自建数字资源的元数据管理实践及启示
——以国家图书馆中文图书数字化资源库为例

□ 吴茗 龙伟 肖红 / 国家图书馆 北京 100081

摘要：结合国家图书馆中文图书数字化资源库元数据管理的实践，提出自建数字资源元数据管理的基本要求，从健全的元数据规范和评价体系的建设、元数据与对象数据的关联方式的选择、元数据互操作的实现以及完善的元数据管理平台的构建等四个方面分析所涉及的技术，并探讨具体的解决方案。

关键词：元数据管理，自建数字资源，数字化，查重，互操作
DOI：10.3772/j.issn.1673—2286.2014.03.006

元数据为数字图书馆提供了精确描述数据内容、语义和服务的机制，支持浏览、传递、评估和管理信息资源，不仅是实现资源发现的关键，也是有效充分利用馆藏资源和实现互操作的基础。根据我国近5年来元数据的相关研究可以发现，目前对元数据的理论探讨远远超过其在实际工作中的应用，尤其是新的网络环境下对自建数字资源的元数据具体实践的研究还比较欠缺[1]。本文将结合国家图书馆中文图书数字化资源库元数据管理的实践，对自建数字资源元数据管理的经验和教训进行简单总结，同时也对具有一定代表性和典型性的问题作更进一步的思考，以期为其他项目提供借鉴。

1 自建数字资源元数据管理概述

1.1 自建数字资源的界定

自建数字资源，学术界目前没有一个特别明确的定义，往往表现为图书馆自建的特色数据库。索传军在其《论数字馆藏的质量评价》一文中提到，每个图书馆都有自己的特色馆藏，为了保护或进一步提高馆藏利用率，通常将它们加工成馆藏特色文献数据库。尽管目前各馆自建的数据库不多，但这是数字馆藏建设的重要途径之一[2]。目前国家图书馆数字资源建设的方式主要包括自主建设、引进建设和合作建设[3]。本文所讨论的自建数字资源，特指通过自主建设，有目的地对重点馆藏和一般馆藏进行不同深度的组织加工，结合本馆人员、技术和设备的实际情况，生产出来的数字化产品。

1.2 自建数字资源元数据管理的基本要求

自建数字资源是对馆藏资源进行数字化而形成的文献数据库，为了充分体现出特色，其元数据管理应该在遵循相关国家标准、行业标准的基础上，形成具有可扩展性、个性化、开放性的方案，在资源组织、资源利用、资源整合、长期保存等方面有其特色需求。

第一、在元数据的适应性方面，要依照相关元数据标准及著录规范来进行，在选择元数据时要考虑到各种类型资源的特征，在尊重差异和使用习惯的基础上统一，在保证通用性的前提下，满足不同类型资源的个性化需求。同时要兼顾简单性与重要性，采用的元数据不仅仅是标引内容的易于理解和掌握，便于使用，同时也要能完全揭示资源的特性和特征。要特别注意的是，对资源揭示没有意义或者过于专深的元数据要酌情采用。

第二、在资源的利用方面，为了将不同领域的资源更好地融合，当用不同格式的元数据进行资源描述、检索和利用时，就需要解决不同元数据格式间的释读、转换问题，以确保系统对用户的一致性服务，即实现元数

据的互操作[4]。互操作性主要体现在解决元数据间多元化、非标准化的协调发展问题上。在异构系统间互操作能力的支持上，缓解元数据资源描述的特殊性和一般性矛盾。

第三、在资源整合方面，考虑到资源建设的来源不同，要求元数据体系实行开放扩展机制，可以通过不同的层次纵向或横向地以规范方式进行扩展，适应资源不断发展变化的需求。开放性是指可以基于开放标准对元数据进行交换，并可在开放标准基础上在元数据集间进行元素的复用、继承和元数据集扩展[5]。

2 国家图书馆自建数字资源元数据管理的实践

国家图书馆自2000年以来对馆藏资源进行数字化加工的工作，涉及了文本、图像、音频、视频等多种类型，特色资源库包括普通中文图书、博士论文、民国文献、善本、地方志、年画、讲座音视频等，其中普通中文图书的数字化建设，建设时间最长，数量最大，积累了丰富的建设经验，具有代表性意义，可为其他项目提供较多的参考借鉴。

