数据,并对馆藏资源整合,构建了超大型海量元数据仓储;其元数据来源于不同系统、不同机构,涵盖中西文MARC、DC、XML以及MDF等多种数据格式文件^[8]。CALIS在三期项目建设中,建立面向图书馆联盟“云上的”信息服务协作网络。其中,资源汇集与交换平台负责收集各成员馆提交的网上免费资源、外部购买资源以及由合作方提供的资源,成员馆数据直接汇集到CALIS全国交换中心,或是成员馆数据先汇集到CALIS省级分中心,再汇集到CALIS全国交换中心^[9]。NSTL是在原有依托“印本采集+自主加工”的元数据资源建设模式下,不断向多来源元数据集成方向转变,并于2015年起陆续与数据集成商(如Clarivate Analytics)、出版社(如Taylor Francis等)、图书代理公司(如中国图书进出口[集团]总公司等)达成元数据合作协议。

总体而言,我国在多来源元数据集成方面已有一定研究和实践,但由于当前元数据资源获取渠道相对单一,获取方式相对简单,具有一定局限性,造成元数据集成中的组织管理工作相对粗放,尚未形成标准化、模块化的工作流程,不利于资源发现系统建设。因此,亟待强化前期组织管理工作,提升元数据资源获取质量和效率。

2 多来源元数据集成中的组织管理需求分析

多来源元数据集成中的组织管理目的在于使各类型、各来源渠道均能按照一定标准规范,高效有序地提交元数据资源,并确保资源集成方、提供方、使用方等

各利益主体间的权益均衡,形成良性循环。具体而言,组织管理工作需包含以下三方面内容。

(1) 制定统一的元数据标准规范。异构元数据资源整合有赖于统一的元数据标准,需要在不同标准间建立映射关系,因此在数据准备阶段应向元数据提供方明确可接收/兼容的数据标准、字段必备程度等,甚至提供统一的元数据标准规范文本,从而提高元数据提供方资源供给的准确性。

(2) 固化元数据获取流程。总体而言,元数据获取流程涉及数据准备、数据提交、数据校验三个环节。数据准备阶段包含向元数据提供方明确数据提交要求,由元数据提供方自行判断是否符合提交要求;数据提交则须说明支持的数据打包、传输方式等;数据校验包含格式校验和内容抽查,利用数据校验工具和人工审核方法分时段抽验数据的提交情况。

(3) 明确相关利益主体间的权益关系。集成海量元数据资源的目的在于高效使用数据,明确各利益主体间的权益关系,以利于资源发现系统的平稳构建与使用。因此,在数据准备阶段,就应通过相应协议文本明确元数据提供方式(免费或有偿)、元数据版权归属、元数据使用方式与范围等,以免产生后续纠纷。

3 国外多来源元数据集成实践的组织管理框架

在国外,无论是商业性资源发现系统还是公益机构建设的资源发现服务系统均有多来源元数据集成管理方面取得成功实践。商业性资源发现系统如Summon、EDS、Primo以及Scopus等同时具备发现和计量评价功能的外文数据库产品;由非盈利性组织构建的资源集成平台如澳大利亚国家图书馆Trove系统、日本国立国会图书馆NDL Search、欧洲虚拟博物馆Europeana等,以及专门整合开放期刊、开放图书、开放学位论文等开放资源的集成平台DOAJ、PubMed Central等。综合上述平台的元数据采集规范、流程与方法实践,其多来源元数据集成工作基本遵循图1的组织管理框架,主要目的是固化和细化工作流程,提升多来源元数据集成工作的效率。

总体而言,该组织管理框架分为基础层、操作层和应用层三个层次。在元数据集成方制定统一数据标准的基础上,元数据提供方按照数据准备、数据提交、数据校验等流程实现多来源元数据资源的转换和映射,最终由数据集成方实现元数据汇集和整合,形成有效

