CASHL大型特藏文献增值性服务研究

课 题 单 位： 兰州大学图书馆

课题负责人：魏清华（兰州大学图书馆）

课题组成员：宋戈（兰州大学图书馆）、孙林（兰州大学图书馆）、张继忠（兰州大学图书馆）、薛小婕（兰州大学图书馆）、胡文静（兰州大学图书馆）、王海花（兰州大学图书馆）

结项时间：二零一九年六月

摘　要：在日趋成熟的数字化环境中，特藏文献资源数字化是图书馆信息化、数字化、网络化发展的必然趋势。通过数字化建设，可以为用户提供 CASHL 特藏文献的目次检索与浏览、全文搜索、在线预览、多语种在线翻译等增值性服务。通过数字化等手段促进资源的深度挖掘和服务的多样化发展，以满足用户多样化、个性化、专业化的文献信息需求。

关键词：CASHL　特藏文献　数字化 全文搜索 增值服务

目　录

1 课题背景

特藏资源是图书馆在自身发展过程中形成的具有地域特色、文化内涵或学科特点的馆藏。这些馆藏在内容和载体上有别于其他馆藏，具有独特性和稀缺性；在日趋发展的数字化环境下，特藏资源将会是图书馆的一个核心要素，一定程度上体现了图书馆馆藏的重要价值，历来为图书馆所重视。

图书馆的特藏资源可以从四个要素来界定：一是珍稀性或独特性；二是需要配备专门的空间和设施来对其流通、展示、利用加以限制；三是特藏建设要服务于图书馆及其所在机构的教学科研；四是数字化是建设特藏应该首要考虑的因素。[1]因此，开展特藏资源的数字化建设是图书馆发展的必然趋势。既可以为用户提供在线检索和预览，并方便快捷地开展馆际互借和文献传递服务，实现文献资源的共建共享，又可以减少物理损坏，达到特藏资源长期保存的目的。通过数字化建设，可以为用户提供 CASHL 特藏文献的目次检索与浏览、全文搜索、在线预览、多语种在线翻译等增值性服务。

在“CASHL 特藏++深度服务”项目的实施过程中，针对印刷型文献的特点，以兰州大学图书馆典藏的 CASHL 大型特藏文献——《阿拉伯边界报告系列丛书》为研究对象，构建了基于全文搜索系统的 CASHL 特藏文献数字化建设框架（图 1）。框架共分为三部分：一是元数据加工。通过文字识别软件，将图像文件转换成文本文件，为全文搜索系统开发提供数据基础；二是全文搜索系统开发。利用全文搜索引擎搭建全文搜索系统；三是通过在线程序 API 开发多语种在线翻译和在线预览功能。

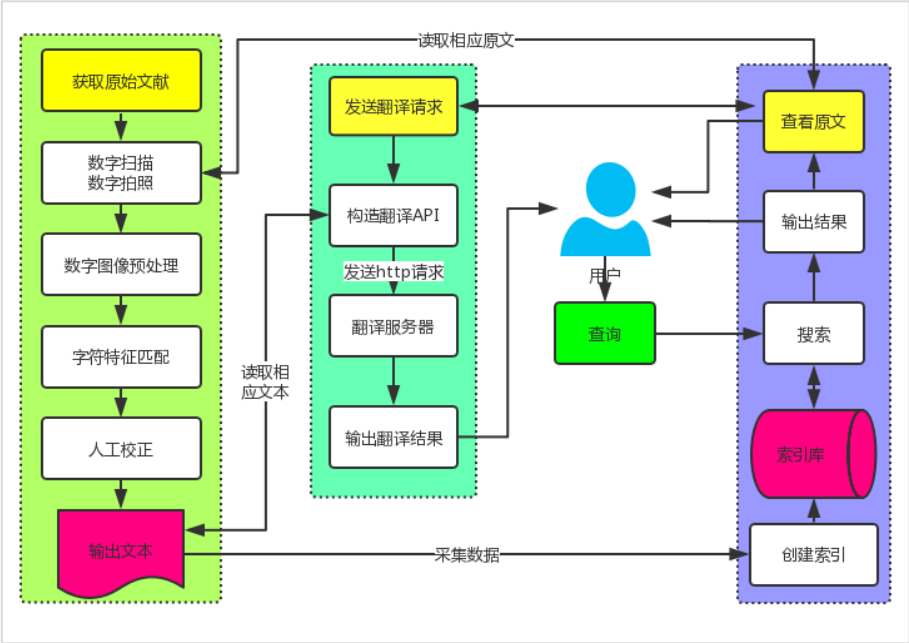

图 1 基于全文搜索系统的 CASHL 特藏文献数字化建设框架

2 研究目标和意义

2.1 研究目标

通过数字化建设，为用户提供 CASHL 特藏文献的目次检索与浏览、全文搜索、在线预览、多语种在线翻译等增值性服务。

2.2 研究意义

在数字人文背景下，本课题对 CASHL 特藏文献的内容进行了深层次的揭示，用户可以通过全文搜索系统通过关键词直接检索文献内容，比目次检索更贴合用户的检索需求；对文献的文本化加工也为今后的知识挖掘、知识组织等工作提供了数据基础；多语种在线翻译、在线预览等辅助性工具为用户使用 CASHL 特藏服务平台提供了便利，增加了用户粘性。

