数字人文视野下西文古籍数据库的研发*

□张毅

摘要 近年来，虽然国内图书馆界对西文古籍的研究不断增加，但大多数研究仅限于馆藏调查、版本分析和文献修复，缺乏关于西文古籍数字资源的研究；而实践方面，还是以纸本借阅服务为主，无法满足读者在线阅览西文古籍的需求。文章以读者需求为中心，在分析国内外西文古籍数据库建设现状的基础上，总结了西文古籍数据库的建设思路，并以华东师范大学图书馆西文古籍数据库建设为例，介绍了其采用开源软件和数字人文技术进行西文古籍数据库开发的过程和经验，可供同行参考借鉴。

关键词 数字人文 西文古籍 开源软件 Omeka-S Open Semantic Search

分类号 G250.74

DOI 10.16603/j.issn1002－1027.2023.02.008

国内馆藏的西文古籍是研究西方思想文化和中西交流史的一手资料，具有重要的历史和学术价值。然而目前国内各馆西文古籍文献的保存和服务的状况却不尽如人意：对于闭架保存的善本西文古籍，读者需要经过预约等复杂手续才能阅览；对于非善本西文古籍，由于缺少专业的保管，存在着破损严重的情况，以上这些问题都不利于西文古籍价值的发挥[1]。

随着数字时代的到来，数字化成为解决纸质西文古籍保存和服务问题的有效手段，本文在分析国内外西文古籍数据库建设现状的基础上，总结了西文古籍数据库的建设思路，并以华东师范大学图书馆西文古籍数据库建设为例，介绍了该数据库的全文高清浏览、可视化、全文检索、分类浏览、在线标注以及知识图谱等等多方面的功能。借助数字化和数字人文等多方面的技术，西文古籍得以重新焕发出历史与文化的魅力，可以被更广泛地传播与利用。

1 西文古籍数据库建设情况调查

1.1 国内的情况

目前，国内图书馆界对于西文古籍的出版时间范围的界定存在着不同的看法[2－4]，但在实践中，通常将1911年之前出版的西文图书视为西文古籍，将1800年之前出版的西文图书归为西文善本。

1.1.1 国内公共图书馆

国内公共图书馆收藏的西文古籍主要来源于晚清民国时期的教会图书馆或者个人收藏者的捐赠。2022年4月，通过在搜索引擎和国内各级公共图书馆的网站以及目录系统中，检索“西文古籍”“旧版西文图书”“外文古籍”等关键字，可查到许多收藏了西文古籍的图书馆。其中收藏最为丰富的是国家图书馆与上海图书馆，均超过5万册。国家图书馆收藏的西文古籍质量最高，大多属于善本，并且建立了独立的西文古籍展示网站①，上海图书馆②和大连市图书馆③专门针对西文古籍开发了独立的书目系统，澳门公共图书馆则有专门推荐西文古籍的网页，提供了部分西文古籍的检索和介绍④。

1.1.2 国内高校图书馆

2022年5月，以检索式“西文古籍 site：*.edu.cn”在必应搜索引擎中检索与高校图书馆有关的西文古籍数据库，发现北京大学图书馆、中山大学图书

* 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”（编号：21BTQ100）的研究成果之一。
张毅，ORCID：0000-0002-0173-6103，邮箱：yizhang@library.ecnu.edu.cn。

① http://www.nlc.cn/nmcb/gcjpdz/xwsb.

② http://search.library.sh.cn/jiuxiwen.

③ http://www.dl-library.net.cn/book/list.php? id=5.

④ https://www.library.gov.mo/zh-hans/library-collections/special-collections/ancient-texts.

