大型文献数字化项目元数据互操作调查与启示*

宋琳琳　李海涛

摘　要　大型文献数字化项目元数据标准的多样性与信息资源需求接口单一性之间的冲突，使得元数据互操作势在必行。通过文献调研、比较研究、案例分析等方法对大学数字图书馆国际合作计划、谷歌图书、欧洲数字图书馆、HaithTrust、美国记忆等八个国内外知名的大型文献数字化项目的元数据互操作情况进行分析，发现映射、集成、协议和 API 较为常用，注册、转换和关联数据的应用范围相对较小。建议我国大型数字化项目的元数据互操作向细粒度和去格式化的方向努力。图 4。表 5。参考文献 25。

关键词　大型文献数字化项目　元数据标准　互操作

分类号　G254.36

Metadata Interoperability in Mass Digitization Project: A Survey and Suggestions

Song Linlin & Li Haitao

ABSTRACT　The conflicts between diverse metadata and information resources make metadata interoperability of mass digitization project a prerequisite. The study investigates 8 mass digitization projects, such as CADAL, Google Book, Europeana, Haithtrust, American Memory, and so on. It finds that mapping, integration, OAI and API are the most used methods for improving metadata interoperability; methods such as metadata registry, conversion, and linked data also play an important role. For metadata interoperability in China's mass digitization projects, we suggest that great efforts should be made in refined description elements and removed metadata format. 4 figs. 5 tabs. 25 refs.

KEY WORDS　Mass digitization projects.　Metadata standard.　Interoperability.

大型文献数字化项目是指大型机构或是多个机构合作开展以创建数字信息资源、提供数字信息服务为目的，通过扫描、拍照等转换技术，将传统的非数字型资源转换成计算机可以读取和识别的数字资源的工作。在大型文献数字化项目中，信息组织的主体主要有政府机构、文化遗产保护机构、非营利机构和 IT 公司；信息组织对象的来源、类型、格式多样，主要包括图书、期刊、报纸、手稿、地图、书籍、乐谱、录音资料、电影、印刷制品、照片和建筑制图等珍贵文献。主客体的多样性必然导致大型文献数字化项目信息组织过程中多种元数据标准并存。这虽然满足了不同资源、不同领域、不同系统及其应用的需要，却为分布式信息环境下的信息资源整合利用带来诸多问题和挑战：一方面不同类型的信息资源倾向于采用特定的元数据方案，另一方面用户又希望通过统一接口获取各类信息资源以满足其需求。元数据的多样性与信息资源需求接口单一性之间的冲突，使得大型文献数字化项目的元数据互操作势在必行。

* 本文系国家社科基金项目“社会化网络环境下信息组织的理论与方法创新研究”（项目编号：10BTQ023）及中山大学青年教师培育计划“大型文献数字化项目的信息组织研究”（项目编号：20000-3161107）的研究成果之一。

通讯作者：宋琳琳，Email：songlinl@mail.sysu.edu.cn

1 文献综述

笔者对2002以来有关“元数据互操作”的文献进行研读分析。就地域而言,国内外学者的研究方式存在很大不同:国内学者多从理论层面进行探讨,归纳总结元数据互操作的方案、技术,国外学者则多结合具体资源或项目,提出具体的元数据互操作实现途径。就内容而言,国内外学者主要关注以下方面:①元数据互操作的要求和目标。Getaneh等认为元数据互操作正朝着社会化和简洁化的方向发展,为此他们构建了一个概念框架,力求深化元数据描述程度、简化元数据格式与结构,并吸纳用户驱动的描述元素,即来自社会化网络环境中的用户参与,从而提高元数据互操作的语义性[1]。Jung-ran Park认为元数据标准的概念歧义和语义重叠是阻碍元数据语义互操作的关键,为提升互操作效果应重点解决上述两个问题[2]。②元数据互操作层次及方式。张晓林将元数据互操作框架划分为数据内容、编码规则、元素语义、元素结构、标记格式、交换格式、通信协议七个层面[3];曾蕾按照实现互操作性的水平把元数据互操作分为模式级、记录级、仓储级三个级别[4];毕强等将元数据互操作划分为语义互操作、语法互操作、协议互操作三个层面[5]。Philip Hider通过对澳大利亚数字馆藏保存机构的元数据标准使用情况以及互操作情况的问卷调查,发现互操作范围主要局限于机构内部,而互操作方式则为尽量保证元数据标准的一致性,并采用新技术解决不同格式之间的差异[6];Lina建立了档案元数据EAD与书目元数据MODS之间的对照[7];Bountouri提出通过建立应用程序和本体两种方式开展元数据互操作[8];Shirley Lim对新西兰GLAM的数字图片的元数据质量与互操作进行调查,认为自建元数据标准以及重用原始描述数据会增加互操作的难度[9]。③元数据互操作质量评估。Weagley利用完整性、准确性、一致性和受控词汇使用情况4个指标来评价元数据互操作的质量[10]。

通过文献综述,笔者得到以下启示:元数据互操作方式多样,不再局限于映射、对照等传统方式,应综合应用多种方式;此外新技术可以解决互操作的语义歧义、模糊等问题,并细化描述粒度。其次,应将元数据互操作的研究与具体资源或项目相结合,才能解决现实中存在的问题,发挥其对实践的指导作用。