3 研究内容

3.1 CASHL 特藏文献及服务现状

3.1.1 CASHL 特藏文献概况

CASHL 特藏文献采取竞争性采购方式进行，每年由各成员馆在本校征集采购目录并上报给 CASHL 管理中心，由 CASHL 管理中心统一甄选和采购。从订购标准看[2]，CASHL 特藏文献至少具有以下特征：（1）稀缺性。订购标准保证了其在国内具有唯一（或极少）复本，通过其他途经基本无法获得；（2）学术性。CASHL 特藏文献都是由各高校的专家实名推荐购买，且以原始档案资料为主，对相关领域的学术研究将产生直接的促进作用；（3）系统性。CASHL 特藏文献学科相对集中，具有相对完整的专题，无法拆分；（4）共享性。这是由 CASHL 的文献保障和服务职能所决定的，各收藏馆需要提供基于特藏文献的文献传递、网络数据库等对外服务。

截至 2019 年 4 月，CASHL 大型特藏文献发订量为 205 种，分别收藏于多所高校图书馆，涉及历史学、哲学、法学、社会科学、语言学、区域学、文字艺术、政治军事、图书馆学、教育学、经济学等多个学科门类（图 2）[3]。

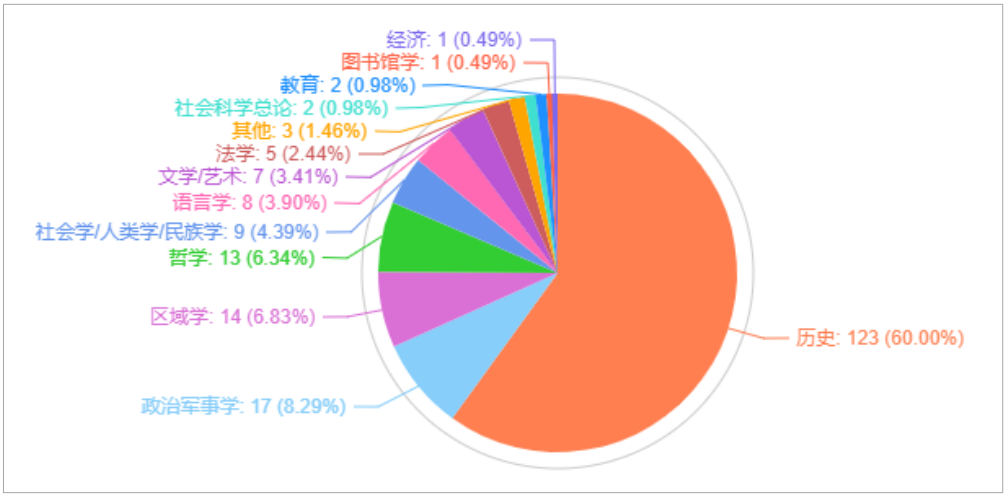

图 2 CASHL 大型特藏文献学科分布

3.1.2 CASHL 特藏文献服务现状

CASHL 资源平台提供了大型特藏文献的文献传递和馆际互借服务；用户可在此平台查询文献，点击“发送文献传递请求”按钮后，输入用户名和密码，即可进入申请信息页面，填写相应的信息后即可提交文献传递请求。元数据质量方面，目前 CASHL 特藏文献的元数据设有题名、出版社、资源类型、学科类型、馆藏地、章节目次等几个表示文献外在特征的字段。资源发现方面，在 CASHL 特藏文献专题页面提供基于文献标题的初级检索，以及首字母和学科的分类浏览（图 3）。可见，目前 CAHSL 特藏文献的组织揭示和服务尚处于粗粒度状态，无论是对用户还是成员馆都带来了不利影响：一是用户对特藏文献内容了解途径较少。用户只能在 CASHL 资源平台的页面中浏览文献的标题，对文献的内容和具体章节基本无从知晓，增加了申请的难度以及时间、经济、人员等各方面的成本。二是不利于文献的保藏。用户每提交一次文献传递申请，图书馆都需要对文献进行数字化扫描或拍照，反复的数字化有可能对文献造成一定的污损。

