数字人文在高校图书馆特藏资源建设中的实践与思考

——以近代中译本全文特藏库建设为例*

张　毅　赵晨鸣　陈　丹

摘　要　电子资源取代纸质资源导致的同质化问题,使特藏资源的价值不断凸显,但受制于资金、技术以及研究范式,图书馆的数字特藏建设并未取得实质性突破。本研究在分析数字人文提供的研究方法与大量开源工具基础上,以华东师范大学近代中译本全文数据库建设为例,借助自然语言处理、可视化、关联数据以及 IIIF 等技术,构建以用户为中心的数字特藏揭示平台,为人文学者提供全新的特藏资源浏览与分析平台,以期对高校特藏数据库的建设提供思路与经验。图 4。表 2。参考文献 20。

关键词　数字人文　数字特藏库　近代中译本　可视化　关联数据

Practice and Thinking of Digital Humanities in the Construction of Special Collection in University Libraries

——Taking the Construction of Special Collection Library for the Full Text of Modern Chinese Translations Version as an Example

Zhang Yi　Zhao Chenming　Chen Dan

Abstract: The homogenization problem caused by the replacement of paper resources by electronic resources has made the value of special collection resources increasingly prominent. However, subject to fund, technologies and research paradigm, construction of digital special collection in libraries has not achieved substantial breakthrough. Based on the analysis of the research methods and a large number of open source tools provided by digital humanities, this study takes the construction of the full-text database of modern Chinese translation version in East China Normal University as an example, and uses natural language processing, visualization, linked data, IIIF and other technologies to build a user-centered reveal platform for digital special collections. In doing so, we can provide humanities scholars with a new platform for browsing and analyzing special collections resources, in order to provide ideas and experience for the construction of special collections databases in colleges and universities. 4 figs. 2 tabs. 20 refs.

Keywords: Digital Humanities; Digital Special Collection Library; Modern Chinese Translation Version; Visualization; Linked Data

特藏资源是具有研究意义的珍稀版本的汇集,或是某类高价值资源的集合,不仅是教学科研的一手资料,更是高校文脉的传承。国内高校图书馆和相关领域研究者都非常重视数字特藏资源的建设,王乐认为特藏资源是图书馆的核心竞争力[1],蔡迎春提出利用可视化、关联数据与文本挖掘等技术进行特藏资源库建设[2],韩冰关注数字特藏库建设的版权与开放力度[3],蔡梦玲建议在贝叶经特藏库建设中借鉴国外特藏库建设经验来优化数据采集与加工[4],李宗富等总结

* 本文系国家社会科学基金项目“高校图书馆特藏资源服务模式及站群系统研究”(项目编号:21BTQ100)研究成果之一。

DOI:10.13666/j.cnki.jnlc.2023.0107

了俄罗斯装饰数字档案库建设中利用 VR、短视频等渠道宣传数字特藏资源的经验[5]。

虽然我国图书馆领域已经意识到特藏资源的重要价值,并有不少研究与实践逐步借助数字人文方法与技术深度挖掘特藏资源的价值。但是由于缺少完善的特藏资源建设软件与服务商,我国图书馆无法便捷、有效、有深度地进行特藏资源建设,定制开发的特藏资源库不仅稳定性差,也不能满足数字时代读者的需求。随着数字技术的不断迭代,在开放社区与开源软件的支持下,数字人文的发展逐渐成熟,可为数字特藏资源建设提供研究范式,面向数字人文的开源软件也为数字特藏资源建设提供了技术支持[6],如 Islandora、Samvera、Goobi 等数字特藏发布系统,可以为数字特藏资源建设提供长期保存、发布、互操作、众包等功能,Loris、Cantaloupe、Mirador 等在线图像浏览开源软件,可实现数字图像的高清在线浏览。本研究对国内外高校特藏数据库项目的建设情况进行梳理,着重关注数字人文对特藏库的支持,归纳适用的数字人文工具,并通过华东师范大学近代中译本全文数据库的建设实践,为人文学者提供全新的特藏资源浏览与分析平台,以期对高校特藏数据库的建设提供思路与经验。

1 数字人文与特藏数据库

1.1 数字人文的定义

当前,数字人文并没有统一的定义,一般认为数字人文是一种全新的信息组织形式,是适应数字化的全新研究范式,帮助人文学者从繁琐的资料收集整理中解放出来[7]。数字人文是计算机与人文学科的交叉研究领域,是一种资源揭示与学术研究的新形式,它利用数字工具打破了印刷媒介在知识生产和传播中的局限[8]。随着计算机技术的不断发展,数字人文涵盖的内容也在不断拓展,从最初的人文计算、定量分析等,逐步发展到大型图像集的可视化、历史文物的 3D 模型、自然语言处理、图像识别等[9]。