在馆藏资源数字化建设过程中，我们能够越来越深刻地体会到，元数据的有效管理和利用，日益成为关系到系统建设质量乃至最终数据管理状况的关键。

2.1 自建数字资源元数据管理的步骤

自建数字资源建设过程大体可分为原始资源的遴选、采集、数字化加工、验收、保存、服务、处置（主要指资源的剔除、销毁）等过程。其中，元数据相当于主动脉，支撑着整个流程的运作。整个过程需要通过元数据对数字信息进行描述，定义其元数据的核心元素，并进行结构化存取，从而实现对数字信息资源的有效管理。不同类型的元数据在信息资源生命周期的不同阶段产生并在不同环节起着重要作用，对元数据的管理也关系到整个过程能否顺利展开以及资源服务性能的优劣。

第一步：针对不同类型的资源，根据建设目标、资源加工的深度和用户需求，围绕粒度和维度、相关加工规范，制定元数据方案。

第二步：配置相应的元数据参数，定义核心元素，研究数字化信息元数据的获取方法，针对不同类型的资源，采集相应的元数据信息，并对采集过来的核心元数据进行清洗、查重和整合。

第三步：要明确不同资源格式（文本、图像、音频、视频等）所需要的不同元数据元素，根据具体需要选取目的元数据。数字化过程中，将生成结构元数据、管理元数据和保存元数据等，不同类型元数据之间的界定，以它们在信息资源生命周期中的不同功能为标准进行区分，在具体的元素设计上允许交叉和复用，但这种重复应以最小限度为前提，应通过元数据的模块化和开放型使不同的元数据可以相互连接和调用。

第四步：元数据的集中保存，分类管理。元数据生成后，需要对元数据进行封装管理和存储，确保对象数据和元数据之间关联的一致性。在数字化加工完成后，按照相关的技术标准，将元数据和对象数据进行封装后，进行长期保存，并提交发布。元数据一般统一保存在专门的系统平台上，并有专业人员对元数据及相关文档按项目类型进行整合、统一管理，保证数字化过程中的可持续管理、回溯，方便为后续工作或其他项目提供参考。

第五步：回溯管理，资源的长期保存与回溯，为文献数字化信息的可持续利用和长期保存保驾护航。

在数字资源的生命周期内，许多环节都会生成甚至修改元数据。因此对元数据的控制应是一个持续的过程，贯穿生产、使用、管理的各个阶段。

2.2 “馆藏中文图书数字化资源库”的元数据建设

作为国家总书库，中国国家图书馆拥有全球最丰富的中文文献。“馆藏中文图书数字化资源库”依托雄厚的馆藏资源，经过多年的理论和实践摸索，已经形成了较为完善的元数据建设和管理体系。

（1）元数据体系的形成

目前国家图书馆的元数据方案采用MARC格式来描述书目数据，在资源的数字化过程中，描述型元数据直接移植了传统文献的书目数据，按照特定的资源建设遴选标准，对其资源相关的书目元数据进行采集，对采集过来的元数据进行清洗、查重、分类、人工审核后，按照一定的技术手段，对MARC数据进行解析，提取相关的字段，同时文献原始载体的物理形态信息、加工信息、存储结构、版权信息等分别按照相关的加工标准，保存到对应的数据库中。

（2）MARC数据解析

对不同格式的元数据，需要根据元数据分析以及字段映射表，将这些不同格式的元数据转换成元数据应用中可以直接利用的元数据格式[6]。以MARC数据为例，由于MARC数据基于ISO 2709格式，识别必须依赖于专门的软件，信息结构上也是千差万别，无法被关系型数据库直接使用，为此，需要开发相应的MARC数据转换系统，这就需要了解原始数据，预先定义转换的对应关系。按照项目需求，提取所需的字段标识符以及子字段标识符。

