

国内图书馆自建人物数据库的现状分析^{*}

徐晶晶, 葛红梅

(中国医学科学院医学信息研究所, 北京 100005)

摘要: 从收录对象、资源类型、服务功能等方面分析国内图书馆自建人物数据库的现状, 发现人物数据库建设虽然取得了一定成就, 但是还存在着资源收集不齐全、数据库完成度不高, 资源组织加工深度不够、功能相对单一, 互动性、开放性、共享性较差等问题, 进而从规范建库方案、深化资源组织、拓展服务功能等方面提出相应建议, 指出人物数据库建设应往规范化、关联化、可视化、社会化的方向发展, 变人物数据库为人物知识库。

关键词: 人物数据库; 特色数据库; 数据库建设

中图分类号: G250.74

DOI: 10.3772/j.issn.1673-2286.2015.12.009

人物数据库是对在某一方面乃至多个方面, 对人类社会做出特殊贡献的典型人物相关信息的集中收录、组织和展示, 一般具有较强的学术性、纪念性、史料性和观赏性^[1]。随着人们对人物研究工作的不断深入和社会对人物资料需求的不断扩大, 越来越多的图书资料机构将人物数据库作为一种特色资源进行研究和建设。

1 调研方法和对象

由于图书馆坐拥丰富的文献资源和文献处理经验, 在人物相关文献资源的搜集整理方面, 相比一般的公众网站、商业数据库, 在资源深度、数据规范性和获取方便性上更具优势。因此, 本文着重研究图书馆在人物数据库建设方面的工作。通过文献调研和网络搜索, 得到公共图书馆和高校图书馆自建的人物数据库28个(见表1), 主要调查了人物数据库的收录对象、资源类型、服务功能、访问方式等。

2 人物数据库建设现状分析

2.1 收录对象

从收录对象上可以将人物数据库分为单一人物数

表1 国内图书馆自建的人物数据库列表

建库方	人物库名称
天津图书馆	津门群星
上海图书馆	中国电影明星录
南京图书馆	百年人物
南京图书馆	纪念徐悲鸿诞辰100周年
金陵图书馆	孙中山研究
金陵图书馆	郑和研究专题
苏州图书馆	苏州名人
安徽图书馆	安徽院士
安徽图书馆	纪念雷锋50周年
安徽图书馆	毛泽东诞辰120周年
湖北省图书馆	湖北名人
上海交通大学图书馆	钱学森图书馆
上海大学	钱伟长特色数据库
湖南省图书馆	湖北近代人物数据库
温州市图书馆	温州历史名人数据库
吉林省图书馆	东北抗战时期人物库
吉林省图书馆	杨靖宇将军专题
暨南大学珠海校区图书馆	留学人物数据库
暨南大学珠海校区图书馆	容闳数字图书馆

^{*} 本研究得到中央级公益性科研院所基本科研业务费专项课题“医科院图书馆特色数据库建设研究”(编号: 12R0114)资助。

续表

建库方	人物库名称
北京大学图书馆	北大名师
清华大学图书馆	清华学者
人民大学图书馆	人大名师
中国石油大学(北京)图书馆	全国石油专家学者学术资源平台
中南大学图书馆	中南大学专家学者库
上海交通大学图书馆	交大名师
北京体育大学图书馆	运动人物库
广州体育学院图书馆	南粤体育名将库
广州图书馆	广州名人

据库、群体人物数据库和专家学者数据库。

2.1.1 单一人物数据库

单一人物数据库如钱学森图书馆、钱伟长特色数据库等,这类数据库的特点是围绕某一特定历史人物,系统全面的收集整理与其生平相关的各种资源,包括生平履历、亲属关系、学术成果、手稿、荣誉、照片、音视频资料以及围绕其开展的研究活动、纪念活动等。通过系统的搜集整理,以达到梳理其人生经历,保存其学术成果,明确其历史价值,传播其学术思想的目的。单一人物数据库侧重于个人资源的深入全面挖掘,尤以钱学森图书馆为代表。