1.2 数字人文在特藏库建设中的应用

研究团队曾于 2021 年 6 月到 2022 年 1 月之间对 U. S. News 世界大学排名前 100 的高校以及国内重要高校的图书馆数字特藏进行了调研[10],访问了 82 个数字特藏库,发现:大部分数字特藏资源的内容集中在历史资料、政治文件、珍稀手稿、照片地图、建筑图片、乐谱等几大类,近年来新建或新改版的特藏数据库更多地加入了关联数据、图像语义、全文检索、知识图谱等数字人文技术,不仅为人文学者提供了数字资源在线高清浏览功能,也扩展了人文学者的研究边界。例如耶鲁大学通过数字人文技术,将威廉·布莱克绘画手稿中的色彩进行量化和可视化研究,为研究者理解威廉·布莱克绘画提供了新的材料与视角①。哈佛大学图书馆以编程接口方式开放全部馆藏记录,邀请数据科学家、图书馆员以及人文学者等对哈佛大学馆藏进行文本挖掘、可视化分析等,从而发现新的研究领域与趋势②。

1.3 面向数字特藏建设的开源软件调查

数据库建设、在线高清浏览、可视化、文本挖掘等的实现,离不开计算机软件的支持。然而软件系统开发周期长,需要大量的技术储备与大量资金,这导致国内至今仍缺少优质的特藏资源数据库厂商,单个图书馆自主开发更是存在很大的失败风险。免费、数量多、类型丰富而且有社区支持的开源软件,成为图书馆数字特藏建设的最佳选择之一。通过调研开放存取存储库联盟(COAR)开发的学术交流技术目录(SComCat)、哈佛大学数字艺术与人文(DARTH)、奥地利数字人文中心(ACDH-CH)、布朗大学图书馆等机构和高校发布的数字人文工具指南[11-14],本研究归纳出了适用于特藏资源建设的开源软件,并划分为采集与整理、保存与发布、资源揭示以及在线浏览四类,详细信息如表 1 所示。

① Yale university library dhlib. https://dhlab. yale. edu/projects/blaketint.

② Harvard LibraryCloud. https://wiki. harvard. edu/confluence/display/LibraryStaffDoc/LibraryCloud.

表1 可用于特藏资源建设的数字人文工具

采集与整理				保存与发布		
本地数据	注释	原生数据采集	语料库制作	数字对象管理	特藏管理发布	图像处理
Zotero	Annotot	Textgrid	CorpusExplorer	Fedora	Omeka	ImageMagick
Tropy	Elucidate	Social Feed Manager		Dspace	Islandora	Apache GD
JabRef	Miiify	Webrecorder			Drupal	Light Room
资源揭示				在线浏览		
图像识别	可视化	数据分析	检索与索引	图像服务器	高清浏览	图像互操作
Giles-E	TimelineJS	Breve	Solr	Loris	Mirador	IIIF
OCRmyPDF	StorymapJS	Lexos	Blacklight	Cantaloupe	Diva. js	
MuPDF	Gephi	TinEye		Digilib	Universal Viewer	

2 具有数字人文功能的特藏数据库建设思路

当前,图书馆服务重心已从纸质资源转向数字资源,图书馆自建数字特藏资源的价值不断凸显,这就要求图书馆具备对海量数字资源进行长期保存、管理以及揭示等方面的能力。全球前100的高校中,已经有82所建立了数字特藏库,并且针对数字媒介的特点,利用数字人文技术对特藏资源进行管理与揭示,数字人文的功能大多可借助开源软件实现。本研究在分析特藏库建设所需要的开源软件基础上,结合数字人文范式,归纳出以下特藏数据库建设思路。

2.1 数字特藏加工与采集

(1)本地数字对象加工管理

数字人文为人文学者提供了对本地海量文档、图片以及视频进行统一加工管理的工具,比如文献管理工具Zotero、图片管理工具Tropy。这些工具不仅可以对本地资源添加标签,而且可以将本地资源导出为XML、CSV等格式。其最为强大的功能是可以直接与特藏资源发布服务器Omeka等平台互通,方便人文学者将本地高价值资源对外一键发布。

(2)特藏资源在线标注

标注是对特藏资源二次加工的重要手段,利用开源软件,构建数字特藏资源在线注释功能,引导读者参与资源在线标注,标注后的数字资源可以为用户提供分类、标签等浏览方式。Annotot、Elucidate、Miiify等开源注释服务器软件符合W3C网络注释数据模型,提供团队合作、数据导出、自动识别实体等功能,可以实现对大量数字特藏进行标注,而且标注数据可以直接发布为关联数据,提高数据的可见性。