图 3 CASHL 大型特藏详细内容

综上所述，目前 CASHL 特藏文献的组织揭示和服务模式还有待深化和拓展。因此，CASHL 管理中心从 2015 年开始设立“CASHL 特藏++深度服务”（以下简称“CASHL 特藏++”）项目，鼓励有条件的图书馆开展大型特藏内容深度挖掘服务，力图通过数字化手段，促进特藏文献的保存和利用，提升 CASHL 的在线服务能力。

3.2 CASHL 特藏文献数字化建设

3.2.1 “CASHL 特藏++”项目概况

截至 2018 年，共有 13 项“CASHL 特藏++”项目被批准立项（表 1），其中 2015 年 2 项，2016 年 4 项，2017 年 4 项，2018 年 3 项。从表中可以看出，在“CASHL 特藏++”项目的推动下，CASHL 在大型特藏文献数字化揭示与服务方面取得了一定成果，初步建成了一批能够较为完整地揭示特藏文献的数据库平台，对文献的基本信息或内容进行了揭示、报道，并且在尊重知识产权的前提下提供了在线服务，用户能够通过网络进行访问。

表 1 CASHL“特藏++”项目一览表（2015—2018 年）

序号	项目名称	资源类型	承建学校	立项时间	完成情况	网址
1	《日本外交文书》深度服务	图书	武汉大学	2015 年	完成	http://apps.lib.whu.edu.cn/wjws/list.asp

2	卫理公会传教士信件缩微胶卷专题资源数字化加工与特色库建设	缩微胶卷	中山大学	2015 年	完成	（仅限中山大学校内访问）
3	《教育史——来自于哥伦比亚大学师范学院图书馆米尔班克纪念图书馆的珍藏》Unit1 内容揭示与目次检索	缩微平片	北京师范大学	2016 年	完成	（仅限北京师范大学校内访问）
4	《希腊罗马作家作品集》平台建设与深度服务	图书	东北师范大学	2016 年	完成	http://bt.library.nenu.edu.cn/main.html
5	《美国外交部档案-越南》深度服务	缩微胶卷	厦门大学	2016 年	完成	http://210.34.4.46/mfilm/
6	《美国公民自由联盟档案》缩微胶卷数字化加工及特藏数据库的建设	缩微胶卷	吉林大学	2016 年	完成	http://202.198.25.162/tsk/wdindex.action （点击“美国公民自由联盟档案”专题）
7	《杜鲁门口述历史全集》深度服务	缩微平片	华东师范大学	2017 年	完成	http://www.lib.ecnu.edu.cn/truman/index.php
8	阿拉伯边界报告系列丛书深度揭示与服务	图书	兰州大学	2017 年	完成	http://202.201.7.42/
9	《影印标点韩国文集丛刊（续）》目次检索	图书	南京大学	2017 年	完成	http://114.212.7.49/korea/
10	《美国和卡斯特罗的古巴》解密档案的揭示	缩微胶卷	南开大学	2017 年	在建	（仅限南开大学校内访问）

	与服务					
11	《移民及难民政策特别委员会文件》缩微胶卷数字化加工及深度服务平台建设	缩微胶卷	东北师范大学	2018 年	在建	http://bt.library.nenu.edu.cn/reelindex/
12	《美国宗教合集》深度揭示与服务	缩微胶卷	四川大学	2018 年	在建	
13	民国期刊目次整理与揭示项目（一期：创刊号）	期刊	复旦大学	2018 年	在建	

3.2.2 CASHL 特藏文献数字化加工

对 CASHL 特藏文献进行数字化加工主要目的是将纸型、缩微型、数字型等介质的特藏文献转换成能够被机算机读取和利用的数据格式，并通过元数据著录，将文献的外表特征和物质特征进行描述和标引，最终形成系统的、可准确揭示文献特征和内容的元数据数字格式文件（图 4）。