2.1.2 群体人物数据库

群体人物数据库的诞生背景是群体传记学(Prosopography),通过对一群历史行为人的生平做集体性的研究,探讨这群历史人物共有的背景特征^[2]。这类数据库的建设意义在于,利用便捷的网络信息技术来管理海量的人物相关数据。国外比较有代表性的群体人物数据库有英格兰圣公会神职人员数据库、盎格鲁-撒克逊英国群体传记资料库、拜占庭帝国群体传记资料库、中国历代人物传记资料库等。其中,郝若贝教授(Robert Hartewell, 1932-1996)发起建设的关系型数据库“中国历代人物传记资料库”(China Biographical Database, CBDB)在中国学界产生的影响最大,该数据库拥有一系列数据结构方式、方法论支

撑及个性化的分析工具^[3],国内群体人物数据库建设颇受其启发。群体人物数据库因为涉及人物较多,数据量大,因此收录范围和主题的确定至关重要,表现出明显的主题特征。目前国内图书馆建设的群体人物数据库有以地域特征为主题的湖北近代人物数据库、津门群星等,以时间特征为主题的东北抗战时期人物数据库,以相同履历特征为主题的留学人物数据库、中国电影明星录等。

2.1.3 专家学者数据库

在群体人物数据库中,有一类是高校为了展示学术实力或进行教师管理而建立的专家学者库,收录本校任职的专家、学者、教职员工的各种学术资源。这类数据库有固定的收录范围和主题,在研究、教育和行政管理方面都具有更为直接的实用价值,在高校特色数据库建设中颇受青睐,可以分出作为第三种人物数据库,即专家学者数据库。目前建设比较成熟的专家学者库有北京大学图书馆的北大名师、清华大学图书馆的清华学者、人民大学图书馆的人大名师、中南大学图书馆的中南大学专家学者库等。

2.2 收录资源

目前人物数据库的建设基本涉及了人物相关资源的所有类型,可大致分为档案资源、学术资源和衍生资源三类(见表2)。资源格式涵盖文本、图片、音频、视频、网页、地图等多种形式。

表2 人物数据库收录资源类型

资源类型	资源内容
档案资源	基本信息(国籍、籍贯、性别、党派、学历、职称、生卒年等)、生平履历、档案、奖励、荣誉、家族亲友关系、传记评论、动态消息等
学术资源	学术论文、会议论文、专著、手稿、专利、科研项目、指导学生学位论文、讲座、讲义等
衍生资源	研究人员、研究机构、纪念活动、纪念场所等

具体到各类型人物数据库,收录的资源类型则有较大差别。一般而言,单一人物数据库资源类型覆盖较为全面,调研的单一人物数据库如郑和、孙中山、钱

伟长、钱学森、徐悲鸿等人物库均全面覆盖人物档案资源、学术资源和衍生资源三大类。

专家学者库主要收集人物的人物档案资源和学术资源,尤其注重人物学术成果的收录,如北大名师库侧重收录学者的专著成果,清华学者库全面收录学者的论文成果,中南大学专家学者库则收录期刊论文、会议论文、著作、指导研究生学位论文、专利、报告、科研项目、获奖情况等多种学术资源。

群体人物数据库在资源类型的覆盖方面整体较差,很多公共馆所建人物库仅是人物简历的罗列,如湖北人物、金陵人物、安徽院士等。一些特定主题的群体人物数据库的资源类型表现出相应的倾向性,往往只收录与主题相关的部分资源,如留学人物数据库收录人物基本信息和地理信息,东北抗战时期人物库收录人物基本信息、相关事件和相关图片。但是总体来说,国内的群体人物数据库在资源的类型和数量上均与其他两类人物库有较大差距。

2.3 服务功能

2.3.1 总体功能分析

调研发现,目前人物数据库提供的功能都比较简单,如表3所示。其中,“高级检索”可检索的字段较少,无组配检索、限制检索、相关检索、二次检索等功能。除上海交通大学的交大名师库可以按照姓名正序/倒序对检索结果排序外,其他人物库均无对检索结果排序、导出功能。

在资源获取方面,钱伟长特色数据库、全国石油专家学者学术资源平台、中南大学专家学者库、交大名师、南粤体育名将库等有校园网访问限制,广州名人库需要广州图书馆的读者账号登录。其他可访问的人物数据库中,除了已经进入公共领域的资源可以直接全文展示外,大多数学术资源为文摘,且不提供资源获取方式,只有人大名师库提供资源全文,清华图书馆提供资源来源链接,北大图书馆提供资源馆藏OPAC链接。