(3)原生数据采集

在数字特藏库建设方面,除了对本馆纸质特藏数字化外,对于在线图片、视频、网页镜像等原生数字资源的收集也非常有必要,利用Social Feed Manager可以从Twitter、Tumblr、Flickr和新浪微博等社交媒体上收集网络资源,而Webrecorder则可以利用Web存档服务来保存网页镜像。采集数据的一个重要目的是构建语料库,利用CorpusExplorer可以将采集的数据形成语料库,同时CorpusExplorer自带网络爬虫,也可以作为数据采集的工具之一。

2.2 数字对象长期保存与在线发布

数字资源相较于纸质资源更易丢失损坏,除了在硬件方面需要做好如双路供电、存储的磁盘冗余阵列(RAID, redundant array of independent disks)等预案外,在软件方面也需要可以预防数据丢失的数字对象管理系统[15],数字对象管理系统通常采用二进制验证的方式来确保数据的安

全。比如 DuraSpace 社区管理的 Fedora 工具①,它在实现数字对象长期保存功能方面非常优秀,而且可与 Drupal 等内容管理系统灵活整合,实现资源的揭示。

资源发布与管理系统是特藏库建设的基础,一旦部署并付诸实施,后期的更换成本会很高,所以对于管理发布系统,需要尽量选择自带资源导出工具与数据接口的特藏管理系统[16],比如 Omeka 与 Islandora。Omeka 直接以文件形式存放数字对象,只进行简单的数据验证,但系统简单灵活、功能丰富,比较适合数据量较少的系统;Islandora 利用 Fedora 实现数字对象长期保存,提供数字对象校验功能,更适合大量数字对象长期保存。特藏资源主要以图像资源为主,因此在特藏资源库建设中,选用那些可对多种类型图像进行格式转化的工具尤为重要,ImageMagick 就是一款可以对图像进行格式转化、拼接、修改等操作的工具②,Apache 服务器可以直接调用它生成需要的图形。

2.3 资源揭示工具

(1)OCR 与图像识别

OCR 与图像识别为特藏资源的深度揭示提供可能,OCR 识别的内容可以作为全文检索的依据,而图像识别则可以实现特藏资源中图像的分类与标注,再借助数据分析工具进行索引、交互式过滤、实体命名等操作,可实现多维数据的分析,还可以以图形方式呈现分析结果。百度、阿里、谷歌等人工智能公司提供的接口对于通用文本的识别更加迅速准确,但费用昂贵且不太适合特殊领域的手稿、古汉字等;而 Giles、OCRmyPDF 等开源数字人文工具的灵活性更高,并且完全免费,图书馆还可以根据需要识别的内容进行个性化配置从而提高识别率。

(2)可视化工具

可视化技术为人文学者对数字特藏资源进行远读提供支持,让人文学者直观了解某个领域的整体面貌,常见的可视化技术有地图、时间线、文本分析等。比如 TimelineJS 结合 StorymapJS 工具可以通过地图以时空检索方式为用户提供资源展示,而对于海量文本资源的可视化则可借助 Lexos 工具,Breve 更适合处理混乱的表格数据,TinEye 则是一种图像可视化工具,可以对海量图像进行分类整理。

(3)搜索工具

为了让用户能够迅速有效地获取需要的资源,在对数字对象进行文本识别后,借助开源软件 Solr 对全部文本进行索引与缓存,可以极大提升用户的检索体验。通过配置 Solr 的 Suggest 模块,还可以使数字特藏库具备检索建议功能,读者在检索框输入第一个字符时,系统就会给出相关可能的检索关键字,帮助读者在不了解特藏库的情况下,发现特藏库包含的资源。

2.4 数字对象在线高清浏览

数字特藏资源一般以图片格式对外发布,因此在特藏库建设中,既要确保数字对象的高清在线浏览,又要保证访问速度,IIIF 标准是能够满足这种高需求的技术③。该标准定义了数字对象在线浏览的 6 种接口标准,最为重要的是图像接口与展示接口。其中,图像接口负责动态发送数据到用户终端,已经有 Cantaloupe、digilib、Loris 等图像服务器支持 IIIF 的图像接口;而展示接口则对获取的数据进行本地浏览,通常使用一组 JS 代码完成,能够完成展示接口的开源软件有 Mirador、Universal Viewer、Diva. js 等。

3 数字人文技术在近代中译本全文特藏库中的应用

译介是我国近代爱国仁人志士救国图存的重要措施,通过翻译西方著作向我国人民介绍西

① Fedora. https://duraspace. org/fedora.

② Imagemagick. https://imagemagick. org/index. php.

③ IIIF. https://iiif. io.