简单地说，就是对非结构化的文本格式进行识别，抽取题名项、责任者项、出版项、馆藏信息项等基本信息，保存在可读取的数据库中，为资源描述、数字化加工等提供较为方便的管理控制，并综合生产过程中形成的一系列新的元数据，以便与遵循OAI-PMH的元数据格式整合，为用户提供高级检索。

（3）数据查重方案

文献数据查重是利用描述元数据的某些特定字段，将各种数据进行整理，有序组织，避免因重复加工造成数据冗余，浪费成本，从数据加工的源头就确保数据的唯一性，减少数据冗余度，从而提升馆藏数字资源建设的整体水平。

首先应针对资源的种类确定查重的原则方法，选择查重条件，但在实际操作中通过题名、责任者、ISBN、出版项等单一条件往往很难满足查询的要求，通常采用多条件组合查询、模糊匹配等方式对元数据进行操作，一般建议采用程序控制和人工校对相结合的方法。在实际工作中，可选的检索点很多，可以说每个著录项都可以作为检索点进行查重，采取灵活机动的检索策略，不同的情况采用不同的检索查重方式，并灵活组合运用，才能避免漏查、误检。

（4）元数据的构成

在数字化过程中，根据加工规范，会形成一系列的数据库表，主要包括：

- 用来描述、识别文献的基本信息表，包含原始载体书目的基本信息，如题名、著者、出版信息、馆藏信息、唯一标识号等；
- 用来记录目次标引的信息表，用于记录册次号、目次、原始页码、标引链接页码等，反映文献结构内部形式特征，同时满足目录检索需求；
- 用来记录原始文献结构的信息表，包括封底、封面、前附页、后附页、目录页等信息；
- 用来记录数字化加工的信息表，包括扫描分辨率、压缩因子、图像参数、存储量、保存位置等信息；
- 除此之外，还有分别反映版权信息、原始载体的缺页插页等物理特征的数据库表。

这些数据库表和书目数据共同构成了整个数字化资源的元数据的完整体系。可以看出，描述型元数据、结构型元数据、管理型元数据和保存型元数据等几种类型的元数据在具体的实践中存在一定程度的交叉和

图1 馆藏中文图书资源库的发布界面

图2 从描述元数据提取的元素以及目录页的定位链接

复用，但这个重复的程度要控制在最小范围内。

在数字化工作完成之后，按照相应的技术标准对元数据和图像进行封装以提交网络发布和长期保存，元数据同时也会导入专门的元数据管理平台，便于管理和回溯。

从馆藏中文图书数字化资源库的检索、保存、组织等方面来看，基本满足了初定的特色要求。

图1为馆藏中文图书资源库的发布界面，图2为从描述元数据提取的元素（可作检索用）以及目录页的定位链接。

3 自建数字资源元数据管理的思考和建议

3.1 健全的元数据规范和评价体系的建设

目前国内数字图书馆建设中对自建资源的元数据标准和评价体系的研究较少，同时，对于除了描述元数据之外的其他类型元数据等的研究与应用也比较少。单就管理元数据和技术元数据来说，其规范及评价体系也是必不可少的，将直接影响着整个数据库系统的组织与管理能力[7]。元数据的评价，其核心在于体现元数据使用的“效用”，要满足特定用户群特定的与潜在的需求[8]。研究和移植已有的国内外成熟的元数据标准体系，制定并完善其元数据的使用规范，使规范具有可操作性和指导意义。

3.2 元数据与对象数据的关联方式的选择

元数据与对象数据的关联至关重要，失去与对象数据的关联，元数据将变得毫无意义，而丢失元数据的对象数据，其价值也将大打折扣。为避免在数据的备份、迁移过程中出现偏差，元数据可以嵌入在对象数据中，也可以存放在结构化数据库中，通过系统与对象数据关联，还可以通过封装的方法与对象数据打包在一起。其中，选择适合的封装策略既能保障对象数据与其元数据间紧密的联系，同时又保证了二者的各自独立性，利于它们的个性化、动态管理和利用，是目前国内外比较认同的关联方式[9]。

