图书馆特藏数据结构化的探索

陈晓亮 苏海潮 刘心舜（厦门大学图书馆）

摘 要 图书馆特藏资源及其数据结构化的建设与开放是参与共建共享国内数字基础设施的切实可行的道路，应采用国际通行的标准与系统。文章介绍了新一代文化机构藏品内容管理系统Omeka S、国际图像互操作框架（IIIF）及其在厦门大学图书馆的应用。

关键词 特藏资源管理 结构数据 国际图像互操作框架 Omeka S

DOI: 10.13663/j.cnki.lj.2019.06.007

The Structured Data of Library's Special Collection

Chen Xiaoliang, Su Haichao, Liu Xinshun (Xiamen University Libraries)

Abstract This paper advocates that developing digitized special collection and building structured data are a feasible way to realize the collaboration and sharing of the national data infrastructure, and that we should learn from international practices. It also introduces a next generation of collection management system and web publishing platform, Omeka S, and International Image Interoperability Framework (IIIF) for digital cultural collection.

Keywords Special collection management, Structured data, IIIF, Omeka S

1 从特藏数字化到数据化的难题与需求

学者们指出数字图书馆从“数字化”到“数据化”，再发展为“智慧化”的趋势[1-2]。笔者在厦门大学图书馆特藏工作中的体会是，认准方向，立足现实，夯实“数字化”和“数据化”的基础，累积推进，逐步过渡到“智慧化”。

经过数字化的文献，让研究者能更便利地查阅，不再受原件保护性管理和纸质单件的束缚，扩大了使用范围，但文献彼此之间仍是孤立的，更缺乏知识点之间的关联，还需要完成从非结构文本到结构化数据的进化。这些结构化数据成为各种各样的数字馆藏、专题、文献的黏合剂，并通过开放的接口，和外部资源形成网状关联。研究者直接使用的可能是数字文献，但在使用过程中，隐藏在数字文献背后的结构化数据及其网络，可以帮助研究者去发现更多可供研究的资料、发现新的研究视角等。

数据化项目最困难也是最重要的问题在于到底想要做什么内容。在信息资源高度同质化的情况下，特藏资源可谓是图书馆参与合作竞争的最大资本，特藏资源的价值在大数据时代更加凸显，特藏资源为王。相应地，特藏资源数据规模越大、开放程度越高的图书馆，越容易保持特色。因此，在投入有限的情况下，图书馆应该优先发展特藏资源及其数据化项目。

根据“联合、开放、数字化”的战略目标，厦门大学图书馆的特藏数字资源建设，在采取自上而下、以点带面的项目驱动模式（如所承担“中国稀见史料”“台湾光复前后（1943–1946）”“厦门大学海疆剪报资料选编”、CASHL特藏项目“美国外交部档案——越南”数据平台、CALIS特色数据库项目“东南海疆研究数据库”等）的同时，以这些项目为网络节点，采取“从内部网络确定中心节点及其关系、对外链接更大网络”的开发与开放模式，建设的重点在“东南海疆研究资料”的网络化全面整合。这样的战略规划，决定了所选择的

* 通讯作者：苏海潮，E-mail：hcsu@mail.xmu.edu.cn

技术方案的国际化方向。

与大部分图书馆类似，在特藏数字资源挖掘和整理过程中，厦门大学图书馆一直面临以下两大难题：

第一，特藏文献的整理、著录、揭示不易。特藏文献类型多样，包括图书、古籍、报纸、期刊、地图、手稿、卷轴、档案材料、图片、书信，著录时描述的项目难以统一。著录时的层级和颗粒也不一致，同时存在资源集合、全册、章节、单篇、单页、文本的不同层级。在不同类型的文献、文献的不同层级之间，还存在演绎、从属、版本、责任等诸多关系。仅仅著录文献本身已经不能满足需求，馆员们还需要揭示文献之间、文本之间的关系和语义。

第二，纸质文献数字化后影像体积过大，阅读查看很不方便。与对普通文献电子版的要求不同，纸质特藏文献电子版的使用者除了对文献的文本有需求，对文献（特别是地图、古籍、文书等）的形式、细节等也有兴趣，因此，更需要阅读查看更高清的影像。问题在于，更高还原度的影像意味着更大的文件体积，这与使用者的网络带宽及计算机终端处理能力相互矛盾。以一幅 75*105 cm 的彩色地图为例，使用 600DPI 分辨率及 8 位深扫描、未经压缩，保存得到的原始 TIFF 彩色影像体积约为 1.36 GB。即使转换为 jpeg 格式，仍有 280 MB。通过网络传输这样一幅影像给使用者查阅，既要同时满足纵观全局和体现细节的要求，还要保证使用者的客户端能轻松加载如此大幅的图片，难度大。

