# 图书馆异构特藏资源整合的数字人文研究需求*

李欣，张毅，汪志莉
（华东师范大学图书馆，上海 200062）

摘要：本文以图书馆特藏资源建设及服务中存在问题为切入点，从特藏资源用于支撑数字人文研究需求角度，结合华东师范大学图书馆在Web Services技术应用、GIS技术应用以及结构化数据分词等方面实践，介绍特藏资源整合、地理位置与标签云可视化检索实现方法。

关键词：异构；特藏资源；数字人文
中图分类号：G250.7
DOI：10.3772/j.issn.1673-2286.2017.11.008

## 1 问题提出

随着图书馆在数字网络时代的快速发展，图书馆学术馆藏的相似度越来越高，具有独特学术性、历史性的馆藏资源成为图书馆持续发展的要素。特藏资源不仅能很好地服务有“专门”需求的用户，其独有的学术价值也是图书馆在信息资源共享中体现优势和竞争力所在。

图书馆特藏资源建设得益于1999年中国高等教育文献保障系统（China Academic Library & Information System，CALIS）启动的专题特藏数据库建设项目，其目的在于全面挖掘、整理和发布国内各高校成员馆的一些未开发、散在各处、难以被利用的独有或稀缺资源、网络原生数字资源等，逐步形成具有学科特色、地方特色或民族特色的专题特藏文献数据库服务群。项目后期CALIS整合75所大学的97个特藏数据库元数据，用以构建统一元数据发布检索平台，检索指向原文所在高校图书馆的特藏数据库门户。因CALIS专题特藏数据库建设的项目制特点，后续没有持续性发展经费投入，各高校特藏数据库建设经费需要自行解决。尽管如此，近20年来各高校图书馆的特藏资源数字化建设仍得到快速发展。笔者2014年开展了全国师范大学图书馆馆藏数字化特藏资源调研[1]，30家图书馆（约占CALIS资助图书馆数量的30%）参与问卷调查；2017年3月笔者基于网络追踪了这30家图书馆网站特藏资源变化情况，特藏资源数据库已经多达164个。

调研结果表明，各图书馆都在力求通过特藏资源建设选题体现馆藏独特性，但在数字化建设方式上还停留在全文扫描和简单元数据加工的数据库建设阶段，提供的服务多依赖于运行多年的不同商业或自建平台，功能相对较单一，以资源数字化保存和提供简单的检索功能为主，资源利用率低；加之内容封闭、存储分散且只针对本机构用户开放，极大地限制特藏资源的价值发挥。相较支持人文学者用于科学研究的环境需求而言，无论在资源数据化组织还是平台新功能拓展方面，都远落后于当下日新月异的新技术发展。如资源如何从数字化向数据化存储过渡、数据间关联关系的建立，以及平台的基本数据分析、可视化、文本挖掘等功能缺失等；同时，各图书馆的特藏资源数据库通常只针对某一特定类别的馆藏资源进行数字化处理，其数量局限于一个图书馆的馆藏，多个图书馆的同一类资源无法在一个平台完整表现，很难构成支持研究的资源权威性。以全国师范大学图书馆教育类特藏资源为例，因教育学科在师范大学学科地位的重要性，多数高校都将其作为特藏资源进行数字化保存，但各图书馆间缺乏同类资源的共享与互补，不能形成相对完整的教育类特藏资源数据库。

* 本研究得到国家社会科学基金项目“图书馆异构特藏资源的数字人文研发与共享模式研究”（编号：17BTQ004）资助。

## 2 特藏资源整合与数字人文的关系

数字人文也称人文计算，1949年Busa使用电脑对神学家Aquinas著作内的字词进行大规模处理，被认为是数字人文的起源[2]。目前学界对其定义尚无权威界定，它是伴随人文学者研究方式的变化而产生的。从资源服务研究角度看，数字人文即结合大量数字资源，运用信息技术来从事人文研究[3]。

数字人文的主要范畴是通过信息技术改变知识获取、标注、比较、引用、取样、阐释与表现的方式[4]，使人文学者从大量重复性工作中解放出来，实现人文研究的创新发展。数字人文的意义在于对大规模文本的深度挖掘和智能分析，因此相关资源的大规模整合以及资源的细粒度、关联性重建，成为图书馆支撑人文研究的资源建设重点。