方的科技、文化以及政治等[17]。本研究团队所在的华东师范大学图书馆收藏有大量近代中译本，2012年，华东师范大学图书馆在参与大学数字图书馆国际合作计划（CADAL）项目时已经完成了1271种近代中译本的数字化。但当时虽然完成了资源格式的转化，并未进行有效的揭示，读者无法便捷地发现并利用这些宝贵的资源，造成资源的闲置。本研究将利用数字人文方法对这批近代中译本进行管理与揭示，同时为人文学者提供研究工具，图1是华东师范大学近代中译本全文特藏库的整体框架。

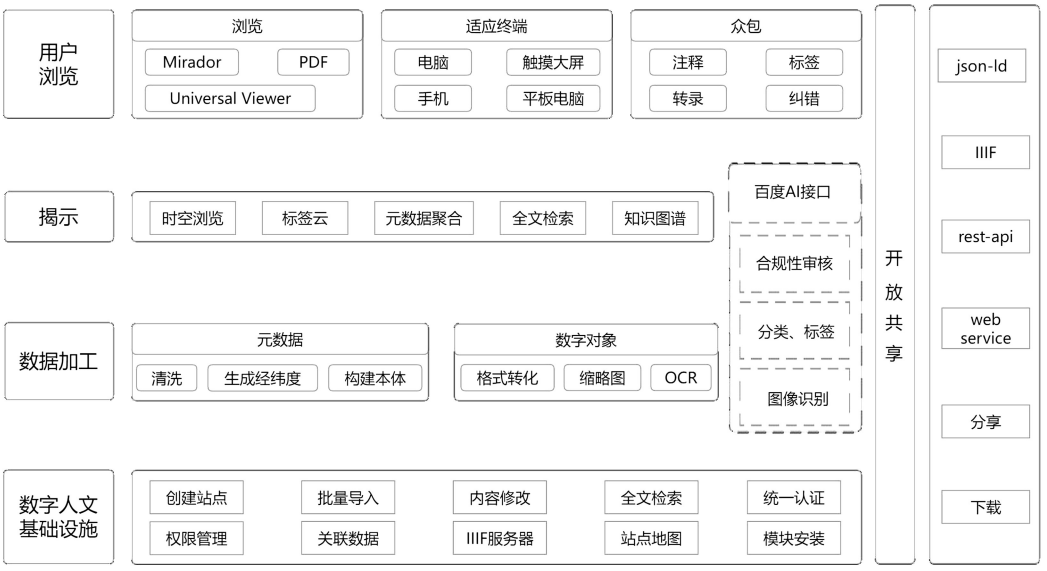

图1　华东师范大学近代中译本全文特藏库框架

3.1　利用开源软件Omeka-S发布近代中译本

数字人文技术为特藏资源的发布提供了多种开源工具，华东师范大学图书馆已经基于开源软件Omeka-S构建了本地数字特藏发布基础设施①，只需要利用Omeka-S站群管理平台创建一个“近代中译本全文特藏库”站点，选择一个适合的前台页面模板，就完成了图书基本发布平台的构建。并且，Omeka-S默认提供的模板都支持响应式页面，在不同尺寸终端上都可以获得不错的用户体验。它可以根据需要创建多个站点，站点之间可共享资源，资源自动以语义方式发布，并且具有非常强大的扩展能力。因此，Omeka-S非常适合作为高校图书馆数字特藏与文化遗产在线发布平台的基础设施，东京大学、塔斯马尼亚大学、巴黎文理研究大学等高校图书馆都采用Omeka-S构建其数字特藏基础设施。

3.2　元数据与自然语言处理

（1）多本体构建词表

为了将近代中译本以关联数据发布，首先构建具有语义的数字特藏资源库，并结合近代中译本元数据特点，选择dublin core（dcterms）、Bibliographic Ontology（bibo）以及SCHEMA.ORG（schema）三种本体来构建近代中译本全文特藏库词表模板，最终词表如表2所示。

① Omeka s. https://omeka.org/s.

表2　近代中译本详细字段信息

dcterms:title	dcterms:creator	dcterms:table of Contents	schema:nationality	bibo:translator
dcterms:publisher	dcterms:created	schema:address Region	schema:record Label	dcterms:contributor
bibo:author List	dcterms:subject	dcterms:rights Holder	dcterms:description	bibo:num Pages
dcterms:identifier	bibo:status	schema:aggregate Rating	Path	Latitude/Longitude