厦门大学图书馆虽然已经完成包括古籍、民国书刊、东南亚华文报章、剪报资料等特藏文献的数字化项目，但是，其中大多数仅仅完成载体形态转换，或是属于出版项目。此外，还有不少非书刊文献脱离于图书馆书目系统之外。因此，厦门大学图书馆的特藏资源建设，迫切需要一个集成的数字资源管理系统，要求这个系统：（1）能够描述形态各异的文献资源，不论是纸本还是电子文献；（2）具备一定的灵活性，或详或简地描述文献资源，并揭示资源之间的关系；（3）能够与外部环境的资源关联；（4）在当前技术人员投入不足的情况下，系统的技术框架不必过于复杂，但必须方便数据迁移，为将来向更好的平台转化创造条件。

2　特藏数据标准与系统国际化

针对以上问题，结合现实需要，厦门大学图书馆采用了符合国际化方向的核心技术解决方案：以 Omeka S 作为特藏数字资源管理平台，采用基于国际图像互操作框架（International Image Interoperability Framework，简称 IIIF）的 Loris2 图片服务展示高清影像。

2.1　Omeka S 用于特藏数据管理

Omeka S 是美国乔治梅森大学的罗伊·罗森茨维格历史与新媒体中心在 2017 年底正式推出的新一代文化机构藏品内容管理系统，以 GPL-3.0 协议开放源代码。对厦门大学图书馆而言，所选用的 Omeka S 有以下优点[3]：

（1）在设计和开发时，Omeka S 遵循关联数据标准，资源的描述和表示非常灵活。该系统预置了 4 种常用的 RDF 词表：Bibliography Ontology、Dublin Core、Dublin Core Type、Friend of a Friend。描述文献时，可从这些本体中选择合适的词汇，也允许导入互联网上开放的第三方本体。此外，Omeka S 使用 JSON-LD（JavaScript Object Notation for Linked Data）作为数据交换格式。

（2）资源的组织非常灵活。Omeka S 的核心描述对象是“Item”。所有进入系统的 Item，不仅可以同时分属不同的集合（Item Set），也可以直接或间接通过 Item Set 作为一个站点（site）的成员，从而极其方便、灵活地建立专题库和专题网站。

（3）模块化设计及应用编程接口（API），让功能扩展和数据的发布、分享更加灵活。Omeka S 基于 ZendFramework3.0 框架（ZF3）开发，也继承了 ZF3 的组件特性，要扩展系统的功能、设计样式独特的站点主题都不太难。同时，不仅可以通过 HTTP REST API 接口从外部获取数据，即使在系统内部，模块之间也是通过 API 操作数据的。

简而言之，Omeka S 是一个开放灵活的开

源管理系统：它有开放的数据结构和数据接口，能让系统内的资源更方便地和外部语义网世界的开放数据相关联；它也能灵活地增减功能模块、定制用户界面、供第三方调用数据。Omeka S满足厦门大学图书馆目前的需求，也为将来的更新换代留下更大的可操作空间。

2.2 基于国际图像互操作框架（IIIF）的图片服务

IIIF不是一个系统或者一个软件，而是一个图像互操作框架、一套标准和协议。该框架中，定义了操作图片的影像（Image）、呈现（Presentation）、验证（Authentication）、搜索（Search）4种API接口，以及实现这4种接口时服务端或客户端应遵循的标准。它的核心目的在于，只要文化遗产收藏机构遵循IIIF框架在互联网上发布图片，研究者都可以很方便地查看、比较、操作和标注这些影像资源。

得益于社区的力量，不论是服务端，还是客户端，部署IIIF的选择很多，基本步骤可以简单概括为3步：（1）搭建支持影像（Image）API的图片服务器。（2）按照呈现（Presentation）API的协议发布影像资源的元数据。（3）将上述两步揭示出来的影像和元数据，集成到业务系统中[4]。根据各机构具体情况的不同，可能还涉及部署身份验证和搜索两大模块。

