

国家数字图书馆网络资源保存的实践与探索^{*}

张炜, 敦文杰, 周笑盈
(国家图书馆, 北京 100081)

摘要: 互联网环境下, 网络资源已成为人类文明的重要载体。作为我国公共文化服务体系的重要组成部分, 国家数字图书馆一直重视网络信息的保存与服务, 自2003年开展试验性项目以来, 经过十多年的积极探索和发展, 在网络资源遴选、管理平台搭建和服务方式创新等方面已积累丰富实践经验。通过对国家数字图书馆网络资源保存工作实践的分析, 以期为业界提供有益参考和借鉴。

关键词: 网络资源保存; 国家数字图书馆; 互联网; 文化服务

中图分类号: G250.76

DOI: 10.3772/j.issn.1673-2286.2017.06.005

1 引言

人类的发展离不开信息传递与交流, 以互联网为代表的信息技术日新月异, 世界正处在信息化快速发展的进程中, 信息的生产、传播和利用方式也发生根本性变革。据CNNIC发布的第39次《中国互联网络发展状况统计报告》显示^[1], 截至2016年年底, 我国网民规模已达7.3亿人, 相当于欧洲人口总量, 互联网普及率达到53.2%, 以“.cn”结尾的域名总数达到2 061万, 同比增长25.9个百分点。网络信息资源已深入渗透至我国的社会、经济、文化领域, 成为记录人类活动和文明的重要载体和战略资产, 网络信息的及时保存和有效利用对推动我国社会和文化发展、增强国家综合实力具有重要意义。

国际社会高度重视网络信息资源的保存与利用, 西方发达国家或组织密集出台多项专门政策给予支持。我国在2016年发布的《促进大数据发展行动纲要》提出, 要加强互联网信息采集利用, 加强顶层设计, 加强互联网信息采集、保存和分析能力建设, 制定完善互联网信息保存相关法律法规, 构建互联网信息保存和信息服务体系^[2]。从世界范围看, 网络信息的保存保护一直是国内外图书情报界的研究与工作重点, 且进行网络信息采集的机构一般为国家级图书馆。目

前, 国际上已完成或正在推进的网络信息保存项目大约有100个, 从影响力和成果看, 较重要的主要有美国Internet Archive的Archive-It项目^[3]、美国国会图书馆MINERVA项目、澳大利亚国家图书馆Pandora项目、大英图书馆和英国国家档案馆等机构发起的UKWAC项目、法国国家图书馆BnF项目、芬兰国家图书馆Finnish Web Archive项目、瑞典国家图书馆Kulturarw项目等^[4]。以这些项目为代表的各国公共文化机构对国际上或所在地区有重要战略价值的网络资源进行采集, 形成各具特色的网络信息保存体系, 为传承人类文明、促进文化发展提供重要保障。

中国国家图书馆也十分重视网络信息资源的采集与保存。2003年开展“网络信息采集与保存试验项目”^[5], 2007年加入国际互联网保存联盟, 并于2009年成立国家图书馆互联网信息保存保护中心, 通过国内外的广泛合作, 致力于网络信息资源的长期保存和利用服务。经过十多年积累和发展, 已形成涵盖国内外政治、经济、社会、文化、科技等领域重要网站和重大专题网络资源的特色化网络资源建设体系, 数据量达到114.73TB, 成为我国数字图书馆资源建设的重要组成部分, 为政府决策、科学研究和满足民众信息需求提供重要支持。在国家图书馆的示范和带动下, 各地方图书馆也开始根据自身需求开展具有地方文化特色的网络

^{*} 本研究得到国家社会科学基金项目“网络信息采集与保存策略研究”(编号: 15BTQ019)资助。

资源建设,网络信息保存的规模化效益逐渐显现。通过积极实践与探索,国家数字图书馆已初步形成较科学客观的网络资源评估和面向全国的长期保存与有效服务机制,对在互联网环境下,提高国家数字图书馆网络资源建设的规范性和质量水平,促进网络信息综合应用与服务能力提升具有重要意义。本文旨在通过对国家数字图书馆在网络资源遴选、平台技术架构和资源服务等方面进行总结和分析,为进一步改进网络资源的保存与服务工作,不断满足社会各界的文化信息需求提供参考和借鉴。