2.3.2 特色功能分析

虽然目前人物数据库的功能都较为简单,但是在建库思路仍有值得借鉴的地方。

(1) 与实体馆藏关联

人物数据库的建设依托于图书馆的馆藏,而图书馆实体馆藏是图书馆资源的重要组成部分,两者结合,有利于提高图书馆的资源利用率。北大名师库将本馆有入藏的图书资源与本馆OPAC关联;钱学森图书馆(网站)通过360度全景技术(Virtual Reality, VR)模拟出可交互的钱学森图书馆(实体)的三维空间场景,并辅以文字和图片揭示人物生平履历、交游、学术活动信息以及资源全文^[4]。

(2) 自建资源内部关联

人物数据库作为自建资源的一种,在与其他自建资源进行数据关联方面更具优势,通过设置通用数据规范和著录规则,可以方便的进行跨库关联、检索。东北抗战人物数据库是东北抗战文献数据库的一部分,通过相关事件和相关图片字段分别与东北抗战事件库、东北抗战图片库相关联。津门群星通过相关事件、相关故事、相关图片、相关作品字段与事件库、故事库、图片库、作品库等字库链接。湖南近代人物数据库更是设计了人物库和文献库两大子库,人物库包含湖南近代人物数据库、研究人员数据库、家族亲友数据库三个子库,文献库包含著述成果、研究评论库、动态消息库、研究刊物、研究机构库、文献图片库、人物照片库、史迹场所库、音视频库等七个子库,各子库间相互关联又相对独立。

(3) 揭示人物之间关系

人物之间存在着多种关系,如合作关系、亲友关系、师生关系、研究关系等,通过研究人物之间关系可以了解人物学术思想产生缘由、师承脉络等。清华学者库通过数据挖掘和分析,提供以学者为中心的合作者网络;中南大学专家学者库收录指导学位论文资源,侧面反映了师生关系;湖南近代人物数据库作为典型的群体人物数据库,收录湖南近代人物1600余人,家族亲

表3 人物数据库检索功能统计

检索功能	未知	仅浏览,无检索	有检索但不能使用	标题检索、全文检索	字段检索、高级检索	总计
数量	2	8	2	9	7	28
百分比	7.14	28.57	7.14	32.14	25.00	100

友、研究人员300余人,人物之间关系的揭示是该数据库建设的一大特色和难点。该库利用知识图谱建立数据库内部人物与资源、人物与人物之间的关系,通过建立名门望族和人物籍贯两个辅助词表,并在个人元数据中标注与其相关的亲友和研究人员,将人物和家族、亲友、研究人员关联起来,形成网状结构的群体人物数据库^[5]。

(4) 人物与地理信息关联

人物数据库中有许多关于籍贯、民族、国籍等地域信息的描述,对群体人物进行地理空间分析,能够快速发现空间分布模式与社会关系网络,从而支持基于群体传记法的历史研究。留学人物库将人物与地理资料关联,利用GIS技术揭示人物的地域特征,如留学名人足迹功能,将留学人物个体的历史足迹显示在Google地图上,从而反应其在一段历史时期内的时空关系;留学人物分布功能将用户查询到的一批留学人物的地域信息显示在Google地图上并加以分析,从而得到地理空间分布模式^[6]。

(5) 人名清理和统计分析

人物数据库资源收集的一大难点是人名的清理,即要解决怎样区别同名人物,怎样将同一人物的不同署名的作品汇集的问题。为此,清华学者库借鉴汤森路透researchID的理念,建立了清华学者学术唯一标识(ThuRID),以清华大学购买的学术文献资源为基础,利用数据挖掘的理念和方法,自动甄别出清华大学目标学者,并为挖掘到的清华学者建立学术唯一标识(ThuRID),获取目标学者的学术文章、与目标学者紧密关联的合作者、期刊会议等信息,应用开放链接技术定位到全文。在此基础上,提供发文趋势、合作者关系、发文期刊、关键词的统计分析,采用可视化方式直观展示学者的学术历程,以及以学者为中心的科研网络^[7]。