馆、厦门大学图书馆和河北大学图书馆等都有馆藏西文古籍的介绍，但未见相关数据库的介绍。在中国知网期刊数据库中检索到，于燕妮总结了中国人民大学图书馆对馆藏的2450册西文古籍进行数字化加工和著录的经验，但未见相关专题数据库建设的说明[5]。此外，北京师范大学图书馆的晚清民国教材全文库零散收录了与教科书相关的西文古籍，但在校外不能访问全文。

1.2 西文母语地区的情况

对西文母语地区的西文古籍数据库建设的调查以高校图书馆为主，笔者于2022年4月，对U.S. News全球高校排名前100的英国、美国、澳大利亚以及德国等国家的高校图书馆进行了调查。有57所高校的数字图书馆中有专门的西文古籍集合，其命名一般为善本集合(Rare Book Collection)[6]，尽管部分善本集合是所在高校数字图书馆平台的子网站，但一般也具备数据库主页、检索框、分类等独立的网站功能，所以本研究也将其作为西文古籍数据库处理。通过对这些西文古籍数据库的详细分析，发现注重用户体验和开放共享，以及数字人文工具的应用是其主要特点，具体如下：

1.2.1 普遍采用IIIF技术

西文母语地区高校的数字图书馆普遍采用了国际图像互操作框架(International Image Interoperability Framework，IIIF)技术发布高清数字对象。IIIF具有图像动态加载功能，可以根据终端屏幕尺寸大小，为读者提供图像的最佳分辨率。如牛津大学博德利数字图书馆收藏的意大利语古籍*Entomologia Britannica*一书共有588页①，每一页的尺寸为3830*5327像素，整本书的存储空间超过1GB；剑桥大学数字图书馆收藏的西文古籍*Gospel Lectionary*(*Saturdays*，*Sundays and Weekdays*)有356页②，每页图像尺寸为1288*2000像素，整本书需要700M存储空间。二者均采用了IIIF的动态加载技术，读者可在低延迟下获得最佳分辨率的浏览体验。此外，采用IIIF技术不仅可以实现西文古籍数字对象的高清在线浏览，而且还可赋予数字对象开放共享的能力。

1.2.2 对外提供编程接口

被调研的部分西文古籍数据库采用了标准的资源描述本体，对外提供数据编程接口，使其成为整个社会数据基础设施的一部分。如哈佛大学图书馆通过应用编程接口开放其西文古籍元数据与部分全文的光学字符识别(Optical Character Recognition，OCR)数据[7]，共计49589册。同样提供编程接口的数字图书馆还有牛津大学博德利数字图书馆中的西文古籍集合[8]。

1.2.3 使用开源软件，并作为开源软件贡献者

西文母语地区的高校图书馆在构建西文古籍数据库时，采用了大量的开源软件，例如，斯坦福大学西文古籍数据库的后台系统，使用Solr进行数据索引，使用Blacklight实现分页浏览[9]。包含大量西文古籍的剑桥大学数字图书馆使用Bootstrap与JQuery构建响应式页面，使用OpenSeadragon作为IIIF图像查看器[10]。剑桥大学、美国西北大学等学校的数字图书馆还将自己的源代码提交到GitHub共享[11-12]，供个人与组织下载使用。曼彻斯特大学图书馆在剑桥大学的帮助下，利用开源的剑桥大学数字图书馆系统构建了曼彻斯特数字馆藏库，并收藏有大量西文古籍[13]。

1.2.4 全文检索

由于古文字与印刷质量的问题，目前基于现代英语的机器学习技术在西文古籍文字自动识别方面仍存在一些困难。但是，对西文古籍进行全文文字识别已经成为一种趋势，也是读者呼声较高的功能。例如，牛津大学数字图书馆已经对部分图书进行了全文文字识别，并计划未来逐渐转录所有的数字馆藏[14]，哈佛大学图书馆于2019年，开发了针对所有数字馆藏的全文检索工具[15]，南安普顿大学数字图书馆提供全文检索，而且可以定位检索结果到章节[16]。

1.2.5 可视化

可视化能够将抽象的事物用生动的形式展示，为人文学者提供一种全新的研究工具，例如瑞士苏黎世联邦理工学院的西文古籍数据库，采用地图方式展示古籍图书的出版地分布情况[17]，不列颠哥伦比亚大学西文古籍数据库则以时间线的形式进行资源揭示[18]，能够清晰地在时间尺度上对西文古籍的数量与作品类型进行分析。

① https://digital.bodleian.ox.ac.uk/objects/552a61bc-e238-452d-af4d-02aaaf05bdeb.

② https://cudl.lib.cam.ac.uk/view/MS-FITZWILLIAM-MCCLEAN-00004.