2 网络资源遴选

2.1 网络资源保存范围确定

信息资源指在人类社会活动中积累的以信息为核心要素的资源集合。网络信息资源区别于传统信息资源,是指以电子数据形式存储,借助计算机通过互联网进行传播的各类信息资源^[6]。鉴于网络信息的海量性和复杂性,确定合适的网络资源保存范围是决定资源采集效益和内容质量的重要因素。结合国内外网络信息资源保存工作实践,在开展规模性采集过程中,既要做到网络资源的重点采集,也要防范垃圾信息的堆积。在保证信息采集全面性的同时,也应注重资源内容的权威性和保存价值。围绕我国公共图书馆的职能定位和服务需求,国家数字图书馆通过全面采集和重点采集的方式实现对不同领域和范围的重要网络资源的有效采集。同时,在采集过程中充分考虑采集内容的延续性和采集后数据的有效利用,不断修正网络资源采集范围。目前,国家数字图书馆全面保存的网络资源主要包括:党中央各部门及直属事业单位网站;全国人大、政协,及其他与国家立法、决策相关机构网站;最高人民法院、最高人民检察院网站;国务院各组成机构及其直属单位和行业协会网站;各民主党派中央机关网站;中央级社会团体网站等。重点保存的网络资源包括:我国经济领域重要机构、组织、研究单位网站;我国重要的科研教育机构网站;我国历史文化、传统艺术、地理人文、科技科普、国学研究类网站和专题;国家一级博物馆、省级及以上主要美术馆、艺术馆、展览馆等公益文化类机构,以及重要图情组织机构网站;我国重要历史人物和事件的纪念馆及专题网站;我国社会保障、劳动就业、人才法律、人口健康等领域重要机构和网站;国土资源

监测与利用、环境气象、生态保护、防灾减灾等领域的重要机构、研究单位和专题网站;国家重大项目和工程建设网站;国际重要组织机构、国外主要国家和周边国家的官方网站,及其公开发布的重要文件和资料等。

2.2 网络资源评估和遴选

与传统文献资源相比,网络信息资源通常呈现出质量的不均衡性与内容的不确定性,这将严重地制约图书馆数字资源服务向深度应用发展。所以,网络信息资源保存与应用服务的第一步需要开展网络资源评估,确定网络资源质量,选择优质的网络资源进行采集和保存。

2.2.1 网络资源评估

国家数字图书馆在总结业界相关研究成果的基础上,利用定性定量相结合的方法构建较科学合理的网络信息评价指标和评估流程。通过对国际上各种数字资源评价方案的深入调研和分析,国家数字图书馆采用德尔菲法对待采集的网络资源进行分析和排名^[7],评估流程如图1所示。

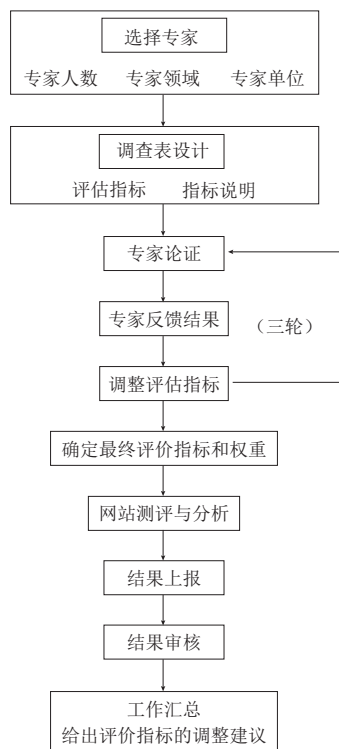


图1 网络资源评估流程

2.2.2 网络资源评估指标确定

基于以上流程,国家数字图书馆采用定量指标和定性指标相结合的综合评估方法对网络资源进行评估。定量指标属于客观性指标,是通过软件工具或相关评测网站能够对网络资源给出具体评测值的指标,评测过程不需人工干预;定性指标属于主观性指标,是评分者根据评估说明并结合资源具体情况给出评测值的评价指标,主要面向网站的内容表现形式与特征。初步确定的网络资源评估指标,如表1所示。

表1 网络资源评估的定性指标和定量指标

指标	一级指标	二级指标
定性指标	信息内容评价指标	准确性
		权威性
		广度和深度
		目的性
		客观性
		原创性
		创新性
		时效性
	编排与设计指标	结构设计
		外观设计
		多媒体效果
		兼容性
	易用性评价指标	互动性
		导航设计
		界面友好设计
		浏览速度和质量
		可检索性
		稳定性
定量指标	影响力指标	网站排名
		引用次数
	访问指标	被访问次数
		用户访问深度
	质量指标	死链接检测
		网站更新频率
		链接数量

2.2.3 网络资源评估流程确定

在初步确定评估指标的基础上,经过多次专家组

论证,确定最终的评估指标及其权重。

如图2所示,评估指标及其权重的确定共经过三轮专家论证。第一轮调查表的内容是向专家征询意见,以百分比形式展现专家所选各项指标的比例;第二轮调查表由第一轮的指标和专家新增指标构成;第三轮调查表展现各项指标的专家所选比例,依据各项指标最后得分进行权重分配。在最终确定的网络信息评价指标体系基础上,由工作人员对国内外符合采集范围的网站二级指标进行打分,并对所有网站按照加权分数进行降序排列,得到最终的网站评估排名,审核后的网站列表即作为最终采集任务列表。