### 2.1 特藏资源与数字人文

数字人文既包含资源又涉及信息技术，因此其研究领域涉及文理交叉学科。计算机科学、信息科学与图书馆学是数字人文研究的基础学科，语言学、文学、哲学、历史、艺术、社会学等是数字人文研究的应用学科，两者不断交叉与融合，逐渐衍生出数字人文研究的新方向如数字艺术、数字史学等[5]。

本文调查的164个特藏数据库所涉及的类别归纳如表1所示，其中古籍善本、民国图书、方志、地区类专题、多媒体、教育类等资源与数字人文研究的应用学科密切相关。因CALIS专题特藏数据库建设的项目制特点，使资源后续的内容建设和维护投入不足。而这些资源正是构成数字人文基础设施的重要组成部分，图书馆在数字人文基础设施建设过程中应该首先从这部分资源入手。

表 1 师范大学图书馆馆藏数字化特藏资源数据库数量调查

个

| 古籍善本 | 民国图书 | 方 志 | 地区类专题 | 多媒体 | 教育类 | 其他专题 | 教参/学位论文/文库 |
|---|---|---|---|---|---|---|---|
| 11 | 5 | 7（含17个子库） | 9 | 15 | 28 | 31 | 22/22/14 |

### 2.2 特藏资源整合与数字人文

目前各图书馆特藏资源数据库平台都存储了大量可计算的基础数据对象，如数字、文本、格式化数据、图像、声音等。作为数字人文的重要数据来源，整合不同图书馆的同类特藏资源，可以更充分有效地汇聚分散、孤立、封闭的特藏资源，创建可促进人文研究的数据集或大规模结构化数据，从而扩大人文学者的抽样范围，提升资源所支撑的研究权威性。

2008年正式对公众开放的欧洲多媒体在线图书馆（Europeana）项目作为典型的资源整合案例，集合了欧洲各大数字资源门户网站和搜索引擎，其元数据采用资源描述框架（Resource Description Framework，RDF）存储，目的是方便在语义环境中通过关联数据实现资源的有效揭示，提高资源可用性。

成立于2008年的Hathitrust[6]项目，通过与高校及公共图书馆合作，整合合作馆的数字馆藏，并向所有成员馆用户开放资源获取服务。2011年，印第安那大学与伊利诺伊大学建立HathiTrust研究中心[7]，为学术研究提供文本分析和可视化的工具。

2011年，美国数字公共图书馆（Digital Public Library of America，DPLA）项目与Europeana开展技术合作，建立可互操作的数字模型、资源规范，可开放获取的馆藏资源并共享源代码。2013年正式对公众开放的DPLA项目，将美国档案馆、图书馆、博物馆、文化遗产机构、私人收藏机构等分散的资源进行统一集合，旨在探索如何建设一个开放的、分布式的在线资源网络。这些典型案例的特点主要表现在力求资源的完整性、开放性、关联性（底层关联数据设计要求），并向具有数据分析功能、表现形式多样化的系统过渡。

在数字人文研究方面，以人文学者需求为主导建设的资源研究平台在近年来备受关注。如由哈佛大学费正清中国研究中心、北京大学人文社会科学研究院、台湾“中研院”历史语言研究所合作开发的中国历代人物传记资料库，复旦大学历史地理研究中心的中国历史地理信息系统（其数据资源来自复旦大学、亚洲空间信息网络澳大利亚中心、格里菲斯大学、哈佛燕京学社），以及台湾大学数位人文研究中心的台湾历史数位图书馆系统。

这些人文研究资源平台的建设具备以下特点：（1）以内容合作的特征突出资源建设的完整性，提升资源快速一站式获取能力；（2）丰富检索型数据库功能，通过分析工具的应用达到对数据深度挖掘的目的；（3）由目录数据库到扫描图像与光学字符识别文本过渡，使文献资源便于全文检索、文本挖掘、词频统计等，有助

于研究者发现除目录外的更多内容；（4）多角度的精细化元数据加工，以揭示文献内容和形式的多种属性。这些研究平台的建设，为图书馆开展基于异构特藏资源的整合与重建提供多角度参考。