2 研究思路

调查结果显示，西文母语地区特别重视西文古籍数据库建设，以最大程度地开放共享为建设理念，依托开源软件构建多种数字人文工具。随着人工智能技术的不断发展，西文古籍全文 OCR 与实体识别技术逐渐得到推广，使得全文检索成为可能。国内图书馆界也开始重视西文古籍的重要价值，但国内的西文古籍全文数据库建设还停留在理论研究阶段。本研究将借鉴西文母语地区建设西文古籍数据库的经验，以读者需求和学科发展为导向，探索构建西文古籍数据库。

2.1 西文古籍数据库应具备友好的用户体验

2.1.1 自适应多种访问终端

第 50 次《中国互联网络发展状况统计报告》显示[19]，截至 2022 年 6 月，国内手机接入互联网的比例已达到 99.6%，超过了台式电脑、笔记本、平板电脑的总和。西文古籍全文数据库平台可采用响应式网页设计方式，自适应手机、电脑、平板等多种访问终端，满足读者多元化的访问需求。

2.1.2 优化页面布局与提高响应速度

2019 年，华东师范大学图书馆在全校范围内对图书馆主页改版的需求进行了调查，读者反馈意见最多的是图书馆主页内容繁杂，响应速度慢。西文古籍数据库也可以借鉴这一调查结果，页面设计以简洁为主，并根据用户使用反馈不断优化。系统的响应速度决定着用户的留存，在设计西文古籍全文数据库时，可采用动态加载与异步通讯等措施确保响应速度。

2.1.3 提高搜索引擎的收录

当前，搜索引擎依然是用户获取信息的重要入口，将 Json-LD 嵌入到西文古籍数据库的网页中，可使得资源更容易被搜索引擎所收录。增加西文古籍揭示平台与图书馆主页、电子资源导航以及学校主页之间的超链接，也能有效提高搜索引擎的收录量。

2.2 数字人文技术的应用

仅仅将西文古籍数字化并在线发布，仍属于传统纸质资源服务模式的简单升级，不能有效释放西文古籍的独特价值。数字人文是数字技术与人文学科的交叉领域，由大量开源工具组成的数字人文软件基础设施，可赋予西文古籍全文数据库更多的功能（如可视化、众包以及文本挖掘等）[20]，能够协助人文学者挖掘出西文古籍所蕴藏的潜在知识。

2.2.1 可视化

数字人文常用的可视化方法有图表、关系网络、地图、时间线等，其中地理信息系统（GIS）是比较成熟的数字人文研究工具，结合时间变量，可为人文学者提供时空层面的内容揭示。

2.2.2 众包

众包可以有效解决西文古籍能见度低的问题，W3C 于 2004 年专门成立了 Web 注释工作组（Web Annotation Working Group），并于 2018 年发表了 Web 注释数据模型、词汇表及注释协议等三份正式推荐标准[21]，这三份标准的发布，标志着数字资源众包时代的到来。

2.2.3 文本挖掘

在西文母语世界中，已经形成了大量西文古籍的语料库，比如维基百科开放数据、哈佛大学数字图书馆云等。利用这些成熟的语料库资源，结合机器学习算法，可以精准地对西文古籍进行文本挖掘，更加深入地揭示其所蕴藏的知识。

2.3 采用开源软件

Islandora、Samvera、Omeka、Goobi 以及剑桥数字馆藏平台等系统，是较为主流的数字资源管理平台，它们的底层架构也多基于开源软件构建，其中包括 Mysql 和 PostgreSQL 等关系型数据库实现的元数据存储，Solr 和 Blacklight 等工具实现资源的发现，Bootstrap 和 JQuery 等技术进行的用户界面开发，以及基于 LeafLet 框架构建的时空可视化展示等功能。在图像处理方面，这些平台大多采用 ImageMagic 进行图像处理，并借助 Mirador、Universal Viewer 等工具实现符合 IIIF 标准的图像在线浏览，同时还利用 Loris、Cantaloupe 以及 IIPImage Server 等图像服务器工具发布可动态加载的图像。

2.4 利用商业人工智能平台

已经较为成熟的商业人工智能平台，能够对西文古籍进行自动分类、标签抽取、内容审核以及图像识别等分析，其分析结果经图书馆校验后，可作为西文古籍元数据的补充，提升西文古籍的能见度；导入相关专业领域的语料库还能进一步提升人工智能平台识别的准确度。