2 调查对象与内容

2.1 调查对象

结合大型文献数字化项目参建主体的国别、性质与类型,本文选择了大学数字图书馆国际合作计划(China Academic Digital Associative Library,CADAL)、中国国家图书馆数字图书馆工程(数图工程)、谷歌图书(Google Book)、欧洲数字图书馆(Europeana)、开放图书馆(Open Library)、HaithTrust、加州数字图书馆(California Digital Library,CDL)、美国记忆(American Memory,AM)共八个国内外知名的大型文献数字化项目作为调查对象,通过登录各项目网站及查阅与其信息资源建设相关的文献,实地调研CADAL项目并对其工作人员就该项目信息组织中采用的标准、技术和具体工作流程等问题进行访谈,进行分析研究。

2.2 调查内容

元数据互操作是指不同元数据格式间的信息共享、转换和跨系统检索等相关问题,为用户提供一个统一的检索界面,确保系统对用户的一致性服务。元数据互操作框架和层次是本次调查的参考依据,其中曾蕾的三级互操作框架,按时间顺序涵盖了从元数据标准构建、元数据记录产生到检索应用的数字资源建设全过程,又兼顾了信息资源描述的不同深度,如元素、记录、框架模式等;其归纳的每个级别的具体实现方式可以基本呈现目前元数据互操作的发展现状。参考上述内容,本文根据调查情况将其细化为七种主要方式,具体应用情况如表1所示。

调查显示,目前大型文献数字化项目所采用的互操作方式不再局限于映射、对照等传统方式,而是多种方式综合应用。

表 1　大型文献数字化项目元数据互操作方式调查结果

方式 / 项目	模式级		记录级		仓储级		
	映射	注册	转换	复用与集成	协议	API	关联数据
CADAL	√				√	√	
数图工程	√			√	√		
AM	√	√	√	√	√		
Google Book				√		√	
Europeana	√	√	√	√	√	√	√
Open Library	√			√	√	√	
HaithTrust	√			√	√	√	
CDL	√			√	√	√	

3　调查结果分析

3.1　常用元数据互操作方式

3.1.1　映射

元数据模式级互操作通常发生在数据记录被创造出来之前，是对现有元数据的派生与修改；相较于其他互操作方式，元数据映射在项目创建的初始阶段应用，可从根本上提高互操作的范围，因此被大型文献数字化项目广泛采用。元数据映射又称元数据对照，是指两个元数据标准的元素之间直接转换，通过一对一、一对多、多对一及多对多等多种映射方式，解决语义互换及统一检索问题。

通过调查发现，八个知名大型文献数字化项目在数字对象描述方面均结合实际需求，自建了新的元数据标准；其中 87.5%①的调查对象采用了元数据映射，但由于各个项目的需求和发展目标不一致，所映射的对象也存在差异，并呈现以下特点（表 2）：

（1）MARC、DC 作为通用的元数据标准是最常用的映射目标。与 MARC 映射主要是为方便与原始文献建立关系，如 CADAL、数图工程、Europeana 等；与 DC 映射主要是为了满足数字资源采集、检索和使用的需求，如 CDL、Open Library 等（见表 3）。

表 2　大型文献数字化项目自建元数据标准的映射对象

项目名称	描述方式	映射对象
CADAL[11]	自建元数据标准	MARC
数图工程[12]	自建元数据标准	DC CNMARC MARC21
Europeana[13]	自建元数据标准	EAD MARC
Open Library[14]	自建元数据标准	DC
HaithTrust[15]	自建元数据标准	PREMIS
CDL[16]	自建元数据标准	DC
AM	自建元数据标准	DC ONIX FGDC UNIMARC GILS

① 某种元数据互操作方式应用率 = 调查对象中已采用该方式的项目数量/所有调查对象的总量；87.5% 的调查对象即意味着有七个项目采用元数据映射的方式开展互操作。

表 3 Open Library 元数据标准与 DC 元数据标准之间的映射

OL metadata	DC	OL metadata	DC
author	dcterms:creator	subject_place	dcterms:coverage
contributions	dcterms:contributor	subject_time	dcterms:coverage
title	dcterms:title	genre	dcterms:type
subtitle	dcterms:title	language	dcterms:language
by_statement		description	dcterms:description
physical_format	dcterms:format	table_of_contents	dcterms:tableOfContents
other_titles	dcterms:alternative	notes	dcterms:description
work_title	dcterms:alternative	LC_classification	dcterms:subject
edition		ISBN	dcterms:identifier
publisher	dcterms:publisher	LCCN	dcterms:identifier
publish_place	dcterms:publisher	URL	
pagination	dcterms:extent	source_record_loc	
number_of_pages	dcterms:extent	source_record_id	
DDC	dcterms:subject	publish_date	dcterms:date
subject	dcterms:subject	publish_country	

(2)特定功能需求,大型文献数字化项目肩负着促进资源利用、推动数字资源长期保存等多项使命,因此也应兼顾元数据互操作。HaithTrust 为了实现其对数字资源长期保存的功能,将其自建的元数据标准与保存元数据标准(Preservation Metadata Implementation Strategies, PREMIS)进行映射。