(2)内容智能审核、标签与分类

政治敏感、暴恐、违禁以及色情等违规内容是影响近代中译本全文特藏库发布的重要因素,在无法实现对大量图书的合规性进行人工判断的情况下,本研究调用百度人工智能接口①,利用其自然语言处理能力对每一本图书的题名、作者、摘要以及目录等信息进行分析,标记出可能存在违规内容的图书,同时生成每一本图书的标签与分类。基于 Microsoft. NET Framework 3.5 技术框架创建 Winform 程序,来调用百度自然语言处理接口,对本地近代中译本进行批量分析,百度自然语言处理接口的 TextCensorUserDefined 方法用来进行内容审核,Keyword 方法用来生成标签,Topic 方法用来对近代中译本进行分类。百度人工智能接口赠送 50 万次调用,完全满足本研究的需要。

3.3　数字对象加工与批量导入

(1)海量 DjVu 图片批量转换

参与 CADAL 项目是高校图书馆数字特藏资源的重要来源,CADAL 项目于 2001 年启动,当时 DjVu 是一种比较先进的格式,被 CADAL 项目作为数字对象存储方式,因此 DjVu 是当前很大一部分高校图书馆数字特藏资源的存储格式之一。但 DjVu 已不再被主流图像标准支持②,也无法在浏览器上直接查看。基于此,本研究经过大量的调研与测试,最终选择免费的 DjVuToy 工具③,它是目前最为简单且运行稳定的 DjVu 图片转换工具。本研究中需要格式转化的图片数量超过 25 万张,图片格式选择 TIFF。在测试过的主流图片格式中,TIFF 格式占用空间最小、清晰度最高。

(2)批量数据导入代替手工录入

要将 1271 种图书的 25 万张图片手工录入近代中译本全文特藏库,不仅很难保证数据的准确性,而且需要大量的人力与时间。本研究采用 Omeka-S 平台提供的资源批量导入功能,数据导入需要用到 CSV Import 与 File Sideload 两个模块,File Sideload 可以让导入程序 CSV Import 找到已经存放在服务器上的数字对象,数据导入过程中 CSV Import 会实时调用 Imagemagick 工具生成每一本书的缩略图。Omeka-S 默认将上传的文件重命名,并全部存放在同一个目录下面。当一个目录中存放了大量文件时,读取文件的速度就会下降,这会给用户带来不好的浏览体验。而系统自动重命名后的文件,因为文件名没有原题名,使后续手动修改时也不容易找到。所以在服务器上保持与上传文件相同的目录结构非常有必要,比如近代中译本以文学、小说、文化等类型将图书分别存放于不同的目录,要在服务器上也保持这样的目录结构,需要用到 Omeka-S 的 ArchiveRepertory 工具。

3.4　资源揭示

(1)全文检索与检索建议

Omeka-S 中配置百度 OCR 接口,在系统空闲时对近代中译本进行文本批量识别,并将识别出的文字存放在每一本图书的相应字段,配合 Apache Solr 生成索引目录,进而实现全文检索、检索建议等功能。

① 百度大脑. https://ai. baidu. com.

② DjVu. http://djvu. org.

③ DjVuToy. https://www. cnblogs. com/stronghorse.

(2)交叉分类浏览

近代中译本全文特藏库可依据资源的标签、出版年份、主题、国家等字段进行交叉筛选,用户可以精确定位到感兴趣的内容,交叉分类浏览页面如图2所示,默认情况下对整个数据库中的所有资源进行分类展示。

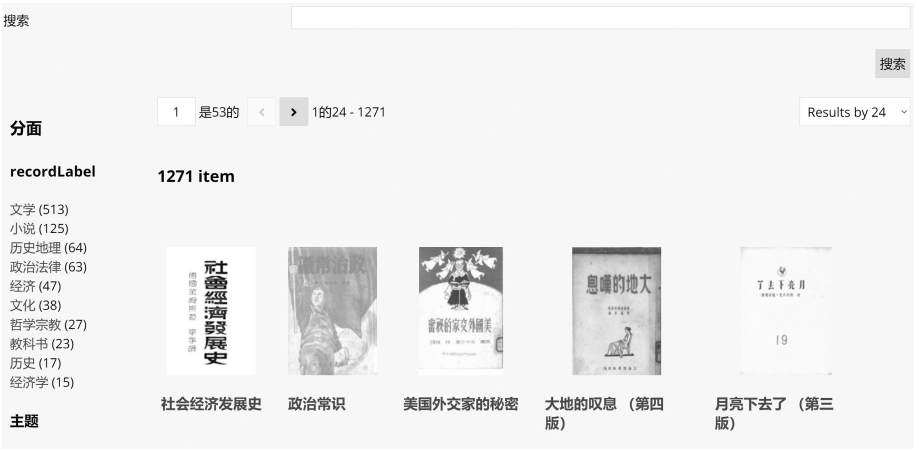

图2　近代中译本全文特藏库分类检索

(3)元数据聚合浏览

由于每一本图书都有多种标签,无法通过分类浏览方式全部展示,而将这些聚合字段嵌入到每一本图书的详细页面中,用户就可以通过详细页面的聚合字段浏览具有相同属性的全部资源,这是用户查找同类资源的有力线索。