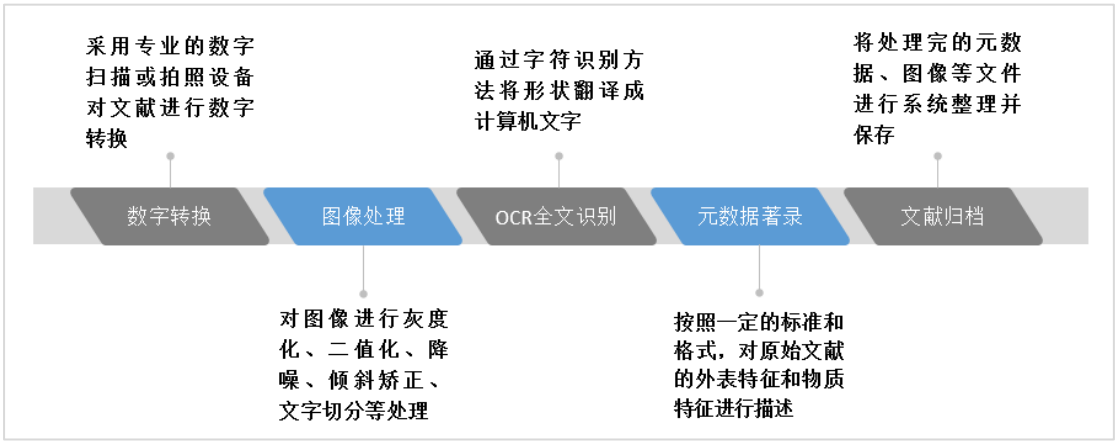

图 4 CASHL 特藏文献内容数字化处理流程

（1）数字转换

通过数字转换设备对文献进行数字化扫描或拍照，并在计算机中进行存储，实现文献内容载体向数字的转变。不同存储介质的文献采用的数字化转换设备也各不相同。对于印本资源，基本采用高精度扫描仪或照相机；而对于缩微胶卷、缩微平片，则需要采用专门的缩微数字化扫描设备。

（2）图像处理

图像处理主要指数字图像处理。数字图像处理(Digital Image processing)是通过计算机对图像进行去除噪声、增强、复原、分割、提取特征等处理的方法和技术。一般来讲，对图像进行处理的主要目的有三个方面：（1）提高图像的视感质量，如进行图像的亮度、彩色变换，增强、抑制某些成分，对图像进行几何变换等，以改善图像的质量。（2）提取图像中包含的某些特征或特殊信息，这些被提取的特征或信息往往为计算机分析图像提供便利。（3）图像数据的变换、编码和压缩，以便于图像的存储和传输。数字图像处理常用方法有图像变换、图像编码压缩、图像增强和复原、图像分割、图像描述和图像识别几个方面[4]。

在 CASHL 特藏文献的数字化加工过程中，经常出现文本图像亮度过亮或者过暗、页面倾斜、黑边、清晰度低等情况。这主要由两个方面因素造成：一是加工设备设备方面的问题。批量化加工设备一般都采用统一的分辨率、亮度、格式等设置，这只能满足大部分图像要求，对于个别图像需要单独处理；二是原始文献本身的问题。原始文献可能存在倾斜、字迹模糊等问题。这些问题都需要通过图像处理技术进行处理，以最大化改善图像质量。图像处理完成后，需要根据项目需求按照一定的图像格式保存在存储设备中。目前比较流行的图像格式有光栅图像格式 BMP、GIF、JPEG、PNG、TIFF 等，以及矢量图像格式 WMF、SVG 等。

（3）OCR 文字识别

文字识别，又称为光学字符识别（Optical Character Recognition，OCR），旨在将扫描或拍摄得到的手写或打印文本识别成可编辑的计算机文字。文字识别是对图像文件进行文本化的方式之一，其目的是通过自动化批量识别软件将图像型文本文件转换成计算机可识别的字符，以便于进行更深层次的内容挖掘与知识发现。

目前，“CASHL 特藏++”项目大多采用 ABBYY FineReader 软件来进行 OCR 处理（图 5）。该软件提供非常完善的预处理功能和强大的 OCR 识别能力，支持导出 TXT、WORD、PDF 等各类格式，原件和识别后的文档对照浏览，可实时人工校正，满足文字识别各类使用场景[5]。