3.1.2 数据复用与集成

数据复用与集成属于元数据记录级互操作,主要发生在元数据记录生成之后。很多大型文献数字化项目在建设前并没有发现相似资源的存在,或是没有考虑互操作问题,因此在项目建设过程中,元数据记录已经产生,映射等模式级互操作方式无法有效满足已赋值的元数据互操作需求,这就需要借助重用、集成等方式,实现各个项目的元数据记录间的整合。

复用与集成方式遵循元数据组织模块化原则,一条元数据记录的各个组成部分可以被当作不同的独立单元,按需要将不同元数据源的这些单元组合在一起或重新应用,都将产生新的元数据记录。该方式覆盖面广,不仅涵盖各种标准、词表、应用规范,还包括了来自不同项目的元数据记录。

调查结果显示,87.5% 的大型文献数字化项目采用了复用与集成的方式,但是具体的实现途径各有不同,以 METS 和 RDF 两种方式为主,有的项目多种方式并用(见表 4)。METS 通过将不同元数据源的各个组成部分统一封装到 XML 文件中,实现与外部元数据格式的结合。而 RDF 则通过定义 RDF Schema,利用 XML Namespace 调用已有定义规范的机制,从而直接在 RDF 中引用不同的元数据标准来实现数据复用与集成。两种方法虽然操作原理不同,其最终结果都可以实现不同来源的元数据记录的复用与集成,这也和林海青[17]、Marcia Lei Zeng(曾蕾)与 Lois Mai Chan[18]的研究成果相互印证。

表 4　大型文献数字化项目元数据互操作复用与集成方式实现途径

项目 / 方式	数图工程	AM	Google Book	Europeana	Open Library	HaithTrust	CDL
METS		√		√		√	√
RDF	√			√			
其他方式			√		√		

(1) METS

大型文献数字化项目在数字化加工过程中，实现了文献载体和格式的转变，而以数字形式存在的资源所需元数据元素和原始文献存在区别，因此很多机构就会重新建立新的标准，从而忽视了对原始文献元数据的复用与集成，这必然会造成资源的浪费。METS (Metadate encoding and Transmission Standard，元数据编码及转换标准)的优势在于构建了由不同模块(描述、管理、结构等)组合而成的统一框架，该框架不受模式、词汇、应用程序等限制，按照属性将不同来源的多条记录分别整合到相应模块中，然后统一封装成 XML 文件，从而形成一条新的记录。

CDL 将 METS 作为其建立数字资源仓储和提供服务的基础，其数字对象指南(CDL Guidelines for Digital Objects)认为 METS 可以共享数字对象的共同特征，如内容文件格式、元数据编码标准，并且还包括足够的细节，以使 METS 的创造者和加工人员在创建和处理 METS 编码的数字对象时符合特定的配置文件[16]。在对数字对象进行描述的过程中，CDL 为节约项目建设成本并重用已有资源，充分利用来自原始文献的书目数据以及合作方的描述元数据，并整合到 METS 框架中，从而作为最终的描述元数据记录。

(2) RDF

相比于 METS 的模块化整合策略，RDF(Resource Description Framework，资源描述框架)将 METS 框架中的元数据包打散成单个元数据元素，着眼于具体元素的描述记录，通过调用和引用的方式实现集成。RDF 应用于元数据数据复用取决于两个方面：首先是命名空间(Namespace)，元数据元素的产生与归属由命名空间定义，所以命名空间是辨识元素来源、理解元素语法语义特征的主要依据。其次是 RDF 规范(Schema)，规定了利用 XML namespace 方法调用已有定义规范的机制，从而可以直接在 RDF 中引用各种元数据定义。

Europeana 的数据模型 Europeana Data Model (EDM)基于多种目的综合应用多种元数据标准，如将 OAI ORE 用于不同数字对象及其衍生形式组织管理，将 DC 用于描述，将 SKOS 用于概念词汇的选择与表述(见图 1)。应用 RDF 可以灵活调用上述元数据标准中的可用元素，不仅可以将不同的元数据标准集成与复用，而且可以保存原始数据并支持互操作[19]。

此外，为了适应语义网的发展，Europeana 将其数字资源的相关数据都采用 OpenLink Virtuoso 或 4Store 等 RDF 存储方式，其目的是为了方便在语义环境中，Europeana 的元数据可以通过关联数据有效揭示，提高资源可用性[20]。

(3) 其他方式

有的项目并没有利用现有的框架或规则，而是根据项目建设的需要，采用其他方式集成元数据。以 Google Book 为例，其拥有超过 100 个书目数据的来源，包括图书馆、出版商、零售商，以及评论和图书封面的聚集者，收集了 8 亿条数据记录，包含超过 1 万亿个数据字段。Google Book 将这些获得的记录转化为简单的数据结构，但没有添加 URI，并通过几种不同的方式转化为 Google 可以利用的状态。初始的元数据结构储存在一个类似于 SQL 的数据库中，用于简单检索。其大体流程为：首先利用解析算法进行处理，从中抽取特殊意义的信息如记录号、条形码等；再利用聚类算法、文本相似度匹配等技术进行综合，择优筛选并最终形成一条元数据记录[21]。