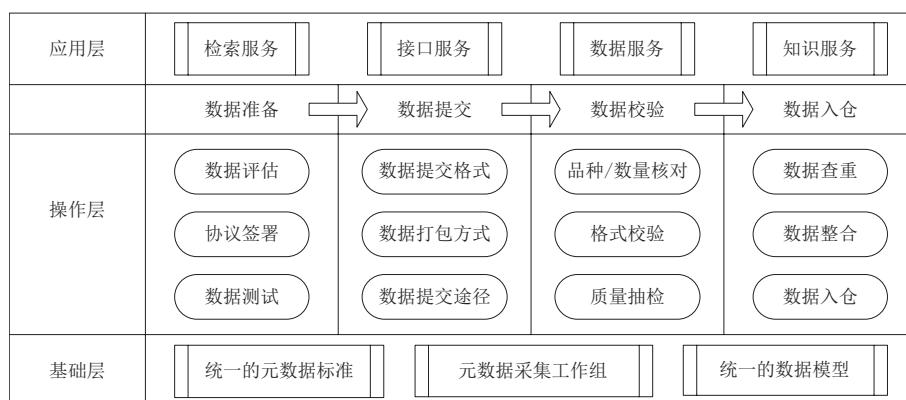


图1 面向多来源元数据集成的组织管理框架

的元数据资源集合，以满足各类用户需求。

3.1 基础层

在多来源元数据集成工作前期，须先构建一个元数据采集工作组，专门负责多来源元数据资源采集工作。如Trove的宏观调控和运行由Trove小组负责，小组成员是来自澳大利亚国家图书馆的高级管理人员及来自地方图书馆的工作人员。Trove小组主要由四部分组成，分别负责其战略方向、资源供给、资源构建、数据提供及维护，各部门间相互协作，职责分明^[10]。此外，制订统一的元数据标准是多来源元数据集成工作需要解决的重要问题，也是后期不同来源、遵循不同标准的元数据进行映射转换的依据，如日本国立国会图书馆NDL Search、欧洲虚拟博物馆Europeana的发现系统均已制定统一的元数据标准。特别是Europeana参照METS和RDF等相关标准和规范，设计开放、跨领域的Europeana Data Modal数据模型，在此基础上确立揭示数字文化资源不同概念间关联关系的语义知识本体模型，通过执行本体与关联数据融合驱动的元数据语义化操作，对跨机构、跨领域的数字文化资源进行关联与整合，形成一个语义关联数据网络^[11]，为其后期开展深层次知识服务提供数据基础。

3.2 操作层

3.2.1 数据准备阶段

该阶段是元数据集成方与提供方交互的初始阶段，需要相互了解数据基本情况及数据提交要求，包括数

据评估、协议签署及数据测试。(1) 数据评估内容包括是否为正规出版物、涵盖资源品种、元数据遵循标准及提供的数据格式是否在资源集成方可兼容范围内。如在资源品种层面，PubMed Central对期刊遴选的基本要求包括拥有注册的ISSN号，出版商在生命科学领域最少拥有2年出版经历，每种期刊至少包含2.5篇同行评论论文；在元数据层面，PubMed Central推荐出版商提供的数据应遵循NISO的JATS标准，并同时提交XML格式的期刊论文全文且不接受HTML格式。(2) 协议签署的目的主要是明确资源集成方与资源提供方的权责关系。如资源提供方在Trove系统提交元数据前须签署《国家发现服务数据贡献理解备忘录》，该协议文本对知识产权、费用、双方义务均进行明确约定。(3) 数据测试是在正式数据提交前进行双方磨合。相比而言，日本国立国会图书馆的元数据操作流程较严谨，测试工作被细分为样本数据提取、系统连接和协调测试两部分内容，以确保后期工作的顺利开展。

3.2.2 数据提交阶段

在双方确认合作、完成初步数据测试后，将进入数据正式提交阶段。对元数据集成中的组织管理工作而言，明确元数据提交方式和提交途径是此阶段的主要内容。一般而言，元数据提交方式包括OAI-PMH数据收割、API、网络爬虫、FTP传输或HTTP等方式。元数据提交途径指数据提交的组织结构，既包括资源提供方（资源建设方的单一结构），也包括多层级的网络结构。如在Europeana的数据提交网络中，资源提供方提供数据到聚合器（项目组或组织），随后处理该数据并将其提供给Europeana。CALIS也采用类似的数据汇集