## 3 特藏资源整合技术与方法

图1为具体化的数据集成系统框架，由三部分组成。网络层负责对分散在网络中的不同服务站点的异构数据源进行收割，数据源可以是关系型数据库、Excel表格，也可以是半结构化的XML文档等多种格式。数据层负责对各种异构数据提供统一的表示、存储和管理，以实现逻辑或物理上的有机集中。集成后的异构数据对用户而言，是统一和无差异的，用户能够透明、有效地对数据进行操作，从而实现全面的数据共享需求。应用层/表现层负责响应用户的具体请求。

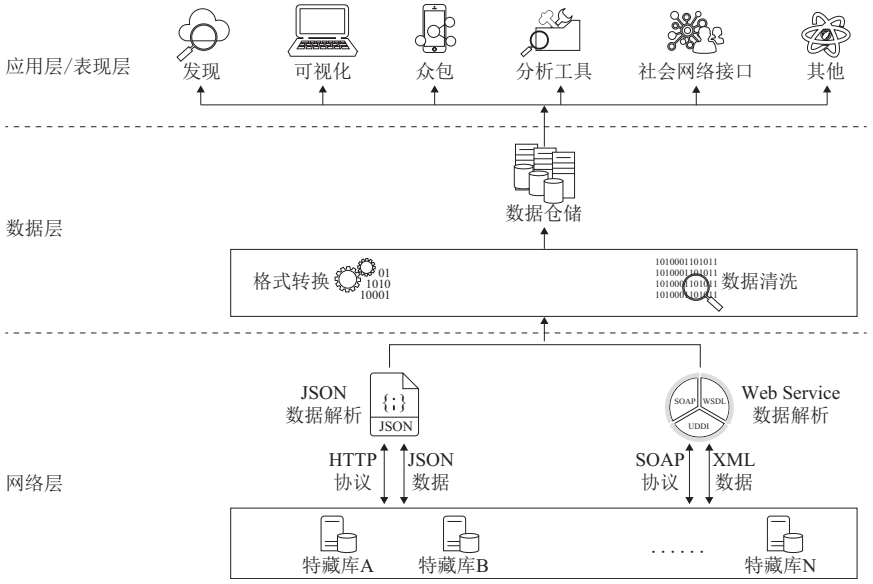


图 1 基于异构特藏资源整合的数字人文研究环境架构

### 3.1 地理信息系统方法

地理信息系统（Geography Information System，GIS）常应用于历史地理资源的整合，即以GIS为整合平台，整合多个专题数据库资源。如中国历史地理信息系统、中华文明时空基础架构、台湾历史文化地图等整合越来越多含有空间信息的专题数据库。申斌等设计的莆田历史人文地理信息系统，以GIS为平台整合文献（民间文献、地方档案、书籍）与田野调查资料（实物、建筑、仪式、音声），构成一个跨越史料文类、主题、数据类型的数字人文系统，也可以说是一个时空史料综合体[8]。这种整合主要以空间方式关联相关专题数据，属于图1中应用层整合。

GIS作为一种对地球表面空间地理数据进行采集管理和分析显示的软件系统，常被运用到传统历史地理研究[9]。此外，GIS系统在资源可视化检索、呈现，以及基于空间特征资源的聚类等方面有广泛应用。

### 3.2 关联数据应用

Berners-Lee等提出关联数据概念，指出关联数据在语义网中使用统一标识符（Uniform Resource Identifier，URI）和RDF发布结构化数据，是构建数据间链接的最佳实践方式[10]。以HTTP URI作为各种对象的统一标识符实现全网域范围唯一标识和定位，便于与Web上的其他数据集建立关联。关联数据在实现数据集成与共享中具有的优势主要体现在两点。第一，URI标识与复用。利用URI标识不同的实体对象，关联数据描述的粒度更加细化、语义化，并可跨领域得到更广泛的数据参引。第二，RDF描述与链接机制。采用RDF三元组“主语-谓语-宾语”的形式，关联数据描述科学数据及其间关系，通过RDF链接可以形成提供数据集成与共享的关联数据网络[11]。

应用关联数据技术实现资源整合，主要基于异构数据源共有核心元素，通过选取适合具体应用场景的

词表建立共享核心元数据本体，实现异构元数据间的语义整合和互操作；通过定制化扩展共享核心元数据本体，建立专门的元数据本体，实现特定资源语义化描述。这种基于数据层的整合不仅可用于实现数据细粒度的语义化、关联化的集成与共享，也可以为关联数据网络提供可用数据。