元数据编码和传输格式METS（Metadata Encoding and Transmission Standard）由于具备支持资源的互操作、适应多样化的应用环境和有着长远的应用前景等优势，成为目前国际领域影响最大、使用最为广泛的数字资源元数据封装方法[10]。

3.3 元数据互操作的实现

考虑不同类型资源，或者与其他合作建设的馆外资源的进行融合时，可能存在着多种不同的元数据方案，在编码、格式、内容等方面存在较大差异，要想整合这些资源，利用这些资源，便于管理，通过统一接口获取各类信息资源，保证向用户提供一致的服务，更易于系统的开发与用户的使用，就要解决不同格式元数据之间的相互转换、相互融合的问题，即元数据互操作。目前实现元数据的互操作有三个技术途径：一是采用字段映射和对照。如DC与MARC、DC与EAD等。二是需要借助重用、集成等方式，实现各个项目的元数据记录间的整合。三是通过协议、聚合和值共现映射等开展仓储级的元数据互操作[11]。最常采用的是字段映射和对照方式，即在两个元数据标准的元素之间直接转换，建立元数据字段映射关系表，与具体的资源或项目相结合，设计适合的编目及数据转换平台，提供统一的编目和审校环境。

3.4 完善的元数据管理平台的构建

建立元数据管理平台，其功能应包含元数据的维护及查询、元数据整合、元数据批量修改、元数据的分析及应用、元数据版本管理等，建议加强对技术人员和业务人员的相关技术培训。在整个元数据应用过程中，建立规范的工作流程和管理规则，以利于工作人员统一地对元数据进行管理和监督以及探查，全面提升数据质量。

参考文献

[1] 叶静.从2006 --2011年我国核心期刊载文分析看我国元数据研究新进展[J].科技情报开发与经济, 2012(14):126-128.

[2] 索传军.论数字馆藏的质量评价[J].中国图书馆学报, 2004(4):43-46.

[3] 全国数字图书馆建设与服务联席会议.数字图书馆资源建设指南[EB/OL]. [2013-11-21]. www.lsc.org.cn/Attachment/Doc/1275990326.pdf.

[4] 毕强,朱亚玲.元数据的标准及其互操作研究[J].情报理论与实践, 2007(5):666-670.

[5] 周波.高校科学数据元数据初探[J].图书馆学研究, 2012(1):45-53.

[6] 肖珑,申晓娟.国家图书馆元数据应用总则规范汇编[M].北京:国家图书馆出版社,2011.

[7] 董蓓.DC元数据在专题特色数据库建设中的应用[J].图书馆工作与研究,2010(4): 42-44.

[8] 陈学清,陈成桂,杜芸,等.网络信息资源编目元数据的选择与评价[J].图书馆工作与研究,2008(7):65-68.

[9] 程妍妍.国际电子文件元数据封装方法VEO和METS的比较研究[J].现代图书情报技术,2011(10):7-11.

[10] 程妍妍.基于METS的电子文件元数据封装研究[J].科技档案,2011(4):19-24.

[11] 宋琳琳,李海涛. 大型文献数字化项目元数据互操作调查与启示[J]. 中国图书馆学报, 2012(01): 27-38.

作者简介

吴茗（1975-），女，馆员，硕士研究生，国家图书馆，发表论文数篇。E-mail: wum@nlc.gov.cn
龙伟（1966-），女，副研究馆员，国家图书馆，发表论文数篇。
肖红（1982-），女，馆员，硕士研究生，国家图书馆，发表论文数篇。

The Practice of the Metadata Management of Self-developed Digital Resources ——Taking the Chinese Books Digital Resources Digital Resources of NLC as an Example

Wu Ming, Long Wei, Xiao Hong / National Library of China, Beijing, 100081

Abstract: Based on the practical metadata management of Chinese books, this paper gives the basic principles of self-developed digital resources, and analyzes the following four issues of metadata: specification and evaluation system construction, approach selection of objects associative, metadata interoperability, and establishment of management platforms.

Keywords: Metadata management, Self- developed digital resources, Digitalization, Duplicate checking, Interoperability

（收稿日期：2013-11-26）