3 人物数据库建设存在的问题

虽然目前国内图书馆人物数据库建设数量较多,角度各异,个别项目不乏亮点,但是整体看来仍存在一定问题。

3.1 资源内容不齐全,数据库建设完成度不高

人物数据库建设的一大问题在于资料收集的准确

性和完备性,准确性是指收集的人物资源要确定为该人所有,即同名人物的区分。完备性要求汇集与人物有关的各种形式名称的所有资源,古代人物有名、字、号、谥号、官职代称、地名代称等,现代人物发表论文存在姓名翻译的多种形式,如姓前名后、姓后名前、韦氏拼音、汉语拼音、双字名中间有连接符等,其他还有诸如笔名、更名的问题。人名的区分与鉴别给人物资源的收集带来较多麻烦,再加上资源类型庞杂,资源来源多样,造成资源收集不齐全,数据库完成度不高,具体表现为完成建设的人物较少,或资源类型较少,如清华学者库基本实现了设计的功能,但是目前仅完成了150余位人物的数据,更多的群体人物库中的许多人物名称下并无实质内容,即便是单一人物数据库在人物资源的完备性上也仍有待提高。

3.2 资源组织加工深度不够,功能相对单一

目前的人物数据库更多的是作为资源的提供者,多数公共馆所建数据库仅是通过浏览和简单检索提供简单的人物信息,高校馆建设的专家学者库一定程度上收录的资源较为丰富,但是对资源的描述和组织仍是过于简单,一些图书馆对常见文献资源类型如图书、论文、专利等使用了MARC、DC等元数据描述,但是更多的非传统文献资源如项目、机构、事件、图片、视频、音频、新闻报道、史迹场所等由于没有相应的元数据规范,仅提供资源的名称和简要说明,缺乏深层次的信息挖掘,内容加工深度不够,从而导致检索功能简单。

3.3 互动性、开放性、共享性较差

人物数据库建设的目的是充分发挥图书馆的资源优势,最大限度地满足用户的信息需求,为更多用户提供个性化服务,在更广范围内实现数据库资源共享。但是从目前的调研来看,多数资源较丰富的数据库均存在各种形式的访问限制,资源获取不便;提供个性化辅助研究和社会化分享评论等增值服务的数据库很少,仅北京体育大学运动人物库有评论和评分功能,钱学森图书馆有社交软件分享功能,北大名师提供人物点击排行榜功能,其他诸如文献传递、在线咨询、知识发现、资源推送、个性化定制服务均没有,数据库使用体验和效果较差。

4 人物数据库建设的发展建议

4.1 规范建库方案

建库前做好可行性分析, 慎选主题, 严格控制人物收录范围和资源收录范围。建库时尽量采用通过认证的软件平台和系统, 在数据格式上以国际和国内标准为依据, 遵照数据交换协议和数据加工整理所采用的标准规则, 建设兼容性强、标准化高的人物数据库。通过版权转让协商、链接资源来源、数据库分层次发布和访问权限分级等方式解决版权带来的访问限制问题。提供数据开放接口, 支持用户或第三方系统获取特定资源, 以支持面向公众和科研需求的数据共享和再利用。

4.2 深化资源组织

建立人名唯一标识系统和人名规范库, 由图书馆员、人物本身、人物相关人员多方参与, 共同对人物信息进行清洗、区分, 做到人物与资源对应关系准确, 不遗漏、不错乱。人物数据库涉及各种类型的资源, 如专著、论文、专利、项目、机构、事件、图片、视频、音频、新闻报道、史迹场所等, 拥有大量的人物、事件、项目、机构、地理、学科、出版等数据, 对这些资源建立相应的元数据规范和著录规则, 按照规则进行元数据标引, 并从中对相关内容进行本体化实体抽取标注, 形成以人物、项目、机构、地理、学科、出版、事件等为知识内容的元数据仓储, 构建不同类型的知识库, 将内容资源全面系统地串联起来, 构建人物知识体系^[8]。同时将人物数据库资源与图书馆已有实体资源、自建资源、商业资源整合, 构建统一检索发现平台。

4.3 拓展服务功能

在元数据标引的基础上, 将数据研究和技术支持相结合, 整合底层数据资源, 利用智能化的知识分析统计功能, 通过推理机制在人物数据库中寻找满足条件

的数据资源, 为用户提供可视化形式的面向问题的直接答案, 满足用户的知识需求, 小则能够回答用户的事实型和数据型问题, 大则能够提供趋势分析报告。利用RSS订阅和资源推送功能主动推送资源, 加入用户对信息资源的组织(标签)和评价(评论与注释)等知识化功能, 提供用户问答、收藏、分享、评分等互动功能。引入情境智能设计, 根据用户访问的授权方式、具体资源存在与否等类似情境自动做出识别判断操作, 提供给用户最合理便捷的资源检索和获取服务。

