北京大学图书馆长期保存系统建设与探索

□张乃帅* 孙超

摘要 数字资源作为图书馆馆藏资源的重要组成部分，其采购经费在图书馆资源建设经费中所占的比重越来越大。与纸质资源相比，数字资源对存储介质及网络的依赖性非常强。一旦存储介质损坏或者因各种原因导致网络中断，数字资源将无法获取和使用。文章以北京大学图书馆长期保存系统建设实践出发，从系统建设概况、长期保存的技术保障、长期保存的实践探索、长期保存实践中的问题与未来展望等方面全面介绍了长期保存系统建设情况，并对数字人文资源的长期保存难点进行了探索。

关键词 数字资源保存 长期保存 数字人文
分类号 G250.74
DOI 10.16603/j.issn1002－1027.2019.02.011

数字资源作为馆藏资源中的重要部分，其采购经费在图书馆资源建设经费中所占比重越来越大。以北京大学图书馆（以下简称北大图书馆）为例，2014年购买数字资源的经费占资源建设经费的38%，2016年已上升至65%，比重大幅上升。澳大利亚的维多利亚大学图书馆在其2016－2020年的战略规划中提到，到2020年该馆新购的信息资源将是100%电子化的。作为馆藏资源中的重要组成部分，各图书馆越来越重视数字资源的揭示，越来越多的图书馆建设了资源发现系统，以期能够最大限度地揭示馆藏数字资源，提高数字资源的使用率，更好地服务读者。然而，与纸质资源相比，数字资源对存储介质的依赖性非常强，存储介质非常脆弱，一旦受到破坏或者损伤，所承载的内容就无法获取和利用，这使得数字资源面临着非常大的消失和不可获得的风险[1]。

2000年12月，美国国会为国家数字信息基础设施和保护计划（National Digital Information Infrastructure and Preservation Program，以下简称NDIIPP）拨款1亿美元，用于收集、保存重要的数字内容并确保其长期可用，建立和加强合作伙伴网络，并协同开发一系列的工具和服务技术框架，用于支撑长期保存。该计划由美国国会图书馆领导，通过与美国国家科学基金会、斯坦福大学、州政府等众多机构建立合作伙伴关系，对WEB信息、音频、视频、数字期刊、电子书、数字电视、州政府数字信息等多种类型的数字资源开展长期保存研究和实践。该计划还建立了完善的资助制度，鼓励对新型数字资源开展保存研究和实践[2][3]。作为一个国家级的项目，该计划建立起了成熟的合作保存机制，形成了广泛的社会参与，并不断把新型数字内容纳入保存体系当中，具有很好的借鉴意义。

“大量拷贝确保数据安全”（Lots Of Copies Keep Stuff Safe，以下简称LOCKSS）项目是由斯坦福大学图书馆发起的开源的、由图书馆主导的长期保存系统，其系统设计原则是大量拷贝确保数据安全。LOCKSS系统的参与者包括出版商、图书馆和用户。出版商通过发布LOCKSS权限声明和资源清单对允许保存的内容进行限定；图书馆在本地部署LOCKSS BOX，根据出版商的权限声明和资源清单获取和存储出版商的内容，并将本地LOCKSS BOX注册加入到LOCKSS分布式保存网络；用户在出版商内容因故（网络拥塞、退订、自然灾害、战争等）不能访问时通过本地LOCKSS BOX获取内容。一旦数据摄入完成，LOCKSS BOX中的内容将不再依赖数据库商，通过不断与分布式保存网络中其他LOCKSS BOX节点中的相同内容进行对比及同步，

* 通讯作者：张乃帅，ORCID：0000-0002-8041-3571，邮箱：zhangns@lib.pku.edu.cn。

LOCKSS BOX 确保本地保存的内容始终是正确的。目前，已经有超过 530 家出版商加入了 LOCKSS 全球保存网络，另有大量机构创建了 LOCKSS 私有网络保存机构的特殊数字内容[4]。LOCKSS 保存系统有众多的出版社及图书馆参与，在长期保存领域具有很大的影响力，值得国内保存系统学习和借鉴。

“柱廊”（Portico）项目不同于前述长期保存系统，是由独立于出版商和图书馆的第三方提供的保存服务。Portico 保存服务是非营利机构 ITHAKA 的一部分，截至 2018 年 8 月 25 日，Portico 已与 554 家出版社和 1013 家图书馆开展合作，获取授权保存期刊 31379 种、电子书 1246248 种，已保存期刊 26808 种、电子书 918893 种[5]。