在关联数据应用方面，上海图书馆在家谱数据库以及历史地理数据的开放应用与服务方面，基于知识本体和关联数据技术进行大量实践研究，特别是在应用领域本体解决语义异构集成方面，对数字人文研究环境建设具有较高参考价值。

### 3.3 应用程序编程接口技术

异构数据一般指类型相同但在处理方法上存在差异的数据。在内容上则指不同数据库系统间的数据存在异构现象（如SQL Server和Oracle数据库中的数据），或指不同结构数据间存在异构现象（如SQL Server数据库数据和XML数据）。分布式异构关系数据库的整合始于20世纪80年代初期，目的是屏蔽各数据库结构、组织方式等方面的差别，为用户提供访问资源的统一接口[12]。由于数据存在多种异构性，因而需对数据进行整合处理[13]。

应用程序编程接口技术（Application Programming Interface）主要用于数据在网络环境下的互操作，属于纯粹信息技术方法的网络层整合技术，通常用于结构性异构数据的整合。以Web基础架构HTTP/SOAP协议为依托提供数据应用接口，目前已经成为数据互操作的成熟技术，用于实现基于网络的数据集成与共享。基于HTTP协议的JSON格式接口，是一种简单的数据互操作接口，数据传输效率高，不仅易于用户阅读和编写，也易于机器解析和生成。SOAP是一种基于XML数据格式可以传输复杂数据类型的协议，SOAP协议具有可扩展和独立于编程模型的消息处理框架，可通过多种底层网络协议使用。

## 4 研究实践

本文结合华东师范大学图书馆地方志数据库建设项目组（以下简称“项目组”）应用Web Service实现数据库间的互操作与结构性异构数据的集成，并利用GIS进行可视化展示，采用分词技术对现有数据库进行重建与功能拓展。项目组仅在网络层的数据整合方法、数据层的分词处理，以及应用层/表现层的数据可视化呈现与检索方法进行小范围技术实现，为深层次数据挖掘作准备。

### 4.1 数据整合方法研究

以高校图书馆目前自建数据库平台的资源为整合对象，本文通过调研国内师范大学图书馆的特藏资源数据库平台环境，发现目前国内主流平台包括TPI、TRS、Apabi-DESi、Apabi-TASi、IDL-ETD、DIPS、超星、麦达，以及一些基于开源商用数据库系统开发的平台，所涉及的数据库以SQL Server和MySQL为主，只有TPI平台使用封闭的自建数据库[1]。而对于标准商用数据库系统，无论是开源的数据库（如SQL Server、MySQL、Access、PostgreSQL等），还是收费的数据库系统（Oracle、Sysbase等），通过SOAP和HTTP等网络通信协议，以标准数据库连接方式，基于数据库开放权限进行数据库底层操作完全透明。面对封闭的自建数据库系统（如TPI平台），用户无法直接对数据库进行操作，需要通过开发商提供的封装函数才能实现对数据库的有限访问。

项目组选取本地TPI平台的学位论文数据库进行数据获取研究实践，平台提供Web Service接口方式，因其封闭的数据库模式，必须通过封装函数连接后台数据库，实现基于网络的数据互操作。典型应用如在高校主页服务器与TPI服务器进行数据互操作，实现对高校主页下“教师名录”关联相关教师指导学位论文元数据的推送。高校主页可以看成图1中的“数据仓储”，TPI平台的数字资源即网络层中的特藏资源之一。麦达的学位论文数据库和自建特藏数据库发布平台采用分开存储方式，后台数据存储采用SQL Server数据库，但某些字段中存储XML格式数据，用户直接应用SQL语句操作有一定障碍，数据互操作由公司提供开发的JSON接口格式。Web Service和JSON两种接口方式对应的数据格式分别为XML和JSON，因此在数据获取后的格式解析方面，相应处理脚本稍有不同。