大数据时代, 信息技术的发展, 人物数据库不应该只停留在百科式的人物信息陈列, 应该往规范化、关联化、可视化和社会化的方向发展, 利用语义工具和数据挖掘手段, 全面收集整理人物资源, 揭示人物数字资源中概念内容和实体间的关联关系, 实现不同资源、事件、人物、记录之间的链接和集成, 构成并可可视化反映人物知识网络, 变人物数据库为人物知识库。

参考文献

- [1] 蔡璐. 浅谈人物数据库网站的核心表现——以湖南近代人物资源库建设为例[J]. 图书馆, 2010(3): 107-108.
- [2] 群体传记学[EB/OL]. [2015-11-20]. <http://isites.harvard.edu/icb/icb.do?keyword=k35201&pageid=icb.page499404>.
- [3] China Biographical Database(CBDB)[EB/OL]. [2015-12-09]. <http://isites.harvard.edu/icb/icb.do?keyword=k35201&pageid=icb.page145374>.
- [4] 钱学森图书馆[EB/OL]. [2015-11-20]. http://www.qianxslib.sjtu.edu.cn/d4/flash/TourViewer_qxs_d4.html.
- [5] 张文勇. 湖南近代人物数据库建设研究与实现[D]. 长沙: 中南大学, 2013.
- [6] 黄勇. 留学人物数据库GIS应用的设计与实现[J]. 现代图书情报技术, 2012(5): 91-95.
- [7] ThuRID (Tsinghua University Researcher ID) 服务目标[EB/OL]. [2015-11-20]. <http://rid.lib.tsinghua.edu.cn/thurid/constructionGoals.pdf>.
- [8] 周杰, 苏静, 曾建勋. 下一代数字图书馆的发展思考[J]. 图书情报工作, 2013(8): 35-39.

作者简介

徐晶晶, 女, 1989年生, 中国医学科学院医学信息研究所助理馆员。

葛红梅, 女, 1979年生, 研究方向: 图书馆信息组织、图书馆元数据建设, 通讯作者, E-mail: ge.hongmei@imicams.ac.cn。

The Figures Database Construction in Domestic Libraries

XU JingJing, GE HongMei

(Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100005, China)

Abstract: The objects of collection, types of resources and functions of the figures databases built by domestic libraries were analyzed, which showed that although certain successes have been achieved in the construction of figures databases, there are some problems needing to be solved, such as the incompleteness of the construction, the insufficient organization of the resources and the poor accessibility of the database. Suggestions are put forward to for the solution of these problems, such as regulating the process of database constructions, sufficient organizing of the resources and expanding database functions, and directions were pointed that figures databases should be built in a standardized, associated, visualized and socialized way.

Keywords: Figures Database; Characteristic Database; Database Construction

(收稿日期: 2015-11-27)

2016年全国知识组织与知识链接学术交流会 征文通知

为了探讨网络环境下知识组织与知识链接的新理念、新思路、新方法, 中国科学技术信息研究所、国家科技图书文献中心、中国科学技术情报学会拟于2016年下半年召开第七届“全国知识组织与知识链接学术交流会”, 由《数字图书馆论坛》承办。特向国内图书情报界及相关领域的专家学者征文。具体会议时间、地点另行通知。

一、会议主题

知识组织

- 知识组织研究进展
- 大数据中的数据清洗、组织和分析
- 事实型数据识别与分析
- 叙词表、本体等知识组织体系的构建和应用研究

知识评价

- 科学计量与评价
- 引文分析、主题分析
- 专利分析与利用
- Web科技资源评价

知识链接

- 科研实体关系揭示
- 网络资源链接及其关联分析
- 面向项目研发产出的关联分析
- 数据关联挖掘和揭示

知识服务

- 知识服务、知识管理研究进展
- 数字科研环境与开放共享服务
- 用户分析与个性化用户服务
- 知识图谱及可视化分析

二、联系方式

会议邮箱: KOLink@istic.ac.cn

会议网址: <http://168.160.16.186/conference>