与国外长期保存现状不同的是，目前国内各图书馆在数字资源的长期保存方面投入的经费及关注度远远不够，并未引起足够重视。

北大图书馆于 2016 年承建国家数字科技文献资源长期保存体系（National Digital Preservation Program，以下简称 NDPP）北京大学节点建设项目，并以项目为依托，组建了由馆长牵头、两位副馆长分头负责的长期保存项目团队。项目团队成员来自信息化与数据中心及中国高校人文社会科学文献中心（China Academic Social Sciences and Humanities Library，以下简称 CASHL）管理中心，在资源谈判、软件开发及运行维护领域积累了丰富经验。同时，以项目为依托，除了完成项目约定的国外重要数据库的国内保存以外，逐渐向馆藏资源辐射，与资源建设中心合作探讨馆藏数字资源的长期保存事宜。

本文将以北大图书馆在长期保存方面的工作实践为基础，阐述长期保存体系的建设经验，从系统建设概况、长期保存的技术保障、长期保存的实践探索、长期保存实践中的问题与未来展望等方面进行介绍，以期能为更多图书馆的长期保存系统建设提供经验和借鉴。

1　长期保存系统建设概况

众所周知，大部分外文数据库的服务器都位于境外，且在境内没有镜像服务器。一旦因网络拥塞、自然灾害、战争、政治因素等原因导致出境网络中断，外文数据库将无法访问。这将使大量经费购买的国外数据库无法产生科研和社会效益，直接影响我国的科研、教育和创新环境，对国家科技自主创新能力和国家科技安全造成影响。为此，科技部于 2013 年批准由国家科技图书文献中心（National Science and Technology Library，以下简称 NSTL）牵头组织实施，以 NSTL 主要成员单位和少数重要高校为核心，进行国家保存体系的建设工作，NDPP 应运而生。NDPP 由管理机构和保存节点构成，管理机构为 NSTL，保存节点包括中国科学院文献情报中心、中国科学技术信息研究所和北大图书馆。

保存节点每季度召开例会，汇报各节点在资源谈判、资源保存方面的工作进展及存在的问题，并就已发现问题的解决进展进行说明。NDPP 还建立了完整性检查制度和审计制度，确保各节点对签署保存协议的数字资源进行了准确、完整、有效的保存。保存节点还形成了联合谈判机制，对部分配合度低、谈判进展缓慢的数据库商开展联合谈判。

作为 NDPP 的参建节点和唯一的高校保存节点，北大图书馆重点保存基础科学、跨学科领域和高科技领域的数字资源，也涉及社会科学相关资源的长期保存，同时承担探索新型数字资源如数字人文资源长期保存方案的任务。根据项目组成员所承担的任务不同，北大图书馆组建了权益谈判团队、系统运行团队和软件开发团队，分别承担数字资源的保存权益谈判、保存系统的稳定运行及新增数字资源的摄入插件开发等任务。

北大图书馆长期保存系统采用了由保存体系承建单位中国科学院文献情报中心研发的基于 Fedora 仓储的数字资源长期保存系统（Digital Preservation System，以下简称 DPS）。有关 DPS 的系统架构，付鸿鹄等在《分布式数字资源保存系统与技术架构研究》一文中已经详细论述，在此不再赘述。

经过两年多的实践和探索，北大图书馆在资源权益谈判、插件开发和资源保存方面均取得了较大进展，与 Emerald 期刊数据库、ProQuest 硕博士论文数据库签署了长期保存协议，开发 Emerald 数据摄入插件一个，保存 Emerald 期刊 305 种、258506 篇，获取 Proquest 硕博士论文 71.6 万篇。并根据工作需要，开始在馆藏数据资源和新型数字资源长期保存方面开展研究和探索。

2　长期保存的技术保障

长期保存作为一个复杂的系统工程，需要来自技术、政策、组织等多个层面的保障。其中技术层面

包括系统部署、网络安全、系统备份、插件开发和数据更新等，用于确保数据真正做到“长期”保存，可谓长期保存系统的基础。

2.1 系统部署

长期保存系统建设的第一步是系统部署。系统部署需要根据 DPS 系统要求，结合馆内的网络、存储、服务器环境，制定部署架构及方案，确保长期保存系统在系统性能、网络安全等方面满足设计需求。最终，北大图书馆将 DPS 系统部署在两台物理服务器上，一台服务器部署 web 服务器、数据库及索引服务，另一台服务器直连存储服务器，用于数据存储。长期保存系统的首要任务是对资源进行可靠保存，平时不对外提供服务，为确保服务器的可靠稳定，在长期保存系统前端与校园网之间架设了防火墙，对服务器进行严格的访问控制。