### 4.2 地方志数据库重建

从数字人文角度看，资源整合只是研究环境建设的初级阶段，如何提升人文学者资源发现能力，使其能通

过新型资源研究环境实现对资源的比对、统计和分析功能。项目组在原有特藏地方志数字资源数据库基础上开展数据库重建工作，通过数据重组、技术应用实现功能拓展，构建数字人文基础研究环境。

#### 4.2.1 分词技术实现标签云功能

分词技术主要用于对结构化、半结构化以及非结构化数据的细粒度化加工，是较成熟的信息处理技术。汉语分词方法主要依据词典分词法（字符串匹配）、统计分词法和理解分词法[14]。应用分词技术处理数据可有效提升数据细粒度，为数字人文研究的数据统计分析与深度挖掘工具应用提供方便。中国科学院计算技术研究所开发的商业化软件汉语词法分析系统ICTCLAS（Institute of Computing Technology Chinese Lexical Analysis System）[15]，是功能较强的非开源系统。主要功能包括中文分词、词性标注、命名实体识别、新词识别，不仅支持用户词典与繁体中文转换，还支持GBK、UTF-8、UTF-7、UNICODE等多种编码格式。此外，有许多开源软件可供使用，如类似开源LAMP平台的基于字符串匹配PHP Analysis分词组件工具[16]。

标签云是一种流行的可视化检索手段，具有较强的直观表现力。标签云数据来源可以是用户的标注、资源的元数据及全文。地方志数据库标签云数据取自题名以及主题字段内容，通过分词处理形成标签云数据；其采用开源LAMP平台的字符串匹配PHP Analysis分词组件工具进行分词，为实现标签云检索功能奠定基础。通过计算分词计数的四分位数来为不同词频的分词赋予权重，以标签大小设定字体大小。

目前项目组只针对结构化数据进行分词处理，在检索效果与呈现方式上有所改善。在非结构化数据以及文本分词方面还待进一步探索，如通过文本挖掘技术进行相关词频统计分析等功能，进一步对数字人文研究所需要的相关功能进行开发。

#### 4.2.2 GIS应用实现多元表现

GIS技术的应用模式有多种，包括桌面型、嵌入型、移动型以及Web型。Web模式的GIS应用较常见，如谷歌地图、百度地图、高德地图等。资源细粒化是数据库重建过程中的重要工作，主要通过增加元数据的时空信息、不同时期地点名称变化及关联数据的映射等数据处理，实现数据的多维展示。

根据地方志检索平台典型的B/S架构特点，项目组选择适合Web模式的GIS技术方案，地图平台选用高德地图，其云图API[17]通过云平台可提供基于位置的服务（Location Based Service，LBS），包括后端云图位置数据存储服务及前端云数据图层插件。使用者可以利用云数据图层插件将存储在LBS云中的数据作为一个图层叠加到地图上，利用云数据检索接口对自有数据进行空间检索。高德云图位置数据存储服务提供通过地理位置名称自动匹配地理位置经纬度的服务，使用者可以不必提供准确的经纬度信息，极大地减轻数据准备工作。高德地图平台API插件提供的地图检索、测距、热力图、区域面积计算等强大的接口功能，只需按接口要求准备相应所需数据，通过平台少量的配置及客户端编码，即可实现平台多维度可视化功能。

研究的下一步将在数据精加工基础上实现时间轴功能，并结合地域面积伸缩方式开展精细化、可视化研究。

## 5 结语

资源整合是数字图书馆发展的进阶，也是数字资源重建的一个侧重点，它是数字人文研究驱动下图书馆数字资源建设的重要内容，也是开展服务创新的基础。

图书馆参与数字人文研究，切入点无外乎资源和服务两方面。基于资源建设就是要通过更好地组织资源，并针对人文学者研究需求开展数据库建设和已有资源重建，进而实现从资源检索系统向研究环境建设的过渡。资源整合既能有效克服资源分散和孤立存储带来的资源难发现问题，又有助于提升量化分析的准确性。本文只是从异构特藏数字化资源的数据获取与可视化发现角度进行一些技术实现的尝试性研究。作为数字人文研究的通用研究方法，整合资源的多角度及关联关系揭示、可视化检索与展示、时空分析、文本分析以及社会关系分析等功能的实现，已经逐渐成为数字人文研究环境建设的必要部分，这些功能将为人文学者带来研究视野的拓展。

开展数字人文方面的思考与实践是学术图书馆必须把握的一个发展机遇，也是未来学术图书馆开发自身特藏馆藏的最佳途径。图书馆不仅是服务的提供者，更是合作者，要主动融入数字学术工作的生命周期，成为数字人文研究社群的一份子。图书馆作为资源组织

和管理者、研究工具和服务平台的提供者、跨学科协同合作的中立方，具有更大的发展空间[18]。

## 参考文献


[1] 全国师范院校图书馆联盟文献资源建设调查问卷[EB/OL].[2017-09-14].http://www.sojump.com/jq/4055719.aspx.