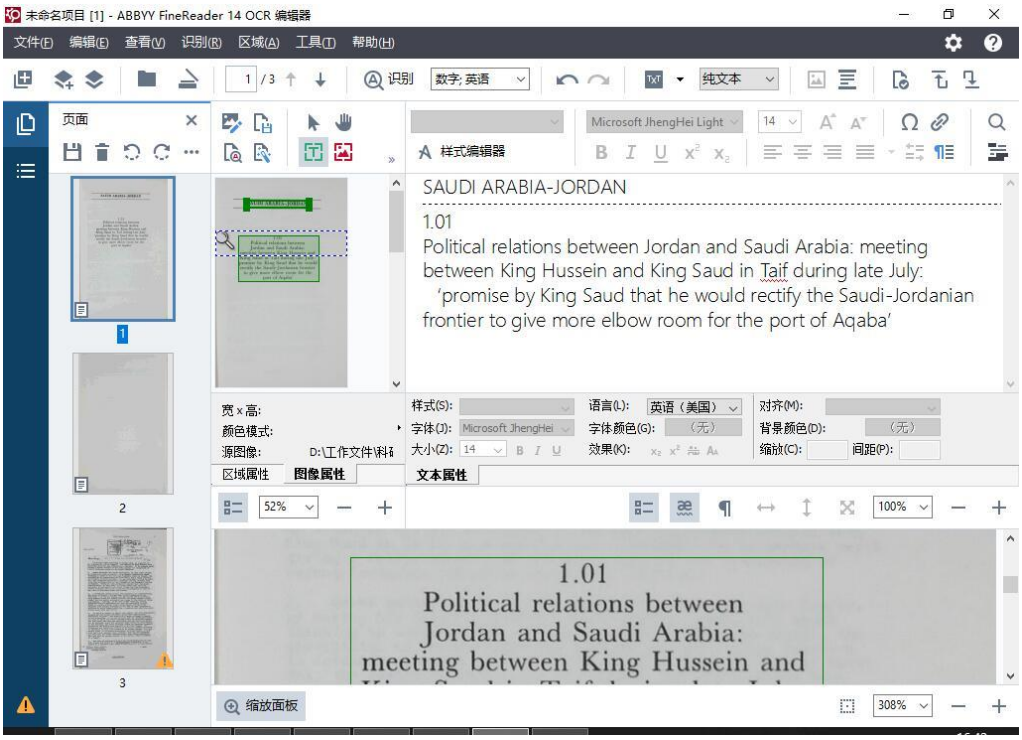

图 5 ABBYY FineReader 14 OCR 界面

（4）元数据著录

元数据是关于数据的描述性数据信息，说明数据内容、质量、状况和其他有关特征的背景信息，其目的是促进数据集的高效利用。在“特藏++”项目实施过程中，不同的文献所采用的著录标准和格式有所不同，需要根据项目的实际需求和文献内外部特征来决定。例如武汉大学图书馆通过对《日本外交文书》的内容体例进行剖析，确定目次结构和检索点，其元数据字段主要有章节号、章节标题、页码、事项、时期等[6]；南京大学图书馆承担的《影印标点韩国文集丛刊》目次检索项目[7]，以韩国文集为主要内容，通过作者姓名、拼音首字母、文章标题等字段将文集内容和作者进行了详细标引和著录。

总体而言，除了考虑统一和跨库检索等一般性需求外，还应尽可能地加强标引深度，或针对性地设置特殊的检索项目，例如可以增加学科、年代、地点、文摘、全文等检索字段来实现多种途径检索。

（5）文献归档

指将处理好的元数据、图像等文件进行系统整理、保存、备份。元数据可通过 EXCEL、MySQL 数据库等软件进行处理和保存，图像可以按照章节、日期、时间等分类保存。

3.3 全文搜索系统开发

全文搜索系统指应用全文搜索技术建立起来的系统体系。全文搜索系统依赖全文搜索引擎，往往需要根据自身特点和要求进行定制开发，因此产品化的商业软件较少，多为利用开

源的基础平台进行二次开发[8]。Solr 是目前比较流行的、开放源码的、基于 Lucene、Java 的全文搜索服务器，它提供了灵活的、简单的接口,让开发人员能够方便、快速地构建搜索引擎[9]。Solr 的主要特性包括：高效、灵活的缓存功能，垂直搜索功能，高亮显示搜索结果，通过索引复制来提高可用性，提供一套强大的 Data Schema 来定义字段、类型和设置文本分析，提供基本 WEB 的管理界面等。因此，研究人员使用 Solr 作为构建全文搜索系统的核心部件。

Solr 的工作原理（图 6）：（1）Solr 对外提供标准的 HTTP 接口来实现对索引的增加、删除、修改和查询；（2）在 Solr 中，用户向部署在 servlet 容器中的 Solr Web 应用程序发送 HTTP 请求来启动索引和搜索;(3)Solr 接受请求后,确定要使用的 SolrRequestHandler，然后处理请求，通过 HTTP 以同样的方式返回响应；（4）默认配置返回 Solr 的标准 XML 响应，也可以配置 Solr 的备用响应格式（JSON）。