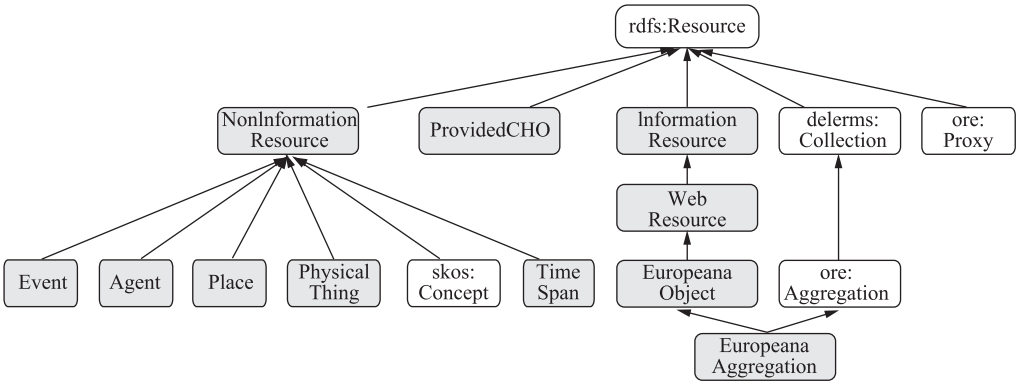

图 1 EDM 的 RDF 构成及来源

3.1.3 协议

大型文献数字化项目通常由多个机构合作完成,数字化成果多为分布式存储,且存在由于规划导致的异构状态。当要实现上述分布式异构资源的集成检索时,项目参加方面临的一个重要问题是检索结果很少以一致、系统和可信赖的格式出现。究其原因还是因为各个检索结果来自采用不同元数据标准的数据库,或是采用一些特殊的描述规则。解决这个问题还需通过协议、聚合和值共现映射等开展仓储级的元数据互操作。

支持元数据互操作的协议有很多种,如 OAI - PMH、Z39.50、ZING(SRU/SRW)等,大型文献数字化项目根据建设的需求选择应用。Europeana 同时支持 OAI - PMH、Z39.50、SRU 三种协议,其统计显示[22]:27% 的馆藏支持 Z39.50 协议,3% 的馆藏支持 SRU 协议;由于有的馆藏支持多种协议,所以超过 70% 的馆藏支持 OAI - PMH 协议。Haithtrust 则参考数据来源选择元数据互操作协议,采用 Z39.50 协议采集 OCLC 和密歇根大学图书馆提供的元数据记录,其他机构多采用 OAI 协议进行采集。

OAI 协议在八个被调查的大型文献数字化项目使用率达 87.5%,而且功能完善。除 Google Book 项目使用 API 调用与解析外,其他项目都支持 OAI 协议。AM 利用 OAI 协议不仅实现了参建机构的元数据互操作及采集整合,并以此为途径将其资源开放给其他相关项目使用。首先,AM 将来源于 AM、全球门户(Global Gateway)、印刷品及图片部在线目录、历史新闻数据库(Chronicling America)和其他参建机构的元数据进行采集整合;然后,针对不同类型文献的元数据分别进行聚类,大致归为图书、手册、地图、海报、影片、音频、期刊等十类;再根据各类别的具体情况,分层次提供基于 OAI 协议的开放采集,其中照片类的元数据集最多,共有 27 个数据库可供采集,而大多数类别中仅有一个数据库开放。同时,为了满足用户对于不同格式元数据的需求,AM 共提供了 DC、MARC XML 和 MODS 三种格式的元数据方便采集,并且还提供了一些遵守 OAI 协议的、预先编制的元数据采集请求编码供有需要的机构参考(见图 2)。

3.1.4 API

作为互联网上开放应用程序接口,API 的功能就是将系统原有的登陆、数据查询、浏览、更新等操作及操作参数和功能调用,按照某种发布协议进行封装,外部程序就是按照封装后的调用方式通过 API 实现与系统的数据交互。协议是 API 实施的基础,它定义了交互数据的方法和格式,常用的发布封装协议有 REST、SOAP、XML - RPC。除此之外,交互数据格式、交互数据内容设置和使用授权等都是 API 应用的关键技术。

75% 的调查对象都通过调用 API 实现元数据记录的互操作。Google 为了提高信息资源描述的质量,获取了超过 100 家书目信息供应方的

```
<?xml version="1.0" encoding="UTF-8" ?>
-<OAI-PMHxmlns="http://www.openarchives.org/OAI/2.0/"      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2012-05-27T13:06:14Z</responseDate>
  <request                verb="GetRecord"                identifier="oai:lcoa1.loc.gov:loc.gmd/g3791p.rr002300"
metadataPrefix="oai_dc">http://memory.loc.gov/cgi-bin/oai2_0</request>
- <GetRecord>
- <record>
- <header>
  <identifier>oai:lcoa1.loc.gov:loc.gmd/g3791p.rr002300</identifier>
  <datestamp>2005-11-21T17:08:59Z</datestamp>
  <setSpec>gmd</setSpec>
  </header>
- <metadata>
-<oai_dc:dc                                          xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:dc="http://purl.org/dc/elements/1.1/"              xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  <dc:title>New railroad map of the state of Maryland, Delaware, and the District of Columbia. Compiled and drawn
by Frank Arnold Gray.</dc:title>
  <dc:creator>Gray, Frank Arnold.</dc:creator>
  <dc:subject>Railroads—Middle Atlantic States—Maps.</dc:subject>
  <dc:description>Description derived from published bibliography.</dc:description>
  <dc:publisher>Philadelphia</dc:publisher>
 ……
  </oai_dc:dc>
  </metadata>
  </record>
```