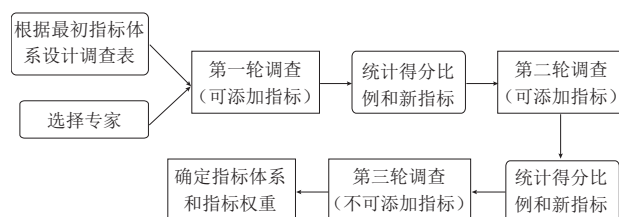


图2 基于多轮专家论证的评估指标确定流程图

3 网络信息资源保存平台的开发与建设

3.1 云管理系统平台技术架构

随着网络信息保存业务的发展,网络信息总量快速增长,传统的软硬件基础架构已不适合海量网络资源的存储与管理。为实现网络信息长期保存的可持续发展,国家数字图书馆加强设施改造和技术创新,建设基于分布式云存储管理平台以实现网络信息的有效保存。

国家数字图书馆云管理系统平台包括采集层、存储层、管理层、应用层四级架构。在采集层,采用分布式采集结构,通过部署多个省级采集节点和机构节点进行具体的采集工作,以完整采集和重点采集相结合的方式由Heritrix软件完成对网络资源的采集和保存;在存储层,通过分布式文件系统实现资源的分布式存储和集中式管理;在管理层,通过云管理平台对采集层多个采集节点进行统一调度与配置,实现工作的高效运行;在应用层,搭建访问服务平台、嵌入第三方信息查询和检索接口,为用户提供资源访问和获取服务,并使用数据分析工具对存储资源进行分析与挖掘。

国家数字图书馆通过硬件基础设施共享来为海量

网络信息管理提供最基本的物理资源(包括计算、存储、数据和网络设备),通过虚拟化技术和集群技术把内存、I/O设备、存储和计算能力汇集起来成为一个虚拟的资源池,以虚拟化环境提供支撑平台硬件系统统一的管理能力。硬件系统处于平台技术结构的最底层,包括硬件设备和虚拟化。硬件设备由现有各类数字资源管理服务器、存储设备、网络设备等构成;虚拟化则运用虚拟化技术对物理硬件层进行管理,向上提供计算、数据存储和网络通信等虚拟资源,其主要目的是将异构的底层物理资源整合成相同类型的资源池(如计算资源池、存储资源池等),以便创造协同、统一的工作基础。

国家数字图书馆网络信息保存平台硬件系统建设方面主要包括计算资源(服务器等)建设和存储建设。计算资源用于运行支撑平台基础软件系统、应用服务集成系统及未来扩展系统等。支撑平台以共享的“云存储池”为基础,通过基础软件系统中的云管理系统实现对全国硬件系统资源的统一管理和调度。

国家数字图书馆网络信息保存平台软件架构主要由平台基础软件系统(即云管理系统)组成,分别在国家数字图书馆的不同层级和机构进行部署,形成一一对接。云管理系统实现对不同层级图书馆硬件系统的综合使用管理并形成统一视图,实现对全国图书馆存档网络信息数据服务的基础支撑作用(见图3)。