其中的影像API，提供了操作原始影像的标准HTTP（S）接口。向这个接口请求的统一资源标识符（URI）通常包括原始影像的唯一标识符和特定的参数，通过参数值的变换，实现对指定影像的操作，并得到符合需求的图片。可指定的参数包括区域、尺寸、旋转角度、色彩和图片格式，完整的URI结构如下：

{scheme}：//{server}{/prefix}/{identifier}/{region}/{size}/{rotation}/{quality}.{format}

如果在服务器imageserver.org上使用loris2发布了自己的数字影像，客户端向它发起请求：把影像abcd1234从坐标点（125，15）截取宽120像素、高140像素的部分，并按比例缩小50，再顺时针旋转90度，转换成灰度影像，最后以JPEG的格式返回给客户端。这时请求的具体URI是：

https：//imageserver.org/loris2/abcd1234/125，15，120，140/pct：50/90/grey.jpg

在这种灵活的机制下，资源所有者只要对原件全尺寸扫描一次，并保存一份原始影像，就足以满足各种不同需求；资源调用者只要知道唯一标识符，就能方便地调用影像，不论是完整图片还是局部细节。影像API解决了查看高清影像便利性的直接需求，如果进一步部署呈现API，并通过语义标注工具和语义互联网世界的结构化数据关联，也可以揭示和操作附着在影像上的知识。

2.3 厦门大学图书馆的实践

2.3.1 资源描述

初始安装的Omeka S默认提供4种词表，包括都柏林核心元素集（Dublin Core）和DCMI类型词表（Dublin Core Type）、反映人物关系的“朋友的朋友”（Friend of a friend，FOAF）、描述书目数据的BIBO书目本体（Bibliographic Ontology）。这4种词表基本能满足描述一般文献资源的需求。但是，考虑到馆藏资源中除了书刊文献外，还有诸如早期毕业论文、剪报资料、往来通信、手稿等其他资源，厦门大学图书馆在自定义词表的同时，也引入第三方开放词表：自定义的词表主要定义标识符和机构信息，如MARC记录号、财产号及学院系别机构；引入第三方开放词表，如“基于BibFrame的手稿及档案本体（SHLSG）”[5]，以应对捐赠关系、往来通信、近代人物、个人档案等描述需求。

引入自定义或第三方的词表后，还需要针对不同描述对象定制描述模板（Resource Template），即选择合适的描述词、指定取值类型，这个过程相当于制定特定类型资源的著录规范。套用事先订制的描述模板，并不意味着无法选择其他词汇，仍然可以任意增减描述项目。这也是Omeka S在描述资源方面时灵活性的体现——描述特定类型资源时很灵活，在描述具体一项资源时也很灵活。表1和表2分别是“早期毕业论文”和“人物”的描述模板样例。

2.3.2 资源彼此关联

馆藏中的资源并不是彼此孤立存在的。以虚拟描述对象——人物为例，人物经常承担多

表 1 “早期毕业论文”描述模板

Vocabulary	Original label	Data type	Alternate label
dcterms	Title	Default	题名
dcterms	Creator	Item	作者
xmutc	指导老师	Item	
dcterms	Description	Default	简介
xmutc	学院	Default	
xmutc	系	Default	
xmutc	学号	Default	
xmutc	XMU Tecang Item No.	Default	特藏编号
xmutc	XMU Call No.	Default	索取号
dcterms	Date Submitted	Default	完成时间
xmutc	文件数量	Default	

表 2 “人物”描述模板

Vocabulary	Original label	Data type	Alternate label	Alternate comment
dcterms	Title	Default	姓名	
foaf	gender	Default	性别	
foaf	lastName	Default	姓氏	
foaf	firstName	Default	名字	
shlsg	字	Default		
shlsg	号	Default		
shlsg	籍贯	Default		
dcterms	Description	Default	简介	
shlsg	生于	Default		
shlsg	卒于	Default		
shlsg	出生地	Default		
shlsg	死亡地	Default		
shlsg	著作	Default		
foaf	is primary topic of	Default	以此为主题的	
shlsg	关系人	Default		
schema	alternateName	Text	别名	原名、曾用名、笔名