图 2　AM 提供的基于简单 DC 格式的 OAI 请求内容(节选)

书目记录,其中很多就是通过调用相关机构的 API 实现的,如通过调用 Worldcat 的 API,Google 可以获取 Worldcat 中几乎全部的书目信息。CDL 通过 API 调用整合了两个项目的资源,即 PubMed 和 NSDL 资助的地球科学门户。此外,Google、Open Library、Haithtrust 还提供 API,方便其他项目的调用。

API 的调用过程主要包括检索匹配、调用、转换格式、加载等环节。以 Open Library 为例,首先要解决的是如何准确获取信息,通过 GET 方式请求将用户的检索词与 API 中提供的元数据进行检索标识的匹配,如 ISBN、OCLC 标识符、LCCNs 号和 OLIDs(Open Library 内部的标识符)。如果 API 中保存相关信息,就将其进行调用。如果只是调用一个 API,就可以将符合要求的数据进行分析整合,通过修改 Html DOM 的方式直接实现客户端浏览器页面的更改;如果同时调用多个数据库,则需要将返回的结果进行综合处理,通常采用将调用返回的多个 XML 数据混合到一个 XML 文档中,并且使用 XSLT 将这个文档转化为一段 XHTML 代码,然后把这段 XHTML 代码加载到相关网页中,从而实现了架构于元数据互操作基础上的资源与服务集成。

利用 API 开展元数据互操作的优势在于:API 是对操作及操作参数和功能调用的封装,与内容无关;而服务提供方通过调用 API 进行解析和链接而获取资源与服务,不必再根据内容的变化而不停维护资源链接,从而大大降低了工作负担。

3.2　其他元数据互操作方式

除了 3.1 总结的大型文献数字化项目常用的元数据互操作方式外,还存在其他的互操作方式,如注册、转换、关联数据等;此次调查中,虽然这些方式的应用率不高,但在元数据互操作方面的作用却不能忽视。

3.2.1　注册

元数据注册系统(Metadata Schema Registry,MR)是对元数据定义及其编码、转换、应用等规范进行发布、注册、管理和检索的系统,以支持开放环境中元数据的发现、识别、调用,以及在此基础上的转换、挖掘和复用。它根据统一的标准模型(ISO/IEC 11197)进行语义、编码、标准解析和转换,按照领域或者主题建立元数据规

范目录列表,映射到各自所对应的物理信息资源,并以 Web 服务的形式在网络上进行发布,通过元数据从语义层面的关联和协同可以有效地进行信息资源的整合,支持智能检索、定题服务、主题聚类、内容挖掘等知识服务,从而实现信息资源的开发和增值。

Europeana 的元数据注册系统 EuMDR(Europeana Metadata Registry)是用来管理和发布该项目参建方所使用的元数据标准和具体元素,并实现不同标准之间的转换。除了实现上述功能外,EuMDR 还将作为系统的组成部分被整合到 Europeana 中,为其他服务提供支持,如 REPOX 可以利用 EuMDR 将原始元数据转换成 ESE 或 EDM 等不同的元数据框架,方便用户查询[23]。

3.2.2 转换

转换主要指已生成的元数据记录从一种格式转为另一种格式。在转换过程中,最重要的问题就是将面临数据失真或丢失。如果转换过程中包含了数据值,当目标格式比源格式包含更多细节元素时,就必须将源元数据记录分解为更细小的单元,如从 DC 到 MARC 的转换,会导致数据失真;反之则会造成数据丢失。此外,如果元数据的取值需要参考受控词汇,也会让转换变得更加复杂;正因为上述情况的存在,在开展元数据标准转换时应该制定相应的操作指南予以辅助。

很多大型文献数字化项目的参建机构很早便开展了本机构的数字化建设,而根据发展的需要参加不同的项目。如密歇根大学图书馆不仅加入了 Google 的图书搜索计划以加快数字化进程,节省图书馆经费,还加入了 Haithtrust 项目以保障信息资源的有效利用。针对这部分参建之前已完成数字化的文献的整合利用,转换不失为一种简便的途径。以 Europeana 为例,其文献资源来自欧洲各国的国家图书馆及相关机构,在建立元数据记录时,通常采用转换的方式,并在 Europeana 的元数据记录的"来源"要素中注明转换过程和采用的转换标准(见图 3)。

```
Title: Welsh Wesleyan Ministers
Data provider: The National Library of Wales
Source: Converted from MARCXML to MODS version 3.4 using MARC21slim2MODS3-4.xsl
```

图 3 Europeana 中 *Welsh Wesleyan Ministers* 的元数据记录(摘录)