2.2 系统安全

DPS 系统采用了大量的开源组件进行建设，而开源组件面临的一项重大挑战是源代码对所有人开放，一旦开源组件出现安全漏洞，漏洞即对所有人可见且漏洞特征将会非常明显。开源组件的漏洞如果被别有用心的攻击者利用，造成的损失不可估量。虽然 DPS 系统位于防火墙后，不会受到直接攻击，但是目前仍与其他服务器处于同一个网络环境，一旦其他服务器存在安全漏洞被攻击者利用，DPS 系统仍将受到威胁。为了尽早发现 DPS 系统存在的漏洞，降低受到网络安全威胁的概率，确保长期保存系统的数据安全，系统运行团队定期对 DPS 系统进行网络安全扫描和渗透测试，如果发现新的漏洞，第一时间与开发团队沟通，获取漏洞解决方案并进行相应的网络安全升级。通过网络安全扫描和渗透测试，北大图书馆共发现命令执行、注入、WebShell 等类型高危漏洞 6 个，通过与开发团队合作，及时封堵了漏洞，清除了潜在威胁。

2.3 数据备份

除了网络安全扫描以外，数据备份是另一项对长期保存系统数据安全至关重要的维护任务，主要应对硬件故障及网络攻击等带来的数据损坏和丢失。目前，系统运行团队根据长期保存系统的特点及备份系统架构，制定了在线磁盘备份和离线磁带库备份两种备份策略，在线磁盘备份可进行快速恢复，保留的备份周期较短；离线磁带库备份恢复周期比磁盘备份恢复周期长，但是能保存较长的备份周期。

目前，北大图书馆仅有一个数据中心，距离金融系统的“两地三中心”运营安全体系尚有较大差距，无法应对灾难级故障。为了提高安全系数，北大图书馆正在规划建设“同城异地数据中心”，将备份数据放置于同城其他校区的数据中心内，避免因一个数据中心遇到灾难级故障导致数据丢失的极端情况发生。

2.4 插件开发

由于不同电子资源的数据类型不同、数据格式不同，这些数据要存入长期保存系统，需要不同的数据摄入插件做支撑。对于 DPS 系统已经支持的电子资源类型如期刊、电子书等，通过分析数据库商提供的样例数据形成新增资源格式分析报告，以格式分析报告为基础，调用 DPS 系统提供的接口开发数据摄入插件。开发完成并测试通过以后，部署到 DPS 服务器，用于新增资源的数据摄入。

对于首次保存的资源类型如 ProQuest 硕博士论文，目前的底层数据模型并不能满足保存需求。通过调研学位论文相关元数据标准，北大图书馆提出学位论文类型电子资源的保存规范，并与中国科学院文献情报中心开发团队进行了深入沟通。后续将在中国科学院文献情报中心开发团队对底层数据模型进行调整后及时开发 ProQuest 硕博士论文摄入插件。

2.5 数据更新

长期保存系统最核心的常规工作是根据保存协议的约定周期定期获取电子资源的更新数据并上载至 DPS 系统。为规范数据来源，北大图书馆统一通过 FTP 服务器向 DPS 系统提供保存资源的数据更新。FTP 服务器上的数据来源，根据数据量大小、数据库商的数据传递策略等多种因素的不同，有多种更新途径，包括硬盘更新、FTP 更新等。如通过硬盘更新数据，在获取硬盘并校验硬盘数据后由项目组成员上传至 FTP 服务器；如通过 FTP 更新数据，则在 FTP 服务器上向数据库商服务器发起 FTP 下载请求获取更新数据。为确保更新数据的安全可靠，通过配置防火墙策略，仅允许 FTP 服务器对外发起请求，不允许外部服务器向 FTP 服务器发起请求，尽量降低 FTP 服务器被攻击的可能性。