种社会角色——可以是其他资源的责任者，也可以是其他人物的研究对象，还可以是其他事件的核心要素。如民国人物“叶长青”，他是图书的责任者、期刊的主编、报纸文章的作者，也是剪报资料（报纸文章）的事件人物，以他为纽带，相关的馆藏资源可以彼此关联，形成网状结构。表 3 和表 4 是人物条目“叶长青”的基本描述信息和它与其他资源条目的关联状况。在 Omeka S 系统中，词表里的不少术语可以从语义上表达资源彼此的关系。在录入这些术语的取值时，也可以通过“文本值”（直接值），“系统中其他资源”（资源项目、资源项目集、媒体文件），URI 将资源在系统内部及与外部其他资源实际链接起来。截至 2018 年 9 月，厦门大学图书馆已经在 Omeka S 中创建和导入约 75 000 条项目，类型涵盖古籍、近现代图

书、民国期刊、报纸、剪报、人物、早期毕业论文等。目前计划通过 API 接口，关联互联网开放数据或第三方平台，如上海图书馆人名规范库、维基数据、厦门人物辞典等。

表 3 人物“叶长青”简要描述信息

Class	Person
姓名	叶长青
性别	男
姓氏	叶
名字	俊生
字	长青 长清
号	长卿
籍贯	福建闽县
生于	1902
卒于	1942

表 4 “叶长青”条目与其他资源条目的关联

Creator		
Title	Alternate label	Class
松柏長青館詩	责任者	Book
国学专刊	责任者	Journal
石遺先生傳	作者	NewsArticle
primary topic		
Title	Alternate label	Class
一封值得公開的私信；葉長青行賄□款經過		NewsArticle

2.3.3 Omeka S 与 IIIF 图像服务集成

2016 年 7 月，厦门大学图书馆开始部署 Loris2 图片服务器，实现了 IIIF 的影像接口。所有完成数字化的、超过百万幅的特藏文献高清级别影像，均可通过这个接口调用。在 Omeka S 中，可以直接添加符合 IIIF 影像接口协议的 URI 来调用图像仓储库里的高清影像。不过，这种方式适合添加单幅图片时使用，当一件特藏资源包括多幅图片时，手动添加难免费时费力。厦门大学图书馆在此基础上，结合自身需求开发专门的 IIIF 扩展模块，并整合 IIIF 图片查看器 OpenSeadragon。这样，数字化得到的影像经过质量审查、确认合格后，就可以通过该模块便捷地与 Omeka S 中的资源条目关联，不需要著录人员额外操作。

2.3.4 进一步努力的方向

图书馆应基于现实条件，选择一个尽量符合需求、现时可行、可扩展的解决方案。如基于经费投入的考虑，选择基于开放源代码软件的自主二次开发；根据技术团队的编程语言，选择相应语言开发的管理系统；根据人力的投入情况，选择相应的技术框架。厦门大学图书馆在特藏资源建设过程中难以一蹴而就，只能逐步递进，选择一个轻量级的特藏著录与管理系统使用，以满足从资源数字化向数据结构化转化的初步需求。至于资源和数据的发布、展示和利用，除了使用 Omeka S 自带的专题网站建设功能，还将通过第三方应用程序调用 API 接口的方式专门处理，以满足更多样的需求。同时，厦门大学图书馆正在研究实现 IIIF 框架的身份验证 API 和呈现 API，届时可在保证特藏数据知识产权的情况下，提供更加开放的特藏服务。

3 基本结论

“双一流”（一流大学、一流学科）建设，要求所在高校有某些学科在国际上一流或者有重要影响，要求高校图书馆具备国际视野，统筹发展。在“双一流”建设的大背景下，高校图书馆发展的根本在于资源[6]，要为学者们提供优越的研究环境，如有特色的文献资源、国际一流学者群常用的系统、标准和研究工具，主动建设和开放具有一定国际影响力的数据库与数字研究平台。

特藏资源及其数据结构化的建设与开放，是图书馆参与国内数字基础设施的共建共享切实可行的道路[7]。厦门大学图书馆的探索表明，应具备国际视野，坚持以特藏资源和图书馆特色为本，采用国际通行的系统和标准，累积地推进特藏资源数字化、数据化和智慧化进程。

参考文献

[1] 曾蕾.从“数字化”到“数据化”到“智能化”——图档博领域的智慧数据发展趋势[EB/OL]. (2017-12-06)[2018-06-02]. http://www.lib.fzu.edu.cn/adls2017/down/ZengLeiADLS17SmartData.pdf.