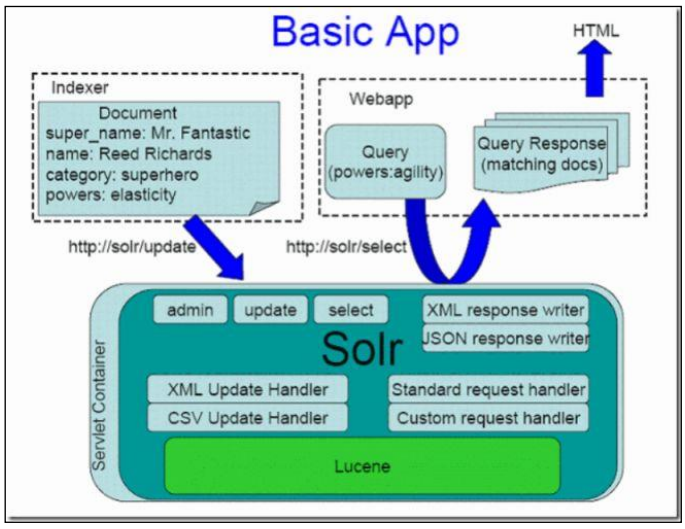

图 6 Solr 的工作原理

全文搜索系统按照搜索流程主要分为五部分（图 7）：

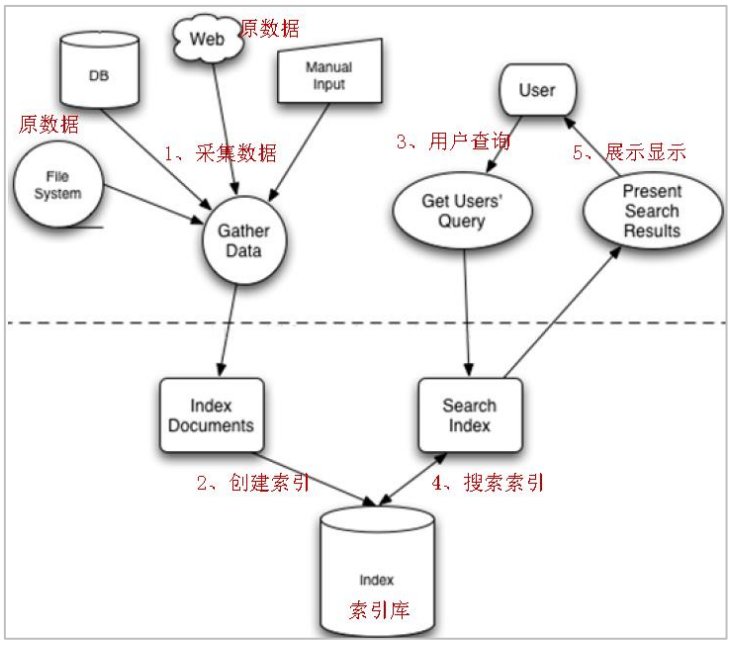

图 7 全文搜索流程

（1）采集数据。Solr 支持从数据库、网页、文档、手工录入等方式对数据进行采集。

（2）创建索引。在配置文件 schema.xml 中定义需要索引的字段，如标题、全文、路径、日期等。Solr 经过分词处理后采用倒排法创建索引并存储到索引库中。

（3）发起查询请求。用户在搜索页面输入检索词并发送 POST 请求。Solr 支持完全匹配查询、模糊匹配查询、多域查询、通配符查询、组合查询等各种查询方法。

（4）搜索索引。Solr 接收到用户发送的请求后，到索引库中进行搜索并排序。

（5）返回并显示结果。用户将 Solr 返回 XML 或 JSON 等格式的检索结果进行解析，并在浏览器界面上按照规定样式显示出来。通常需要进行关键词高亮显示、分页、链接原文等操作（图 8）。