2.5 整合西文母语世界的同类型资源

西文母语世界中已经有大量开放的西文古籍资源，对于其中以 IIIF 格式发布的相关资源，可以将其整合到本地数据库来丰富本地资源，对于以关联

数据形式发布的数据集，则可用来对本地西文古籍的元数据进行校验与丰富。

3 西文古籍全文数据库的建设实践

以华东师范大学图书馆西文古籍资源为对象，基于开源软件与云开放平台构建西文古籍全文数据库。

3.1 西文古籍文献详情

华东师范大学图书馆（以下简称华东师大馆）特别重视西文古籍资源的数字化，目前，已有超过2000册的西文古籍被数字化。这些西文古籍在出版时间上的分布如图1所示，可以看出，华东师大馆所收藏的西文古籍出版时间主要集中在1890年到1911年之间，这段时间也是我国西学东渐的开始，其出版社与出版城市集中分布信息如表1所示，其中，上海是国内出版西文古籍较多的城市。

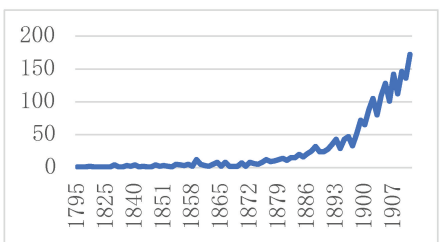

图1 西文古籍出版时间分布

表1 西文古籍出版社与城市信息

出版社	图书数量	城市	图书数量
Charles Scribner's Sons	98	London	792
Houghton Mifflin Company	77	New York	716
The Macmillan Company	74	Boston	243
Longmans, Green, and CO.	66	Oxford	54
At the Clarendon Press	47	Chicago	50
Macmillan and CO.	39	Philadelphia	45
G. P. Putnam's Sons	38	Edinburgh	38
D. Appleton and Company	36	Washington	33
Houghton, Mifflin and Company	36	Cambridge	32
Macmillan and CO., Limited	36	Leipzig	25
American Book Company	34	Shanghai	23
Henry Holt and Company	29	Glasgow	10

虽然华东师大馆已经拥有大量数字化的西文古籍，然而长期以来这些数字西文古籍还没有被有效地揭示，不能充分发挥出这些珍贵收藏的价值。

3.2 西文古籍数字资源的管理与发布

3.2.1 构建西文古籍全文数据库系统

通过对众多开源数字资源管理系统的分析，华东师大馆最终选择Omeka-S来构建西文古籍数据库。Omeka-S具有清晰的文献管理与发布逻辑，系统面向语义网开发，底层数据采用关联数据组织，内置多种元数据本体，其开源社区中有丰富的扩展模块，符合开箱即用系统的要求。图2是采用Omeka-S发布的西文古籍全文数据库的首页，该数据库的Omeka-S系统运行在Centos7.5的虚拟机上，服务器的配置（内存32G，16核CPU）可以基本满足图像处理与数据发布的需求。

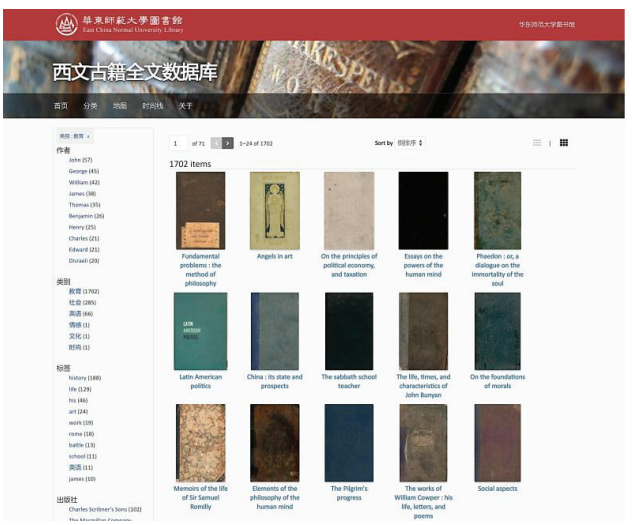

图2 华东师大馆西文古籍全文数据库的首页

3.2.2 批量导入西文古籍全文数据

使用CSV Import插件能够将本地数据批量导入到西文古籍数据库，并可调用ImageMagick软件自动生成所需的缩略图。采用弗吉尼亚大学图书馆开发的档案汇编（Archive Repertory）插件，可让导入的数据以原始路径与文件名存储于服务器，提高了数据库的访问速度，也方便了后续批量更换文件。