方式, 构成“成员馆—CALIS省级分中心—CALIS全国交换中心”三级数据汇交体系。

3.2.3 数据校验阶段

资源集成方在获取各渠道数据后, 需要进行相应的格式校验和质量抽检工作以确保提交的质量。一般而言, 格式校验可通过程序自动完成。如PubMed Central的在线格式校验工具^[12]可基于自身格式规范对数据进行验证, 对不符合规范的数据项进行警告或报错。质量抽检更注重元数据的内容层面(如作者与机构的对应关系、关键词拆分准确性等), 以人工抽查的方式确保数据质量的进一步提升。

3.2.4 数据入仓阶段

数据入仓涵盖前期数据查重、数据整合和数据入仓三个环节。不同来源、不同渠道获取的数据必然在种类、卷期乃至篇级数据方面存在交叉重复现象, 通过在母体、卷期、篇级、渠道四个层面对数据进行查重和集成整合, 增加元数据厚度, 形成有机的元数据资源。如某商业发现系统对印本资源与电子资源数据主要通过以下三个方面进行匹配和融合: (1) 每个ID至少有一个唯一标识符匹配, “ISBN to ISBN” “EISBN to EISBN” “ISSN to ISSN” “EISSN to EISSN” “LCCN to LCCN”; (2) 记录必须有且仅有一个出版年, 允许印本资源出版年份与电子资源出版年份差距在一年内; (3) 使用模糊逻辑匹配题名, 至少匹配一个题名。

3.3 应用层

基于统一的元数据资源, 目前主要应用方式仍处于资源层级的检索服务, 由此衍生出当前市场上众多资源发现系统产品, 却远不能满足用户需求, 相比于全文数据库的语义化出版粒度明显滞后。因此, 随着海量元数据资源的知识组织深度不断加深, 及关联数据等在元数据集成中的应用, 提供基于作者、机构等海量科研实体的分析服务将成为可能。同时, 基于中央元数据仓储的个性化数据定制以及应用系统的嵌入, 也是元数据集成的一种服务形式。如日本国立国会图书馆NDL Search提供API接口服务, 支持其他公共图书馆、博物馆、美术馆等的系统建设与服务。

4 对我国多来源元数据集成工作的建议

(1) 拓宽元数据获取渠道, 重视版权归属处理。

目前, 我国公益机构的元数据集成实践主体主要包括国家图书馆、CALIS以及NSTL, 或局限于系统内部合作, 或基于自身数据基础重点开拓图书代理公司、出版社等渠道, 尚未在系统内外达成广泛合作, 这与国外部分国家图书馆的建设机制恰好相反。如澳大利亚国家图书馆Trove系统集成澳大利亚的图书馆、博物馆、档案馆和其他科学文化机构的5.3亿条有关澳大利亚文化遗产的元数据、全文和链接资源^[13], 包含超过2 000家澳大利亚的图书馆、博物馆、档案馆和其他文化机构的资源^[14]。同时, 在与其他机构合作时, 发现系统构建主体已对合作对象的合作内容及合作要求进行明确表述, 并形成相应的协议文本, 这对我国文献信息服务机构推进跨界合作, 具有重要借鉴意义。因此, 我国在开展相关数据合作时应重视相应合作协议文本的撰写, 对合作内容、版权归属等问题进行清晰界定。

(2) 加快统一元数据标准与其他主流标准映射工作, 固化元数据提交要求与流程。统一元数据标准是在更大范围内进行元数据集成、提高集成效率的有效途径之一。目前, 我国主要文献保障系统内部已形成一定规模的元数据标准, 但仍存在系统性的条块分割。因此, 一方面要加强系统间元数据标准的交流、探讨和宣传, 在统一元数据标准建设方面尽可能达成共识; 另一方面, 要加强元数据标准注册系统建设, 实现单一文献机构、出版社、图书代理公司等元数据标准的相互转换和映射。同时, 我国的大范围元数据集成工作应明确和固化元数据的合作层级(书目级或篇章级)、可提交元数据格式、元数据上传方式(FTP、网盘、HTTP等)、格式转换及校验流程等, 在网站显著位置明确说明, 便于元数据合作方对前期的评估以及后续工作的开展。