通过对该文献转换前后的两条元数据记录对比来看,绝大部分内容可以实现有效转换及使用,但也存在一些差异(见表 5):一是需要经过字段或元素合并才能有效转换;二是部分元素并非完全对应,如 Europeana 的 Type 包含数字图片的一些数据,而 300 字段仅限于对纸质图片的揭示;三是部分内容无法转换,需要再进行人工添加。这也印证了元数据记录的转换确实会造成"数据失真或丢失",因此也制约了其在大型文献数字化项目中的应用。

3.2.3 关联数据

关联数据作为社会网络环境中信息资源整合的新技术,虽然还处于研究阶段,但其超越了格式框架的限制,代表未来网络资源应用发展的方向。

关联数据采用 RDF 数据模型,利用统一资源标识符(URI)命名数据实体来发布和部署实例数据,从而可以通过 HTTP 协议揭示并获取这些数据,同时强调数据的相互关联以及有益于人机理解的语境信息[24]。

关联数据之所以可以用于元数据的互操作,是因为它打破了传统元数据的存在形式,不再局限于某一模式或应用框架中,而是将元数据标准中的每一个元素都用 RDF 三元组的方式进行描述,然后发布在网络中并部署实例数据,利用元素之间的关联,通过 HTTP 协议进行整合。在这种情况下,要进行信息资源整合或是集成检索,就不必再考虑不同元数据标准之间的差异,只需要将符合检索要求的 RDF 三元组进行集成就可获取想要的信息资源。假设图书

表 5　*Welsh Wesleyan Ministers* 的元数据记录转换对照

The National Library of Wales (MARC 21)	Europeana (MODS)
242 245	Title
100	Creator
260	Publisher Publication date
300	Format Type
500 520 533 583	Description
600 650 655	Subject
773	
	Relation; Source; Data provider; Identifier; Language

The Organization of Information 在 CDL 中有书目信息，在 Google Book Search 中有用户对其评论。通过元数据采集或转换，很难采集到 Google 中的用户评价；但是如果采用了关联数据，关于该书的每条信息都可以用 RDF 表示，这样即使是不同的来源，即使其内容为元数据标准规定的非核心元素，只要它们有一个匹配点（见图 4，ISBN 号匹配），便可以实现两个数据集的关联，从而实现元数据互操作的最终目的。

Europeana 在 data. europeana. eu 中已发布了关于文本、图片、视频、音频等类型的 240 万条开放数据，这些数据来自欧洲 15 个国家的 200 多个文化机构，其中很多数据来自于其他机构提供的开放关联数据，如法国国家图书馆、瑞典国家图书馆等。Europeana 收集的数据范围广泛，不仅包括传统的书目记录，还有维基百科描述、用户标签与评论、社会网络活动及用户照片及视频等。经过整理后再以关联数据发布的数据，不仅内容得到丰富，可关联的范围也得到拓展。Europeana 开展关联数据项目的主要目的是帮助 Europeana 成为欧洲文化信息的权威来源及欧洲文化遗产宣传推广的重要渠道[25]。

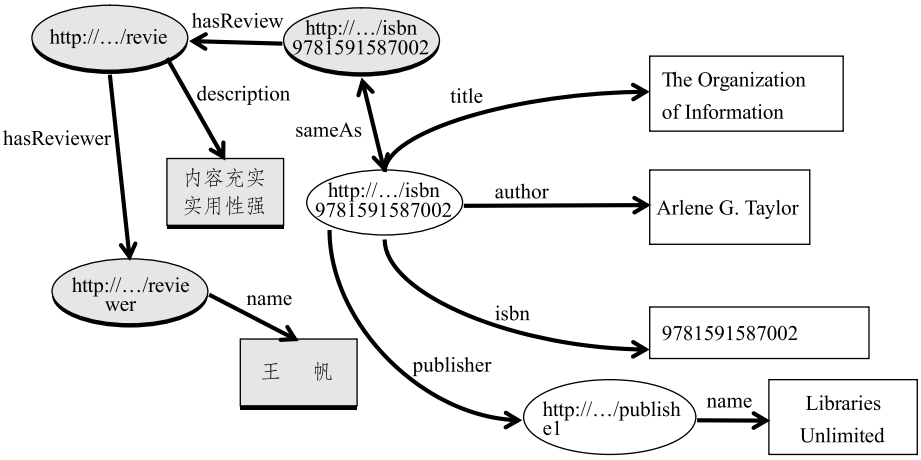

图 4　应用关联数据的元数据互操作

4 对我国大型文献数字化项目元数据互操作的启示与建议

4.1 启示

通过上述调查,笔者认为目前大型文献数字化项目元数据互操作正朝着细粒度、去格式化的方向发展。

(1)细粒度

所谓细粒度,是指随着元数据描述和揭示程度的不断加深,强调对数字对象具体特征的全面描述和各元素内在关系的深度揭示,传统的基于框架模式的转换将逐渐减少,而更倾向于面向具体元素的关联。这就意味着在今后元数据互操作实践中,不仅应该关注各项目采用的元数据标准和框架,更重要的是看其对具体元素的描述方式和开放程度,这才是提高其互操作性的基本保障。