3.2.3 全文在线浏览

西文古籍全文浏览功能基于IIIF接口开发，由服务器端与浏览器端程序组成，服务器端采用IIP-Image Server，提供图像的动态加载功能，浏览器端采用Universal Viewer工具，可让读者在不同访问设备上均获得最佳的体验。

3.2.4 整合开放资源

IIIF技术在西文母语地区的广泛使用，使得大量开放的西文古籍资源可以无缝嵌入到本地数据库

中，还可以对西文古籍不同版本进行对比阅读。例如，本地西文古籍数据库仅收录了莎士比亚作品 *The Plays and Poems of William Shakespeare* 的5册中的1册，严重影响了读者的使用。通过嵌入德国巴伐利亚国家图书馆的5册 *The Plays and Poems of William Shakespeare*[①]，以及耶鲁大学图书馆3册带批注和插图的版本[②]，补充了本地馆藏的不足。整合外部资源到本地，需要大量的手工查找，可尝试将西文古籍数据库的建设与学校的教学过程相结合，利用西文古籍的稀缺性与特殊性来激发读者参与资源搜集和整理的积极性，从而解决需要大量手工查找资源的难题。

3.3 资源揭示

3.3.1 人工智能实现文献审核、分类和标签抽取

西文古籍在数字化时包含的元数据内容较少，并且还可能包含价值观不正确的文献。重新组织人力对西文古籍文本进行整理的成本太高，得益于国内成熟的人工智能基础设施环境，可以通过调用人工智能开放平台接口，对西文古籍文献的题名、描述与目录字段进行自然言语处理，生成每一本西文古籍图书的分类、标签以及审核结果。经过测试，百度人工智能比较符合本研究的需求，以莎士比亚的作品 *A Midsummer-night's Dream* 为例，利用百度人工智能平台进行分析，调用接口进行自动分类的C#代码如下：

第一步：设置APPID/AK/SK，并实例化接口：

```
1.var APP_ID = "App ID";
2.var API_KEY = "Api Key";
3.var SECRET_KEY = " Secret Key";
4.var client =new Baidu.Aip.Nlp.Nlp(API_KEY, SECRET_KEY);
5.client.Timeout = 60000;
```

第二步：分析莎士比亚这部作品类型的方法函数：

```
1.public void KeywordDemo() {
2.var title = " A midsummer-night's dream ";
3.var content = " Magic, love spells, and an enchanted wood provide the materials for one of Shakespeare's most delightful comedies. * * * * * *, all touched by Shakespeare's inimitable vision of the intriguing relationship between art and life, dreams and the waking world. ";
4.var result = client.Topic (title, content);
5.Console.WriteLine(result);
6.}
```

经过运算后，百度人工智能开放平台返回的分类结果是“社会、戏剧”，同样的调用过程，只需改变调用函数，就可以分析出文献的标签与内容审核结果。

3.3.2 构建时空浏览

西文古籍的出版地与出版时间是进行时空浏览的数据基础，利用开源软件 LeafLet 可构建资源的时空浏览。LeafLet 是一个地图操作库，拥有功能丰富的地图插件，其中 MapBox 插件可以自定义需要的地图，TimeLine 插件用来设置时间线样式。由于 LeafLet 只支持经纬度来定位文献在地图上的位置，所以需要事先将地址转为经纬度。浏览时，拖动时间线就可以查看对应的文献，选中某个文献，地图会自动切换到文献的出版地，LeafLet 是专门针对移动端设计的应用，所以能够完美匹配手机等移动终端浏览。