(4)时空浏览

研究团队在加工近代中译本元数据时,定义了每一本图书的创作时间与出版地,在执行导入数据之前已经批量将每一个出版地转化为相应的经纬度坐标,这样每一本近代中译本就可以在地图上展示。地图时空浏览用到了Omeka-S的Mapping模块,配置需要显示在地图中的资源,设置时间线字段、默认缩放比例等,就完成了近代中译本时空检索配置,可视化效果如图3所示。在客户端,用户可以设置时间线的间隔大小以获得最佳的浏览体验,点击时间线中的图书名称,会自动打开这本图书的简介与处于地图上的位置。同时研究团队还对不同的设备做了适配,在移动设备上可以进行触摸操作。

图3　近代中译本全文特藏库时空浏览

3.5 在线全文浏览、对比阅读与众包

本研究根据近代中译本的特点,选择开源软件 Mirador 作为资源在线浏览工具①,它可以对图像资源进行深度缩放、自适应访问终端、更改图像颜色以及注释等操作。其中最为重要的功能是对比阅读,凡是符合 IIIF 标准的特藏资源,都可以直接拖拽到 Mirador 浏览器中打开,即使这些资源不在本地译本数据库中。比如《共产党宣言》对比阅读界面(见图 4),右边的《共产党宣言》收藏于本地近代中译本全文特藏库,是 1950 年《民国丛书》编辑委员会编著的近代中译本,而左边则是德国慕尼黑巴伐利亚国家图书馆收藏的原本德文版《共产党宣言》[18]。图 4 仅显示了两个版本,通过左上角的加号,还可以添加更多的版本进行对比,比如还可以添加斯坦福大学图书馆收藏的延安解放社 1943 年版本的《共产党宣言》等。在对比阅读时,可以将 Mirador 全屏以获取更好的阅读体验。

图 4 不同来源、不同版本《共产党宣言》对比阅读

本研究提供两种众包方式,一种是无需登录,直接在数字对象简介页面进行标签操作,管理员在后台审核之后,这些标签就可以展示出来,这种方式使用了 Omeka-S 的 Folksonomy 插件。另外一种是基于 Mirador 浏览器的注释功能,允许用户在数字对象上直接标注、转录,使用该功能需要用户登录。

3.6 开放共享

(1)关联数据发布

元数据词表已经基于国际通用本体构建,使近代中译本的语义发布成为可能。关联数据的发布采用 Json-LD 数据格式,在近代中译本全文数据库的前台页面中嵌入每一本图书的 Json-LD 格式数据,使每一本图书不仅可以供读者在线阅读,还可以被机器理解,成为全球语义网的一部分。

(2)图像语义

本研究采用符合 IIIF 框架标准的开源软件 IIPImage Server 作为数字对象的服务器②,提供 IIIF 图像接口和展示接口。在展示接口中同样是以 Json-LD 发布每一个数字对象的标签、注释以及内容等,关联数据的介入使近代中译本全文特藏库中的每一张图片成为可以被机器所理解的语义资源。

4 启发与思考

4.1 数字人文是一个宝藏库

数字人文是伴随着计算机技术的发展而产生、发展与成熟的,目前已经积累了大量的开源软

① Mirador. https://projectmirador.org.

② IIPImage Server. https://iipimage.sourceforge.io.

件,成为数字特藏建设的基础设施,比如本研究采用的 DjVuToy 图像批处理工具、内容发布工具 Omeka-S、数字对象发布与浏览工具 IIPImage Server 与 Mirador 等。很多开源软件不仅仅是数字对象的加工管理工具,也是数字特藏资源加工流程与方法的集中体现,比如德国的 Goobi 工具,以工作流的形式提供数字特藏多机构合作加工框架,然后通过在线发布平台统一发布。这些开源软件不仅为没有数字特藏开发经验的图书馆提供发布工具,还可以帮助具有数字特藏开发经验的图书馆改进开发模式。近代中译本全文特藏库的开发过程完全采用开源软件实现,其功能已经远超国内自建数据库厂商所能提供的特藏资源管理功能,而且在资源开放层面更具主动性。

数字人文除了包含大量的开源软件,其开放社区也具有重要价值,通过社区可与全球数字人文研究人员交流,迅速解决数字特藏建设过程中遇到的各类问题。IIIF 是当前影响力较大的一个图像互操作开源社区,一般提供多种加入渠道,比如邮件列表、讨论组、在线社区、日历以及社交网络等,在国内通过 B 站与微信群的形式提供交流平台。近代中译本全文特藏库建设过程中,获得了包括 IIIF、Omeka、Mirador 以及 GitHub 等社区的支持。