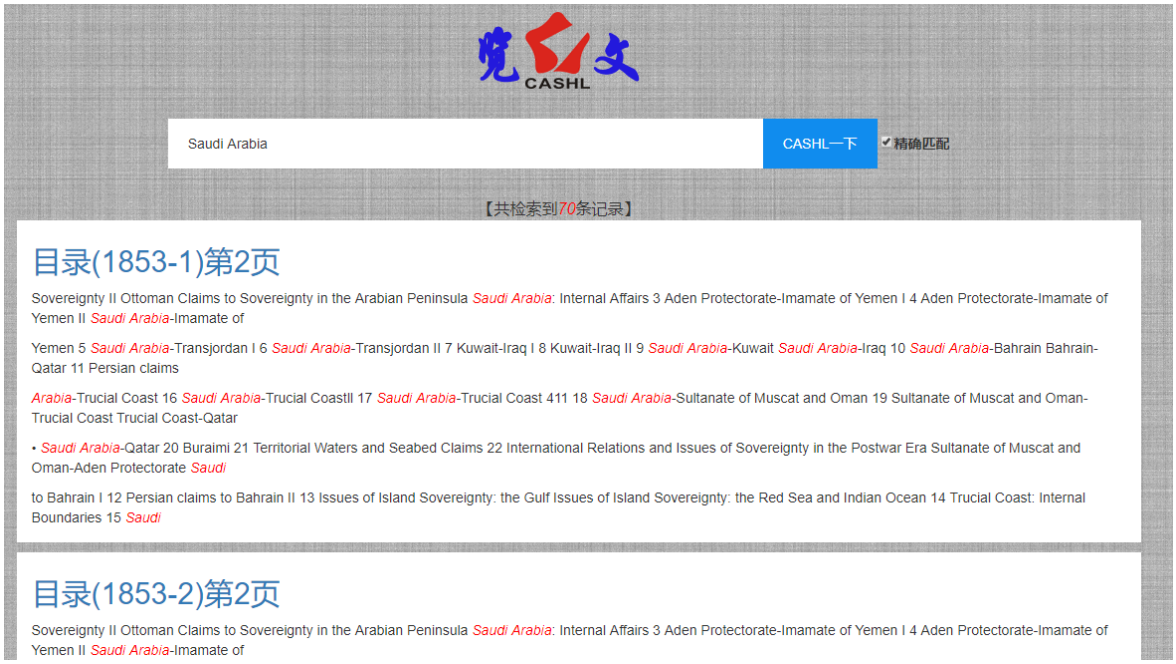

图 8 CASHL 全文搜索系统检索界面

3.4 扩展功能开发

3.4.1 多语种在线翻译

机器翻译，又称为自动翻译，是利用计算机将一种自然语言（源语言）转换为另一种自然语言（目标语言）的过程[10]。国内外对于机器翻译都有比较成熟的产品，如百度翻译、有道翻译、Google 翻译、微软翻译等等。目前主要有基于规则、基于实例、基于统计以及基于神经网络的实现手段。机器翻译作为人工翻译的辅助，是一种很有效的阅读辅助手段。

CASHL 特藏文献全部为外文文献，并且语种丰富，对读者的外语水平要求较高。利用机器翻译来辅助阅读，虽然不可能达到人工翻译的水平，但是由于目前的机器翻译系统大多采用大数据、神经网络等先进技术，使翻译结果的质量有了大幅提升，这使得机器翻译具备了辅助阅读的能力。

有道智云平台是网易有道旗下一个为开发者、企业和政府机构等提供自然语言翻译等服务以及行业解决方案的云服务平台。其自然语言翻译服务采用业界领先的神经网络机器翻译（Neural Machine Translation，NMT）技术，支持中文、英语、日语、韩语、法语、西班牙语等语种的互译功能[11]。

有道翻译 API 接口提供有道的翻译服务，包含了中英翻译和小语种翻译功能。开发者只需要通过调用有道翻译 API，传入待翻译的内容，并指定要翻译的源语言（支持源语言语种自动检测）和目标语言种类，就可以得到相应的翻译结果[12]。

用户需要在平台上注册成为开发者，并获取应用 ID（appKey）和应用密钥等信息。然后，根据平台提供的 API 与自身系统进行开发集成。用户发起翻译请求后，系统判断当前用

户打开的文献名称及页码，然后到服务器中读取对应的文本文件。读取成功后，将文本、源语种、目标语种等信息传入翻译 API，再按照翻译服务器规定的格式发送 HTTP 请求。翻译服务器接收到请求后开始翻译，结束后再用 HTTP 的方式返回 JSON 格式响应。用户解析 JSON 并显示在浏览器上（图 9）。

```
//向服务器提交翻译参数并显示结果
function translate(query){
    var appKey = '4b1db75cb8      ';
    var key = 'CQ1KvJwzK1h70zRufkkjHbY          ';//注意：暴露appSecret，有被盗用造成损失的风险
    var salt = (new Date).getTime();
    var query = query;// 多个query可以用\n连接  如 query='apple\norange\nbanana\npear'
    var from = $("#fLanguage option:selected").val();; //源语种
    var to = $("#tLanguage option:selected").val();; //目标语种
    var str1 = appKey + query + salt +key;
    var sign = md5(str1);
    //使用ajax向有道API提交数据
    $.ajax({
        url: 'http://openapi.youdao.com/api',
        type: 'post',
        dataType: 'jsonp',
        data: {
            q: query,
            appKey: appKey,
            salt: salt,
            from: from,
            to: to,
            sign: sign
        },
        success: function (data) {
            if(data.errorCode=="0"){
            //在页面上显示翻译结果
            document.getElementById("result").innerHTML=data.translation;
            }else{
                alert("翻译失败！ ");
                document.getElementById("result").innerHTML='';
                $("#result").hide();
            }

        }
    });
}
```