(2)去格式化

所谓去格式化,与细粒度相辅相成,在对数字对象描述深入细致的要求下,元数据互操作正在经历由形式到内容的转变,即将对格式一致、兼容、转换的关注逐渐转移到元数据元素描述的可读性与规范化。在统一的描述语言和规则中,即使其描述对象、描述深度、取值的受控词汇、来源框架差异显著,但由于其互操作对象是由关于数字对象的一条完整格式记录转换为描述该数字对象某一具体特征的内容元素,互操作中格式障碍就将降到最低。

4.2 建议

为了适应上述发展趋势,我国大型文献数字化项目元数据互操作应从以下方面着手改进。

(1)将元数据互操作纳入项目建设规划

元数据互操作是信息资源整合的基础,大型文献数字化项目通常由多个机构合作建设,而信息资源整合是项目发展中必须面临的问题。通过上述调查发现,在项目建设初期,开展元数据互操作的方式越多样,工作量和工作难度相对越小。所以,在我国大型文献数字化项目建设过程中,应在项目规划阶段就考虑元数据互操作问题。首先应调查了解相似项目和相关资源,选择或构建一种适合当前资源环境的元数据标准描述方式,既要保证尽可能详尽地描述项目的数字资源,又要与其他通用标准及规则兼容,还可以构建元数据注册系统,吸纳其他相关标准及元素,从而避免项目建成后再去弥补与修正。

(2)推进新技术在元数据互操作中的应用

调查显示,我国的大型文献数字化项目采用的元数据互操作方式较为单一,主要集中在元数据映射、集成和协议等方面,这种情况会直接影响互操作的广度与深度,所以丰富元数据互操作方式是目前我国大型文献数字化项目互操作建设中面临的主要问题。新技术的应用尤其重要。如关联数据在 Europeana 的应用,可以细化信息资源描述的粒度,元数据标准中的每一个元素都可以 RDF 的形式表示实例数据,这样不但可以打破元数据模式和应用框架的局限,提高资源被发现的机率;同时还可以将存在于网络中的各类相关资源进行链接,从而扩大可用资源的范围。

(3)构建逻辑模型

元数据互操作的逻辑模型是指关注元数据内容而非元数据格式的互操作方式,这就要求打破目前元数据互操作的格式界限。就大型文献数字化项目而言,所调查的八个项目均自建了元数据标准,相应的就有八种不同格式,所以元数据的格式无法穷尽,实现不同格式之间的互操作工作量过大,而且范围过于宽泛。如果构建元数据互操作的逻辑模型,则互操作的关注点就从形式转移到内容,强调具体元素的语义描述和规则统一。如利用 XML/RDF 描述每一个元素,由于 XML 是网络环境中普遍应用的结构化语言,在同一种语言下进行转换障碍较少,也便于搜索引擎的理解与反馈,从而可有效实现项目资源与网络资源的整合,以及与外部系统的跨系统、跨平台应用。

(4)构建统一的元数据数据模型

首先,大型文献数字化项目的参建机构性质多样,所采取的元数据标准存在很大差异,如

商业公司以检索为目的，其元数据标准相对简洁；而文化遗产保护机构基于长期保存的目的，元数据著录详实丰富。即使机构的性质相同，由于自身固有的一些特点，所采用的标准也不尽相同，图书馆通常采用 MARC，档案馆采用 EAD，而有的博物馆则采用 DC；在这种情况下，建立一个统一的元数据数据模型显得尤为重要，这也是大型文献数字化项目的元数据互操作面临的特有问题。

其次，大型文献数字化项目所涉及的文献种类和格式多种多样，要实现项目内部数字资源的有效识别、传递和应用，就必须构建统一的元数据数据模型。就目前我国大型文献数字化项目发展的现状来看，针对具体类型的文献，如古籍、拓片、甲骨、家谱、舆图等的元数据标准已经建立，但是统一的元数据框架尚未搭建完成，可以借助于元数据注册系统予以全面收集，也可以构建基于 FRBR 模型的 RDA 要素集，不仅能对所有类型文献的描述元素进行全面呈现，还可以借助于 FRBR 揭示的关系路线，有效建立作品、形式、载体、单件等不同实体间的关系。

(5)开展知识组织系统的互操作

本文所调研的元数据互操作是互操作层面的一个基本问题，也是目前大型文献数字化项目在互操作领域采取的主要措施。除此之外，还存在很多高级别的互操作问题，其中知识组织系统的互操作就是目前本领域面临的难题之一。曾蕾认为："元数据的互操作看似简单，一旦一些元素已经赋值，而且其取值于受控词汇时，就会使情况变得十分复杂，难免会造成互操作过程中数据的失真或遗失。"[18] 目前基于术语表、叙词表、本体等的互操作研究已经广泛开展，也取得了很多成果，但多是针对某一具体领域，如农业、教育行业等；大型文献数字化项目的文献资源涉及学科广泛，年代跨度大且类型多样，需要用到多种受控词汇，可以借助本体、SKOS、NKOS 等数据模型发现知识组织系统中内容概念及其相互关系，利用这些关系可以构建、复用知识组织系统，进而从根本上保障互操作的质量与效果，为信息资源的无缝链接与应用奠定基础。