（下转第91页）

展营利性服务是可以行得通的路径。在国家自然科学基金和人文社会科学基金中提高资助口述历史研究项目，鼓励各高校及科研单位，尤其是图书馆从事口述历史研究，并依靠图书馆现有资源，开展营利性服务，鼓励企业和个人参与口述历史工作，努力将口述历史资源推向市场，推动口述历史向市场化、产业化方向发展[12]。此外，图书馆通过荣誉奖励的方式，如出版捐资人名录、给予捐资人图书馆建筑命名权的方式，积极呼吁校友、公众、社会组织团体捐资也可以作为图书馆口述历史工作资金的来源方式。

参考文献

[1] 尹培丽.口述资料收藏——图书馆的新领地[J]. 大学图书馆学报, 2013(4): 14–18.
[2] 冯云. 我国图书馆口述历史研究综述[J]. 图书馆工作与研究, 2015(2): 21–24.
[3] 周晓燕.基于著作权法视角的图书馆口述文献工作探析[J]. 图书馆建设, 2013(4): 4–8.
[4] 王子舟, 尹培丽. 口述资料采集与收藏的先行者——美国班克罗夫特图书馆[J]. 中国图书馆学报, 2013(1): 13–21.
[5] 王天红. 口述历史: 国家图书馆关注的新领域[J]. 图书与情报, 2006(5): 20–23.
[6] 张一, 谢兰玉. 网络环境下美国图书馆口述历史用户服务研究[J]. 图书馆建设, 2017(3): 66–72, 77.
[7] 王惠玲. 香港经验分享——口述历史档案库的创建兼论图书馆的角色[J]. 图书馆, 2015(12): 10–14.
[8] 韦桥明, 颜祥林. 基于网络调查的美国科技口述历史采集项目浅析[J]. 档案学通讯, 2016(4): 50–55.
[9] 胡立耘. 基于口述历史的图书馆延伸服务[J]. 图书馆, 2015(12): 15–22.
[10] 彭燕, 余弦, 李良嘉. 口述历史在土家族挑花研究中的应用——以吉首大学图书馆的实践为例[J]. 高校图书馆工作, 2017(3): 61–65.
[11] 少数民族口述历史的挖掘与数字化保存模式研究——以武陵山区土家族为例[J]. 图书馆学研究, 2012(10): 43–45, 49.
[12] 朱华顺. 美国图书馆数字人文案例研究及启示——以布朗大学、纽约公共图书馆为例[J]. 国家图书馆学刊, 2016(6): 58–63.

钟　源　中南民族大学图书馆，馆员。中南民族大学民族学与社会学学院，博士生。研究方向：口述历史、开放政府数据。E-mail：363540604@qq.com　湖北武汉　430074
吴振寰　女，武汉职业技术学院图书馆，馆员。研究方向：图书馆新型服务。　湖北武汉　430074

（收稿日期：2018–02–02　修回日期：2018–03–21）

（上接第48页）

[2] 王晓光.数字人文与智慧数据[J]. 上海高校图书情报工作研究, 2018, 28(2): 25, 24.
[3] Omeka.What to expect in Omeka S[EB/OL]. (2016–10–26)[2018–06–02]. https: //github.com/omeka/omeka–s/wiki/What–to–expect–in–Omeka–S.
[4] Quick Start Guide – International Image Interoperability Framework™[EB/OL]. (2017–09–04)[2018–06–02]. http://iiif.io/technical–details/.
[5] 夏翠娟, 张磊, 贺晨芝.面向知识服务的图书馆数字人文项目建设: 方法、流程与技术[J]. 图书馆论坛, 2018, 38(1): 1–9.
[6] 程焕文.浅谈高校图书馆发展趋势[J]. 图书馆论坛, 2018, 38(7): 58–61.
[7] 刘炜, 谢蓉, 张磊, 等.面向人文研究的国家数据基础设施建设[J]. 中国图书馆学报, 2016, 42(5): 29–39.

陈晓亮　厦门大学图书馆特藏部，馆员。研究方向：特藏资源建设、数据保存。作者贡献：论文第1部分“提出问题”、第2部分的撰写与修改。E-mail：sogg@xmu.edu.cn　福建厦门　361005
苏海潮　厦门大学图书馆特藏部，副研究馆员。研究方向：图书馆合作、社会网络分析、特藏建设。作者贡献：第1部分引言及第3部分结论的撰写，论文写作指导与审核。　福建厦门　361005
刘心舜　女，厦门大学图书馆特藏部主任，副研究馆员。研究方向：特藏资源建设与管理。作者贡献：论文内容的讨论与修改。　福建厦门　361005

（收稿日期：2018–07–25　修回日期：2018–11–26）