4.2 数字人文可以重新激活特藏资源

(1)整合异构资源

图书馆在长期的数字资源建设中积累了大量的异构数据,即使在本馆内部,这些数据库之间也无法有效地实现互操作,形成了一个个信息孤岛,从而阻碍了资源的有效利用。近代中译本全文特藏库试图通过构建元数据本体、数据的关联发布以及图像语义等措施,从底层解决特藏资源孤岛问题。本研究还实现了以 OAI-PMH 框架发布资源,可对接 Fedora、Dspace 等系统的接口,最大限度保证系统的开放性。对于已经建成的数字特藏资源,数字人文也提供了大量的中间件以便异构资源的互操作,比如利用 D2RQ 工具① 可以将现有的关系型数据库以关联数据发布;对于数字对象本身的开放共享,则可通过安装 Loris② 与 SIPI③ 等图像服务器将当前的数字对象转化为支持 IIIF 标准的开放图像。

(2)关联数据与知识图谱

特藏资源中包含着丰富的潜在知识,然而仅仅基于扫描获得的资源格式,无法直接对其所包含的知识进行揭示。近代中译本全文特藏库提供了人工智能自动识别与用户标注两种方式来挖掘数字对象中的潜在知识,未来还将充分利用已经发布的关联数据以及更大范围的 LOD-cloud④、DBpedia⑤ 等关联数据资源构建近代中译本知识图谱,以更加深入地分析近代中译本资源。

4.3 众包激发用户智力

高校图书馆面向的都是具有较高文化素质的年轻人,他们是社交网络的原住民,完全能够胜任在特藏资源中添加标签、注释以及转录等众包操作。数字人文也为用户参与特藏资源众包提供了多种解决方案,比如注释服务器 Annotot⑥ 与 E-lucidate⑦ 等。近代中译本全文特藏库已经实现了标签、转录、注释等众包功能,接下来将探索可充分调动用户参与特藏资源众包的方法,比如将读者校园卡号与近代中译本全文特藏库对接,从而可对参与众包的读者给予一定的激励。众包还可以与教学相结合,由教师将近代中译本的标注作为学生的课后作业,不仅可以丰富作业形式,同时也丰富了近代中译本特藏库的元数据。

① D2RQ. http://d2rq.org.
② Loris. https://github.com/loris-imageserver/loris.
③ SIPI. https://sipi.io/.
④ LOD-Cloud. https://lod-cloud.net.
⑤ DBpedia. https://www.dbpedia.org.
⑥ Annotot. https://github.com/PenguinParadigm/annotot.
⑦ Elucidate. https://github.com/dlcs/elucidate-server.

4.4 构建本地数字人文基础设施非常必要

数字人文基础设施可为人文学科研究提供文献、数据以及工具等整体解决方案,使人文学者专注于研究,而不必关心技术实现,从而降低数字特藏研究的门槛,提升研究效率。国内高校在数字人文基础设施建设方面已经有不少实践,比如复旦大学的中国历史地理信息平台[19]、浙江大学的学术地图发布平台[20]等。近代中译本全文特藏库是基于华东师范大学图书馆数字特藏发布基础设施构建的,节省了服务器部署、应用软件安装与系统架构设计等工作。此外,基础设施还提供了资源的批量导入导出、文字识别、自然语言处理、时空检索、众包等功能。因此,近代中译本全文特藏库的构建不需要耗费精力在平台建设上,只需关注元数据与数字对象的整理加工。

5 总结

媒介即信息,数字媒介取代纸质媒介导致信息的内容、组织以及传播都发生巨大改变,给作为知识中心的图书馆带来挑战的同时,也为图书馆打开了数字人文这扇大门。数字人文为知识赋予了新的组织形式,在新的形式中创造着新的信息与内容,为研究者提供了全新的视角和思路。特藏资源作为图书馆馆藏的独特部分,借助数字人文的研究工具和研究方法,得以更深入地揭示与呈现,为人文研究者打开新天地。