3.3.3 利用自带OCR功能的Open Semantic Search实现全文检索

Open Semantic Search 是专门用于对大型文档和图像等内容进行文本挖掘的开源工具[③]，它利用开源文字识别软件 Tesseract-OCR 抽取 PDF 文档或图像中的文字，其在 GitHub 上的 Fork(8400 次)与 Star(49200 次)都很高，是 OCR 领域权威的开源工具之一。在对西文古籍进行全文 OCR 之后，Open Semantic Search 会利用开源机器学习框架 SpaCy 自动识别出文献中的人物、地点、日期、数字、组织等信息，然后提供分类浏览，文献浏览界面如图3所示，识别的准确度需要用户进行验证与调试，对于识别出的实体，利用 Apache Solr 进行索引，以提高检索效率。图3中，每一本西文古籍都进行了人物、机构、位置等信息识别，并且可以按照右侧分类进行筛选，对华东师范大学图书馆西文古籍的全文分析发现，相关人物有莎士比亚、达尔文等，涉及较多的地点有英国、美国、欧洲等。

① https://www.digitale-sammlungen.de/en/search? query=all%3A%28The+plays+and+poems+of+William+Shakespeare%29.
② https://collections.library.yale.edu/catalog/10009029.
③ https://www.opensemanticsearch.org.

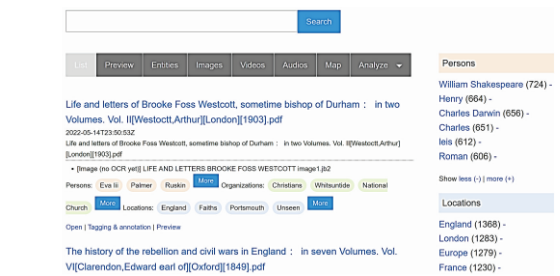

图 3　Open Semantic Search 对西文古籍全文分析的结果

3.4　古英语翻译

西文古籍中有部分文字是“古英语”，和现代英语有比较大的差距，会给读者阅读西文古籍带来困扰，比如莎士比亚的 *The Plays and Poems of William Shakespeare* 一书中，“you”常常被写作“thou”，“has”被写成“hath”，“says”被写成“saith”等。为了便于读者阅读，西文古籍数据库除了将常用的“古英语”单词放在帮助文档中外，还嵌入古英语翻译工具①。

3.5　开放共享

3.5.1　元数据层面的开放

西文古籍全文数据库的系统底层基于关联数据开发，采用都柏林核心、书目框架、Schema.org 三种本体构建元数据词表，并以 Json-LD 格式嵌入每一本西文古籍浏览页面，可以有效被机器理解，使得西文古籍在互联网上具有较高的可见度。除了采用关联数据发布数据外，系统还提供 CSV 格式导出、Restful 方式的编程接口以及 OAI-PMH 的数据收割方式。

3.5.2　数字对象开放共享

在实现西文古籍自适应不同终端浏览时，西文古籍数据库系统使用 IIIF 框架发布数据。IIIF 本身就是为互联网图像数据互操作而生，只要是以 IIIF 发布的数字对象就是开放的，获取数字对象的 IIIF Manifest 连接，就可以无缝地将数字对象嵌入到本地系统，通过符合 IIIF 接口标准的图像浏览器可在线浏览，或者本地缓存后对外发布。

3.6　众包

在进行西文古籍手工编目时，图书馆常常面临人力不足以及领域专家缺乏的问题，为此，采用众包方式可以充分发挥校内师生的智力资源，为西文古籍添加标签与注释数据，提升西文古籍的能见度。利用 Folksonomy 开源插件可以让读者为西文古籍添加标签，标签本身还具有文献聚类功能。前文提到的整合外部不同版本西文古籍到本地系统，也需要用到众包功能，通过轻型目录访问协议（LDAP）功能，西文古籍数据库允许读者使用校园一卡通账号登录，并提交自己整理的数据。

4　思考与展望

4.1　共享西文古籍数据库源代码

华东师大馆的西文古籍数据库已经在校内上线，并且得到读者的认可。虽然该平台是基于开源框架开发，但多种开源软件的协作配合、性能优化，以及部分自主开发等，也需要大量的配置与编码工作，比如：(1)将系统使用的谷歌字体、JS、CSS 等在线资源下载到本地，用以提高系统响应速度；(2)对系统自带的 Papers 模板、分面检索插件、时空浏览插件进行二次开发，以满足西文古籍数据库的需要；(3)解决新版本 Omeka-S 系统的 Mirador 插件因为缺少开源社区维护而无法使用的问题；(4)修复 Universal Viewer 插件侧边栏不显示缩略图问题。