3 长期保存的实践探索

经过两年的建设和努力，北大图书馆长期保存系统在权益谈判、数据建设等方面均取得丰硕成果，并着手探索数字人文资源及馆藏数字资源的长期保存。

3.1 权益谈判

北大图书馆组建了由主管副馆长及CASHL管理中心成员构成的权益谈判团队，负责重要数字资源的保存权益谈判。团队成员均主持及参与高校图书馆数字资源采购联盟(Digital Resource Acquisition Alliance of Chinese Academic Libraries，以下简称DRAA)的日常工作，对数据库资源非常了解，在资源采购谈判方面具有丰富经验。同时借助DRAA理事会等渠道，能够获得DRAA各牵头馆的广泛支持，而且能够扩大保存体系的宣传途径和影响力。

权益谈判团队经过漫长谈判和不懈努力，成功签署Emerald期刊数据库保存协议、ProQuest硕博士论文数据库保存协议。其中，ProQuest硕博士论文数据库保存协议是NDPP中首次签署学位论文类型的保存协议，在保存资源类型和保存数据量上均取得突破性进展。权益谈判团队积极推动与Elsevier公司的谈判进程，目前双方已基本达成一致，即将进入实质性操作阶段。与Taylor & Francis公司的谈判也在持续进行，公司董事会支持NDPP项目，双方正就协议内容展开讨论。在牵头开展电子资源采购过程中，北大图书馆积极推动长期保存谈判，已与“一带一路专题数据库”“南亚研究回溯数据库”“美洲回溯文献典藏数据库”三个数据库提供商达成向北大图书馆提供长期保存数据的意向。此外，权益谈判团队还向Brill发出了保存要约。

3.2 长期保存数据建设

截至2018年8月25日，北大图书馆长期保存系统已完成Emerald 2017年前回溯数据的保存工作，共保存期刊305种、全文258506篇；已获得ProQuest硕博士论文全文71.6万篇，由于底层数据模型及数据摄入插件尚未调整及开发完成，ProQuest硕博士论文还未进行保存。

3.3 数字人文资源及馆藏资源的长期保存实践

数字人文是计算机学科和人文学科交叉研究的一个新领域，由计算人文和人文计算领域发展而来。对数字人文学科本质的认识一直存在不同观点，其中一个被广泛引用的典型解释是：数字人文是针对计算工具与所有文化产品交叉领域的研究[6]。中国历代人物传记资料库(China Biographical Database，以下简称CBDB)是由哈佛大学费正清中国研究中心、北京大学中国古代史研究中心、台湾“中央”研究院历史语言研究所共同主持的学术数据库。截至2018年8月，CBDB共收录41.7万人的传记资料，是数字人文领域具有深远影响力和极具代表性的学术项目。经过沟通，CBDB项目组已同意在北大图书馆设立CBDB镜像站点，将CBDB数据在本地保存。项目组也已原则上同意北大图书馆将CBDB数据长期保存，详细条款正在进行沟通探讨。

Gale数据库整合了多种来源的信息，收录了跨越全球500年历史的大量原始档案一次文献，涉及包括经济、历史、社会、国际关系、文学、地理、政治、法律等在内的丰富的学科主题。北大图书馆于2017年订购了Gale数据库，在订购时即注重数据的本地存储，在签订合同时明确约定全部数据在本地进行备份存储。长期保存系统运行团队已于2018年6月完成Gale数据库平台全部数据的获取和本地存储工作，共存储文件1.82亿个，数据量103T。目前，北大图书馆项目团队正在与资源建设中心、Gale集团探讨将Gale数据长期保存的可行性。

4 长期保存实践中的问题与未来展望

经过两年的探索和实践，北大图书馆在长期保存系统建设方面取得了一定成果，同时也发现了一些问题，制约着长期保存系统的建设和发展。

4.1 数据库商提供的回溯数据和更新数据格式不一致

数据库商提供的回溯数据和后续提供的更新数据，在数据格式方面有时候会存在差异，为此，需要开发两个版本的数据摄入插件，一个版本用于摄入回溯数据，另一个版本用于后续的常规数据更新。这种状况除带来额外的开发工作量，也可能造成同一数据库保存的数据项前后不一致。造成这种状况的原因，一部分跟数据库商原始数据本身存在差异有关，另一部分也跟图书馆和数据库商之间的数据格式约定不严格有关。后续建设过程中，应从权益谈判阶段开始关注电子资源的数据格式，必要时将插件开发人员引入权益谈判团队，尽量从源头避免回溯数据与更新数据不一致的问题。

4.2 部分功能需手动启用

由于系统本身的架构设计原因，北大图书馆长期保存系统的部分功能需要在服务器后台通过执行特定命令开启，无法通过管理界面直接使用。这导致长期保存系统在使用及运行过程中需要进行人工干预，自动化程度有待提高。