(3) 拓展合作内容, 构建多元化元数据资源服务体系。从元数据获取渠道看, 与出版社、集成商的合作可快速获取大规模元数据资源, 与馆藏机构合作的目的在于馆藏信息及后续服务保障。如澳大利亚国家图书馆、日本国立国会图书馆与其他文献信息服务机构的合作主要在元数据资源层面, 为用户提供相应的全文获取渠道(印本资源馆藏位置、电子资源全文链接等)。从现有服务实践看, 国外发现系统尚未完全实现后端服务与本地系统的整合。建议我国大型文献保障

机构在与其他文献信息服务机构开展元数据合作中,除数据本身外,增加原文传递服务整合机制,以及通过API方式将资源嵌入合作方本地文献信息系统,从而增加合作双方黏性,保证资源发现系统功能的稳定性和运行的持续性。

参考文献

- [1] 包凌,蒋颖.图书馆统一资源发现系统的比较研究[J].情报资料工作,2012(5):67-72.
- [2] 蹇天芳,姜爱蓉.资源发现系统功能分析及应用前景[J].图书情报工作,2012,56(7):38-43.
- [3] 聂华,朱玲.网络级发现服务——通向深度整合与便捷获取的路径[J].大学图书馆学报,2011,29(6):5-10.
- [4] 秦鸿.关于发现系统的问题与思考[J].数字图书馆论坛,2012(7):17-20.
- [5] 陈雪,邹小筑,邱彦涛.资源发现系统在科研课题跟踪服务中的应用[J].图书馆论坛,2015(5):68-74.
- [6] 彭佳,郑巧英.信息资源聚合与组织研究——以发现系统为例[J].图书馆杂志,2016(3):80-85.
- [7] 曾建勋,丁道劲.基于语义的国家科技信息发现服务体系研究[J].中国图书馆学报,2017,43(4):51-62.
- [8] 杨东波,邢军.国家图书馆“天津搜索”的设计与实现[J].国家图书馆学刊,2014(3):93-98.
- [9] 王文清,张月祥,关涛,等.CALIS资源整合与云服务[J].国家图书馆学刊,2015,24(2):36-42.
- [10] 肖希明,李琪.公共数字文化服务合作机制研究[J].图书与情报,2016(4):31-37.
- [11] 王萍,黄新平.基于关联开放数据的数字文化资源语义融合方法研究——欧洲数字图书馆案例分析[J].图书情报工作,2016,60(12):29-37.
- [12] File Validation Tools[EB/OL].[2017-11-25].<https://www.ncbi.nlm.nih.gov/pmc/pub/validation/>.
- [13] Trove Content Inclusion Statement[EB/OL].[2017-11-25].<http://help.nla.gov.au/truve/our-policies/truve-content-inclusion-policy#anchor-0>.
- [14] Trove.Current Work Counts by Contributor[EB/OL].[2017-11-25].<http://trove.nla.gov.au/system/counts>.

作者简介

丁道劲,女,1988年生,馆员,研究方向:数字资源建设,E-mail:dingqj2011@istic.ac.cn。
马袁燕,女,1993年生,硕士研究生,研究方向:数字资源建设。
李勃慧,女,1994年生,硕士研究生,研究方向:数字资源建设。

Research on the Organization and Management Models for Multi-Source Metadata Integration

DING QiuJing, MA YuanYan, LI BoHui
(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: In China, the National Library, CALIS and the National Science and Technology Library have developed their own resource discovery system, including integrating the multi-source metadata, but the systems still need to be upgraded in terms of diversity of metadata acquisition channel and standardization of workflow. Based on the metadata integration practice of the national libraries in Australia and Japan, this paper constructs an organization management framework for multi-source metadata integration, which aims to form a unified metadata standard, solidify metadata acquisition workflow, and clarify the interest relationship among relevant stakeholders.

Keywords: Metadata Integration; Discovery Service; Organization and Management

(收稿日期: 2017-11-15)