这样的西文古籍数据库，对于缺少技术储备的图书馆来说，自己部署也有不小的挑战。所以本研究除了共享整个系统的源代码外，未来还计划将西文古籍数据库系统以 Docker 容器和 Vagrant 形式发布为完整的虚拟机（包含源代码的虚拟机，免去了部署源代码、配置服务器的过程），这样，想要采用或者试用本平台的图书馆，只需要简短的几行命令，就能够将整个测试环境部署在本地，来体验完整的西文古籍数据库系统。

4.2　对 PDF 中的图像进行识别

由于需要大量的计算资源，华东师大馆的西文古籍全文数据库仅识别了文献中的文字，并未对文献中的图像进行文字识别，在全文检索时，还无法查找到图片中的文字内容。当前，在没有额外硬件计算资源加入的情况下，需要较长的时间才能完成 163.7 万页西文古籍图书中图像的逐个识别。

4.3　借助领域词表进行文本挖掘

借助领域语料库能够有效提升文本分析的准确度，以西文古籍数据库中莎士比亚的作品为例，

① https://www.oldenglishtranslator.co.uk.

Wiki-Data 中包含莎士比亚的数据集[①]，编号为 Q692，利用 Open Semantic Search 提供的 SPARQL 查询端口构建莎士比亚数据集的 SPARQL 查询语句，从而将 Wiki-Data 中的数据集导入 Open Semantic Search，进行有针对性的文本挖掘，莎士比亚数据库 SPARQL 查询语句如下：

```
1. PREFIXskos: < http://www.w3.org/2004/02/skos/core#>
2.CONSTRUCT {
3.    ? uri rdfs:label ? label;
4.skos:prefLabel ? prefLabel;
5.skos:altLabel ? altLabel.
6.}
7.WHERE {
8.    ? uri wdt:P39 wd:Q692. //Wiki-Data 中莎士比亚数据的编号
9.    OPTIONAL {
10.        ? uri rdfs:label ? label .
11.    }
12.    OPTIONAL {
13.        ? uri skos:prefLabel ? prefLabel.
14.    }
15.    OPTIONAL {
16.        ? uri skos:altLabel ? altLabel .
17.    }
18.}
```

4.4 完善知识图谱

将 Open Semantic Search 识别出的实体导入到 Neo4j 图数据库中，可构建简单的知识图谱，例如在知识图谱中查询“达尔文”，结果如图 4 所示。但由于缺少实体之间的关系数据，目前还不能完全发挥出知识图谱的强大功能，接下来将通过 LOD-Cloud、维基数据等数据集，获取更多实体之间的关系数据，来构建更加完善的西文古籍知识图谱。

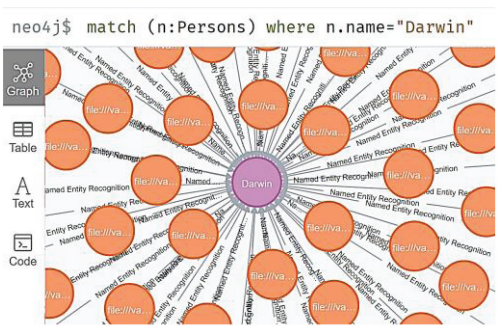

图 4 在西文古籍知识图谱中检索“达尔文”

① https://www.wikidata.org/wiki/Q692.

5 总结

在普通文献资源日益同质化的大背景下，稀有且价值较高的西文古籍，越来越受到图书馆的重视，而性价比高、功能完备的西文古籍数据库软件平台，则能够释放西文古籍的巨大价值。本研究在分析国内外西文古籍数据库建设现状的基础上，利用开源软件，结合开源社区的经验，构建了华东师大馆西文古籍数据库，使用人工智能技术对数据进行深度挖掘，并借助数字人文技术，赋予西文古籍数据库全文在线高清浏览、可视化、开放共享、分类浏览、检索建议等功能，未来还将进一步利用领域知识进行文献分析、构建知识图谱。系统已经具备简单的众包功能，未来还需要不断探索提高师生参与积极性的方法。本研究是对西文古籍文献全文数据库研发的一次尝试，实践过程均采用开源软件，并且将以虚拟机的方式打包共享整套软件系统，希望能够为有西文古籍数据库建设需求的图书馆提供借鉴。

参考文献

1 王雨卉.浅谈非善本西文古籍的开发整理[J].图书馆工作与研究，2011，182(4)：51－53.