图 9 有道智云翻译 API 结构

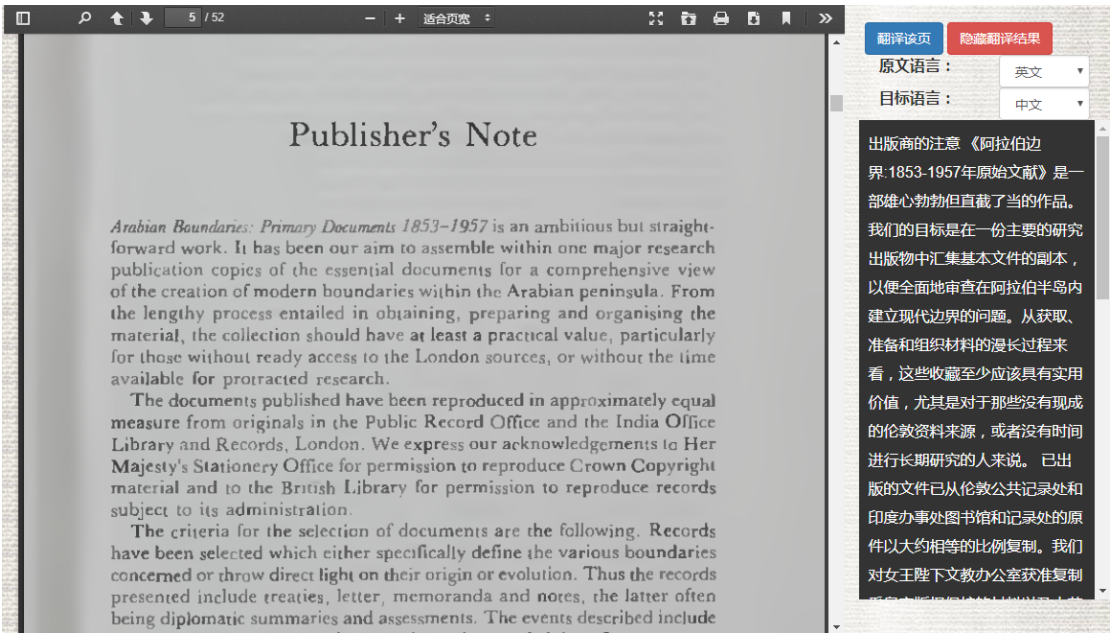

图 10 在线预览及翻译界面

3.4.2 在线预览

CASHL 特藏文献服务平台提供了目次、部分正文等的在线预览和试读，可以帮助用户快速了解文献的目录结构和具体内容，节约了时间成本。如果以图片格式进行预览，可由浏览器直接显示，无需单独配置；如果以 PDF、EPUB 等格式进行预览，则一般需要配置一些插件才能在浏览器中打开和显示。通过支持多种设备的、开源的文档读取解析插件如 PDF.JS[13] 或者 EPUB.JS[14]，可以快速实现 PDF、EPUB 文件的在线预览，避免了下载、打开阅读器等繁琐的过程（图 10）。

4 结语

特藏资源是图书馆在自身发展过程中形成的具有地域特色、文化内涵或学科特点的馆藏。这些馆藏在内容和载体上有别于其他馆藏，具有独特性和稀缺性；在日趋发展的数字化环境下，特藏资源将会是图书馆的一个核心要素。通过数字化等手段促进资源的深度挖掘和服务的多样化发展，满足用户多样化、个性化、专业化的文献信息需求。

参考文献

[1] 黄雯越，王铮. 数字环境下研究型图书馆的特藏建设:内涵、趋势与实践案例[J]. 图书情报工作. 2016, 60(17): 40-46.

[2] 图书馆关于 2019 年度 CASHL 人文社科外文大型特藏文献荐购目录征集启示[EB/OL].[2019-05-15]. https://mp.weixin.qq.com/s?__biz=MzA5NTAxMjk2Ng==&mid=2709118031&idx=1&sn=529ecf81c7a391020e509b231267d887&chksm=b4c8ad4683bf24503039b494473d29bfe1273aadd81ac81b82d85ef69776d9fe96274c5f4eea&mpshare=1&scene=1&srcid=#rd.

[3] 大型特藏文献[EB/OL].[2019-04-06]. http://www.cashl.edu.cn/cashlsearch/result.html?res=spcollection&sphome=1.

[4] 陈洪著. 数字媒体技术概论[M]. 北京：北京邮电大学出版社, 2015:4.

[5] ABBYY FineReader 14[EB/OL].[2019-05-01].http://www.abbyy.cn/finereader/.

[6] 《日本外交文书》目次检索[EB/OL].[2019-05-11]. http://apps.lib.whu.edu.cn/wjws/.

[7] 韩国文集丛刊[EB/OL].[2019-05-11]. http://114.212.7.49/korea/.

[8] 杨新涯著. 图书馆文献搜索研究[M]. 重庆：重庆大学出版社, 2015:195.

[9] Apache Solr[EB/OL].[2019-03-06].http://lucene.apache.org/solr/.

[10] 机器翻译_百度百科[EB/OL].[2019-03-06].https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E7%BF%BB%E8%AF%91/411793.

[11] 有道智云[EB/OL].[2019-02-11].http://ai.youdao.com/.

[12] 有道智云翻译 API 简介[EB/OL].[2019-01-23].http://ai.youdao.com/docs/doc-trans-api.s.

[13] pdf.js[EB/OL].[2019-05-11].http://mozilla.github.io/pdf.js/.

[14] GitHub - futurepress/epub.js: Enhanced eBooks in the browser[EB/OL].[2019-04-18].https://github.com/futurepress/epub.js.