特藏资源的高价值与独特性是信息同质化时代图书馆服务创新的重要基础,结合数字人文范式与开源软件对特藏资源进行获取、保存、管理与揭示,是图书馆利用特藏资源服务读者的有效手段。本研究调查了全球前100高校的图书馆的数字特藏库,发现数字特藏的可视化、OCR、自然语言处理以及注释转录等数字人文方式被大量使用,这也是图书馆适应资源媒介改变做出的调整。本研究开展的华东师范大学近代中译本全文特藏库建设实践使用Omeka-S进行资源发布,采用自然语言处理技术对内容进行审核、标签与分类,提供资源分类浏览、时空可视化、检索建议、元数据聚合等揭示方法,以及在线标注功能来整合读者智力资源;利用Mirador工具满足读者对数字资源的高清浏览需求以及与不同来源资源的对比阅读需求,而且可以适应不同手机、平板、电脑以及触摸阅读大屏等终端。近代中译本特藏库底层数据以关联数据发布,将Json-LD格式的关联数据嵌入到每一本近代中译本中,使得近代中译本不仅可以被读者在线浏览,而且也可以被机器理解,成为全球语义网的一部分。本研究是图书馆借助数字人文理念与技术,进行数字特藏资源建设的一次尝试,未来会将整个系统的源代码打包共享,以期为需要特藏数据发布的图书馆提供借鉴。

参考文献

1 王乐.略论高校图书馆特色馆藏建设的价值与发展方向[J].大学图书馆学报,2020(3):12-17.

2 蔡迎春.数字人文视域下的图书馆特藏资源数字化建设——以“民国时期文献目录数据平台”为例[J].图书馆建设,2018(7):31-36,41.

3 韩冰.“双一流”高校图书馆自建特色数据库调研与思考[J].图书馆工作与研究,2020(10):84-88.

4 蔡梦玲.基于贝叶经信息资源的特色数据库建设[J].图书馆杂志,2021(2):111-116.

5 李宗富,王亿豪.俄罗斯装饰与应用艺术数字档案馆建设特色及启示[J].档案管理,2021(3):117-119.

6 González-Blanco E,et al. EVI-LINHD,a virtual research environment for the Spanish-speaking community[J]. Digital Scholarship in the Humanities,2017,32:171-178.

7 Digital humanities[EB/OL].[2021-09-20].https://en.wikipedia.org/wiki/Digital_humanities.

8 柯平,宫平.数字人文研究演化路径与热点领域

分析[J]. 中国图书馆学报,2016(6):13-30.
9 石志松. 欧洲研究图书馆协会数字人文发展策略探析[J]. 大学图书馆学报,2019(5):24-31.
10 张毅,陈丹. 全球100所知名高校图书馆特藏资源调查与分析[J/OL]. [2022-10-13]. 图书馆杂志:1-13. http://kns.cnki.net/kcms/detail/31.1108.G2.20220517.1740.004.html.
11 scomcat[EB/OL]. [2021-12-18]. https://www.scomcat.net.
12 DARTH[EB/OL]. [2022-01-29]. https://digitalhumanities.fas.harvard.edu/resources/choosing-digital-methods-and-tools.
13 ACDH-CH[EB/OL]. [2021-07-29]. https://www.oeaw.ac.at/acdh/tools.
14 Brown University Library[EB/OL]. [2021-09-12]. https://libguides.brown.edu/Digital Tools.
15 Bountouri L, et al. Digital Preservation: How to Be Trustworthy[J]. Digital Cultural Heritage, 2018,10605:364-374.
16 El-Fakdi A, de la Rosa J L. Analysis of Nature-Inspired Algorithms for Long-Term Digital Preservation[J/OL]. [2022-01-05]. Mathematics,2021,9(18),2279. https://www.mdpi.com/2227-7390/9/18/2279.
17 胡晓进. 商务印书馆与美国宪法在中国大陆之翻译及传播[J]. 政法论坛,2019(2):70-79.
18 Marx, Karl: Manifest der Kommunistischen Partei[EB/OL]. [2021-11-26]. https://www.digitale-sammlungen.de/en/view/bsb10859626.
19 中国历史地理信息平台[EB/OL]. [2021-08-06]. http://timespace-china.fudan.edu.cn.
20 学术地图发布平台[EB/OL]. [2021-12-23]. http://amap.zju.edu.cn.

(张 毅 副研究馆员 华东师范大学图书馆,赵晨鸣 馆员 华东师范大学图书馆,陈 丹 副研究馆员 华东师范大学图书馆)

收稿日期:2022-03-12

《公共图书馆系统古籍类文物定级指南》正式印发

2022年12月19日,文化和旅游部办公厅、国家文物局办公室印发《公共图书馆系统古籍类文物定级指南》,明确公共图书馆领域古籍类文物定级的总体要求、工作依据、工作方法等相关规范及要求,适用于各级公共图书馆和国家图书馆馆藏普通形制汉文古籍类文物定级工作。

资料来源

文化和旅游部办公厅 国家文物局办公室关于印发《公共图书馆系统古籍类文物定级指南》的通知[EB/OL]. [2023-01-19]. https://zwgk.mct.gov.cn/zfxxgkml/zcfg/gfxwj/202212/t20221226_938275.html.

(国家图书馆研究院 提供)