2 董绍杰，卢刚，毕国菊.外文古籍的概念与界定初探[J].图书馆学研究，2010(14)：96－98.

3 张靖，张盈，林明，等.中国大陆及港澳地区图书馆西文古籍保护与修复情况调查[J].大学图书馆学报，2017，35(2)：99－108.

4 北京大学图书馆.北京大学西文古籍收藏[EB/OL].[2022－05－23]. https://www.lib.pku.edu.cn/portal/cn/zy/tszy/xiwentecang.

5 于燕妮.西文古籍题名和责任者著录问题探究——以中国人民大学图书馆 CADAL 项目为例[J/OL].图书馆杂志：1－12[2022－09－19].http://kns.cnki.net/kcms/detail/31.1108.g2.20220811.0852.002.html.

6 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].图书馆杂志：1－13[2022－08－20].http://kns.cnki.net/kcms/detail/31.1108.G2.20220517.1740.004.html.

7 Harvard Library. APIs & datasets [EB/OL].[2022－10－04]. https://library.harvard.edu/services-tools/harvard-library-apis-datasets#librarycloud.

8 Bodleian Libraries.Digital bodleian developer documentation[EB/OL].[2022－10－04]. https://digital.bodleian.ox.ac.uk/developer.

9 Stanford Libraries. Rare books [EB/OL]. [2022 - 10 - 16]. https://rarebooks.stanford.edu.
10 Cambridge University Library.Cambridge digital collection platform[EB/OL].[2022－05－02]. https://cudl.lib.cam.ac.uk/about-dl-platform.
11 Cambridge University Library. Code repository for Cambridge Digital Collection Platform [EB/OL].[2022－05－08].https://cambridge-collection.github.io.
12 Northwestern University Library.Digital collections source code [EB/OL].[2022－10－05].https://github.com/nulib.
13 University of Manchester Library.Manchester digital collections [EB/OL].[2022 - 08 - 11]. https://www.digitalcollections.manchester.ac.uk.
14 Bodleian Libraries.Frequently asked questions[EB/OL].[2022－10－07].https://digital.bodleian.ox.ac.uk/faq.
15 Harvard Library.Full text search[EB/OL].[2022－10－07].https://fts.lib.harvard.edu/fts.
16 University of Southampton.Digital library[EB/OL].[2022－10－08]. https://viewer.soton.ac.uk.
17 ETH Library.E-pics[EB/OL].[2022－10－17].https://ad-provenienz.e-pics.ethz.ch/main/mapsview.
18 University of British Columbia Library.BC historical books[EB/OL].[2022－10－15].https://open.library.ubc.ca/collections/bcbooks.
19 中国互联网络信息中心.第50次《中国互联网络发展状况统计报告》[EB/OL].[2022－11－14]. http://www3.cnnic.cn/n4/2022/0914/c88－10226.html.
20 Wikipedia. Digital humanities [EB/OL]. [2022 - 04 - 02]. https://en.wikipedia.org/wiki/Digital_humanities.
21 W3C中国.W3C发布web注释数据模型、词汇表及注释协议等三份正式推荐标准[EB/OL].[2022－02－23]. https://www.chinaw3c.org/archives/1728.

作者单位：华东师范大学图书馆，上海，200241
收稿日期：2022年6月11日
修回日期：2022年11月18日
（责任编辑：李晓东）

Research and Development of Western Rare Books Database from the Perspective of Digital Humanities

Zhang Yi

Abstract: In recent years, research on Western rare books has been increasing among Chinese libraries, but most of them have been focused to collection investigation, version analysis, and repair, with limited research on digital resources of Western rare books. In terms of practice, the service mainly focuses on paper version materials reading, which cannot meet the needs of readers to access Western rare books online. By analyzing the current situation of Western rare book databases construction at home and abroad and taking the construction of Western rare books full-text database at East China Normal University as an example, the paper introduces the practical experience on implementing open-source software and digital humanities research tools to develop a database for Western rare books, with hope to be used for reference by peers.

Keywords: Digital Humanities; Western Rare Books; Open Source Software; Omeka-S; Open Semantic Search