4.3 底层数据模型兼容性较差

由于DPS系统最初设计面向的保存类型主要是电子书和电子期刊，底层数据模型对其他类型的数字资源比如学位论文兼容性较差。对学位论文类型的数字资源进行保存，首先要调整底层数据模型，然后才可以进行数据摄入插件开发及保存，耗时周期长，时效性较差。

4.4 数字人文资源保存难度大

数字人文研究的基本方法为社会网络分析、文本分析、空间分析和时序分析。社会网络分析是一门对社会关系进行量化分析的艺术和技术，它要求有较高的统计学、数学功底，以及计算机编程技术和能力等[7]。文本分析是指利用数据挖掘、机器学习、统计学、自然语言处理、可视化技术等多学科领域的技术和方法，对文本数据进行抽取进而发现新颖、有趣的知识[8]。空间分析和时序分析经常被结合使用，以地理信息系统(GIS)为依托，利用GIS技术的空间数据采集、时空数据建模、多层地图叠加功能，分析不同时间切面中的地理、社会、自然之间的关系，探索发展演变规律[9]。

通过数字人文研究的基本方法可以看出，数字人文资源除了包括文本、图像、音频、视频等传统数字对象外，还包括图论语言和技术、数学模型、计算机模拟软件、数据挖掘算法、自然语言处理技术及软件、地理信息系统等大量技术工具。这一点与传统数字资源有很大不同。传统数字资源如期刊、电子书等，一般具有规范的元数据标准和全文，长期保存系统只需设计出相对固定的底层数据模型，配合不同的数据库摄入插件，即可完成大部分期刊、电子书数据的保存，而且新增数据相对独立，可以认为与已保存数据没有直接关系。而数字人文资源与人文研究过程紧密相连，是动态变化的、带有时间序列的，变化本身是连续的、不可分割的，甚至这种变化本身也是数字人文所关注的，且每种不同的数字人文资源，其基础数据和所采用的技术工具都存在很大不同。如何设计一种灵活的数据模型，能够在保存数字人文资源时体现其动态变化过程，并能将其依赖的技术工具加以保存或说明，且能满足大部分数字人文资源的保存需求，是数字人文资源长期保存面临的极大挑战，需要经历长时间的探索。

如前文所述，数字资源已成为教育科研的主要资源，世界各国已开始对数字资源的长期保存进行战略部署。但由于数字资源内容增速快、规模大、结构复杂、格式多变，给长期保存和永久利用带来了极大挑战。北大图书馆在参与国家科技部“国家数字科技文献资源长期保存体系”项目的过程中，积累了一定经验，更体会到这是一项复杂的长期的任务，目前尚有许多技术、政策、组织等方面的问题需要解决，需要更多的机构参与进来，共同推动此项工作。

参考文献

1 陆泉，韩雪，韩阳，陈静. 我国数字信息资源长期保存研究综述[J]. 图书馆学研究，2015(4)：2－8.
2 DigitalPreservation[EB/OL].[2018－8－25]. http://www.digitalpreservation.gov.
3 LoC[EB/OL].[2018－8－25].https://www.loc.gov.
4 LOCKSS[EB/OL].[2018－8－25].https://www.lockss.org.
5 Portico[EB/OL].[2018－8－25].https://www.portico.org.
6 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.
7 汤汇道. 社会网络分析法评述[J]. 学术界.2009(3)：205－208.
8 郭金龙，许鑫. 数字人文中的文本挖掘研究[J]. 大学图书馆学报.2012(3)：11－18.
9 夏翠娟. 中国历史地理数据在图书馆数字人文项目中的开放应用研究[J]. 中国图书馆学报.2017(2)：40－53.

作者单位：北京大学图书馆，北京，100871
收稿日期：2018年9月5日

Construction and Exploration of Long-term Preservation System of Peking University Library

Zhang Naishuai　Sun Chao

Abstract: As an important part of the library's collection, digital resources account for a larger proportion of the library's collection development budget. Digital resources are highly dependent on storage media and networks compared to paper resources. Once the storage media is damaged or the network is interrupted for various reasons, digital resources will not be accessible and available. Based on the experience of the digital preservation system construction of Peking University Library, the paper introduces the construction of digital preservation system from the aspects of system construction overview, technical support, practical exploration, problems and future prospects. The difficulties of digital humanity resources preservation are also discussed.

Keywords: Digital Preservation; Long-term Preservation; Digital Humanity