参考文献：

[1] Alemu G, Stevens B, Ross P. Towards a conceptual framework for user-driven semantic metadata interoperability in digital libraries a social constructivist approach [J]. New Library World, 2012, 113 (2).
[2] Jung-ran Park. Semantic Interoperability and metadata quality: An analysis of metadata item records of digital image collections [J]. Knowledge Organization, 2006, 33(1).
[3] 张晓林. 元数据研究与应用[M]. 北京：北京图书馆出版社，2002. (Zhang Xiaolin. Metadata research and application[M]. Beijing: Beijing Library Press, 2002.)
[4] Marcia Lei Zeng, Qin Jian. Metadata[M]. Neal-Schuman Publisher Inc., 2008.
[5] 毕强，朱亚玲. 元数据标准及其互操作研究[J]. 情报理论与实践，2007(5). (Bi Qiang, Zhu Yaling. Research on metadata standard and its interoperability[J]. Information studies: Theory & Application, 2007(5).)
[6] Philip Hider. Australian digital collections: Metadata standards and interoperability [J]. Australian Academic & Research Libraries, 2004, 35(4).
[7] Lina Bountouri, Mannolis Gergatsoulis. Interoperability between archival and bibliographic metadata: An EAD to MODS crosswalk[J]. Journal of Library Metadata, 2009, 9(2).
[8] Lina Bountouri, et al. Metadata interoperability in public sector information[J]. Journal of Information Science, 2009, 35(2).
[9] Lim Shirley, Chern Li Liew. Metadata quality and interoperability of GLAM digital images[J]. Aslib Proceedings, 2011, 63(5).
[10] Julie Weagley, et al. Interoperability and metadata quality in digital video repositories: A study of dublin core [J]. Journal of Library Metadata, 2010, 10(1).
[11] CADAL. CADAL 元数据规范草案(Version 2.0)[R/OL]. [2012-05-27]. http://www.cadal.cn/cnc/cn/

jsgf/CADAL_metadata_2004.pdf. (CADAL. CADAL Edocument metadata (Version 2.0)[R/OL]. [2012-05-27]. http://www.cadal.cn/cnc/cn/jsgf/CADAL_metadata_2004.pdf).

[12] 北京大学图书馆. 国家图书馆核心元数据标准[R/OL]. [2012-05-27]. http://www.nlc.gov.cn/sztsg/2qgc/sjym/files/2.pdf. (Peking University Library. National library core metadata standard[R/OL]. [2012-05-27]. http://www.nlc.gov.cn/sztsg/2qgc/sjym/files/2.pdf.)

[13] Europeana. European library metadata[EB/OL]. [2012-05-27]. http://www.theeuropeanlibrary.org/portal/organisation/handbook/display_en.html.

[14] Open Library. Open library metadata[EB/OL]. [2012-05-27]. http://openlibrary.org/about/infogami-dc.

[15] Hathitrust. Hathitrust metadata[EB/OL]. [2012-05-27]. http://www.hathitrust.org/hathifiles_description.

[16] CDL. Guidelines for digital objects [R/OL]. [2012-05-27]. http://www.cdlib.org/services/dsc/contribute/docs/GDO.pdf.

[17] 林海青. 元数据互操作的逻辑框架[J]. 数字图书馆论坛,2007(8). (Lin Haiqing. The logical framework for metadata interoperability[J]. Digital Library Forum,2007(8).)

[18] Marcia Lei Zeng, Lois Mai Chan. Metadata interoperability and standardization-A study of methodology part II [EB/OL]. [2012-05-27]. http://www.dlib.org/dlib/june06/zeng/06zeng.html.

[19] Antoine Isaac. The Europeana data model[EB/OL]. [2012-05-27]. http://pro.europeana.eu/documents/866205/13001/EDM_v5.2.2.pdf.

[20] Bernhard Haslhofer, et al. Europeana RDF store report [R/OL]. [2012-05-27]. http://www.europeanaconnect.eu/documents/europeana_ts_report.pdf.

[21] Peter Murray. Mashups of bibliographic data: A report of the ALCTS midwinter forum [R/OL]. [2012-05-27]. http://dltj.org/article/mashups-of-bib-data.

[22] Nuno Freire, Diogo Reis. Guidelines for preparing a Z39.50/SRU target to enable metadata harvesting[R/OL]. [2012-05-27]. http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/documents/TELplus-D2.3_v1.0.pdf.

[23] Europeana. D5.1.1 Europeana metadata registry [EB/OL]. [2012-05-27]. http://pro.europeana.eu/documents/12117/1000137/The+Europeana+Metadata+Registry.

[24] Linked data FAQ [EB/OL]. [2012-05-27]. http://structureddynamics.com.

[25] Europeana linked data[EB/OL]. [2012-05-27]. http://pro.europeana.eu/linked-open-data.

宋琳琳 中山大学资讯管理学院讲师,博士。
通讯地址:广州市广州大学城外环东路132号中山大学资讯管理学院。邮编:510006。
李海涛 中山大学资讯管理学院讲师,博士。通讯地址同上。

(收稿日期:2012-03-14;修回日期:2012-05-27)