[2] HOCKEY S.The History of Humanities Computing[M]//A Companion to Digital Humanities.Blackwell Publishing Ltd,2004:1-19.

[3] 于淑娟.台大资讯工程学教授:新技术能为历史研究提供什么帮助[EB/OL].(2015-06-16)[2017-09-14].http://www.thepaper.cn/newsDetail_forward_1340177.

[4] UNSWORTH J.Scholarly Primitives:what methods do humanities researchers have in common,and how might our tools reflect this?[C]//Symposium on Humanities Computing:Formal Methods,Experimental Practice.London:King's College,2000.

[5] 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(11):13-30.

[6] Hathi Trust.Our Partnership[EB/OL].[2017-08-20].http://www.hathitrust.org/Partnership.

[7] Hathi Trust.HTRC Collections and Tools[EB/OL].[2017-08-20].http://www.hathitrust.org/htrc_collections_tools.

[8] 申斌,杨培娜.数字技术与史学观点——中国历史数据库与史学理念方法系统探析[J].史学理论研究,2017(2):87-95.

[9] 朱本军,聂华.互动与共生:数字人文与史学研究——第二届"北京大学数字人文论坛"综述[J].大学图书馆学报,2017(4):18-22.

[10] BERNERS-LEE T,BIZER C,TOM H,et al.Linked data:the story so far[J].International Journal on Semantic Web & Information Systems,2009,5(3):1-22.

[11] 司莉,李鑫.基于关联数据的科学数据集成与共享研究——以Bio2RDF项目为例[J]图书馆学研究,2014(21):51-55.

[12] 李广建,汪语宇,张丽.数字资源整合的实现机制及关键技术——对国外数字资源整合系统的实证研究[J].中国图书馆学报,2007(2):75-80.

[13] 吴业彤.基于XML的异构数据集成的方法研究[J].电脑知识与技术,2010,6(15):3872-3873.

[14] 袁璐,蒙祖强,许珂.依存分析和HMM相结合的信息抽取方法[J].计算机工程与应用,2012,48(9):138-140.

[15] 百度百科.ICTCLAS[EB/OL].[2017-09-18].https://baike.baidu.com/item/ICTCLAS/8609504?fr=Aladdin.

[16] 分词系统简介:PHPAnalysis分词程序[EB/OL].[2017-09-14].http://www.cnblogs.com/sanwenyu/p/4054728.html.

[17] 高德开放平台.产品介绍[EB/OL].[2017-09-14].http://lbs.amap.com/.

[18] 钱国富.英国高校图书馆数字人文服务探析——以兰卡斯特大学为例[J].大学图书馆学报,2017,35(4):30-34.


## 作者简介


李欣，女，1961年生，研究馆员，研究方向：图书馆信息化、数字人文，E-mail：xli@library.ecnu.edu.cn。
张毅，男，1986年生，硕士，馆员，研究方向：图书馆信息化、软件工程。
汪志莉，女，1983年生，硕士，馆员，研究方向：图书馆信息化、数据挖掘。


### Digital Humanities Research Demand of Library's Heterogeneous Special Resource Integration


LI Xin, ZHANG Yi, WANG ZhiLi
(East China Normal University Library, Shanghai 200062, China)



Abstract: Taking the problems in construction and service of the specially collected resources of the libraries as the starting point, from the perspective of the specially collected resources used for supporting digital humanistic, combining with the practices of the East China Normal University Library in the application of Web Services technology, GIS technology and structured data segmentation, the article has introduced the implementation methods of the specially collected resources integration, geographic location and tag cloud visualization retrieval.



Keywords: Heterogeneous; Special Resources; Digital Humanities

（收稿日期：2017-09-26）