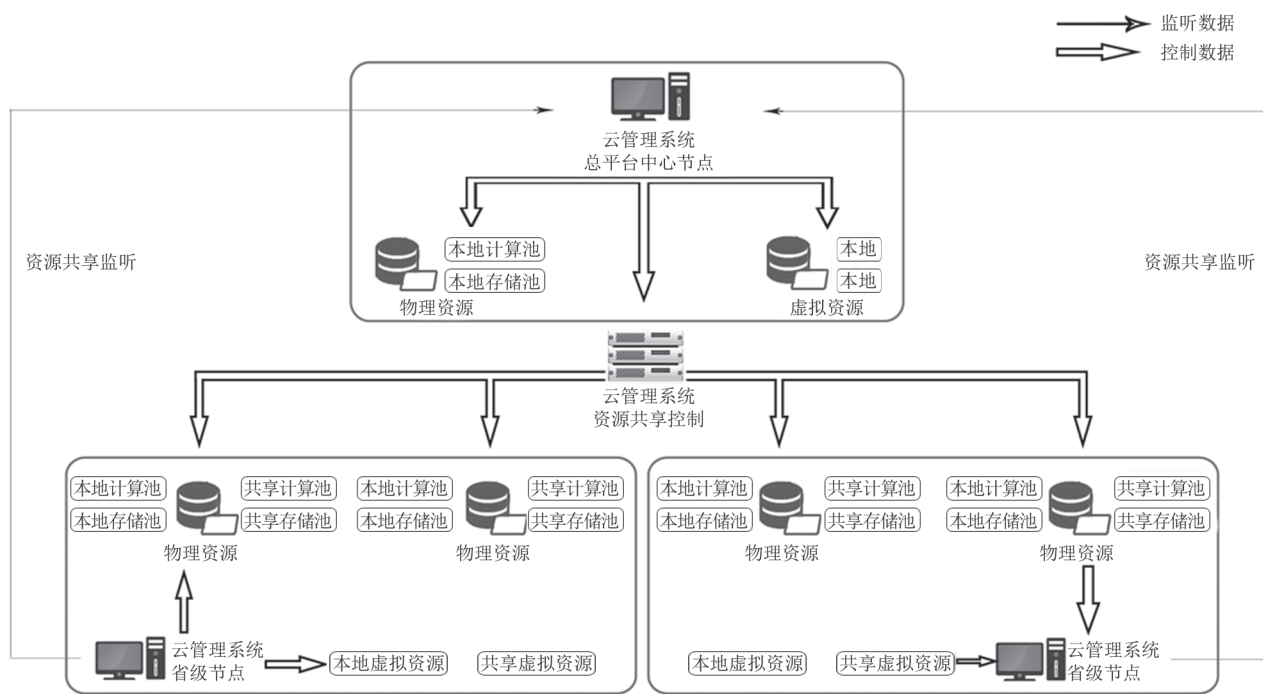


图3 国家数字图书馆网络信息保存平台架构

3.2 分级式资源共享与集成管理

资源的共享和集成是实现国家数字图书馆网络信息采集规模化、管理规范化的重要环节。通过对全国参与网络信息保存的各机构应用系统的“容器”作用,不断“共享”存档资源和应用软件,形成具有一定规模的“资源集成管理中心”,从而推动以网络信息应用系统为服务“窗口”的全国数字图书馆网络信息服务新业态。国家数字图书馆网络信息集成管理

系统架构如图4所示。

统一认证鉴权系统可实现对联建机构系统管理员和公众用户实现统一认证,系统管理员只需登录一次,即可访问接入统一认证的不同机构业务系统和各地存档资源;机构权限管理可实现对各联建机构存档网络信息资源的有效共享和管理。公众用户、各级文化共享机构需要在平台注册登记用户和机构的信息,系统管理员可授权用户与机构不同的访问或管理权限;合作单位在系统提交资源发布申请后,经审核通过即可实

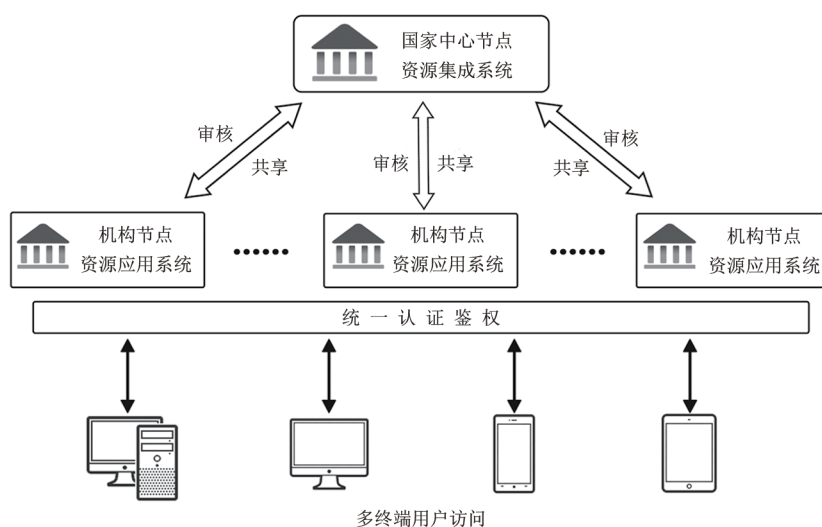


图4 国家数字图书馆网络信息集成管理系统构架

现所有申请单位的资源集成共享；开发者可使用应用构建工具按照应用规范开发出各类应用组件，并发布到内容展现门户或应用门户，供公众或其他机构和加工者使用。

同时，为提高国家数字图书馆的网络信息资源共享能力，系统提供接入国家各省市县及其他文化机构的应用系统能力，包括为各种应用系统提供受控的运行环境（由云平台基础资源提供）、应用监督和管理、应用自动化部署和监控，系统门户整合、业务处理整合、数据交换和数据整合等。

4 网络资源应用与服务

满足社会各界的信息需求是国家数字图书馆服务的根本目标。互联网环境下，国家数字图书馆充分利用网络信息保存成果，以全面采集、组织和分析后的网络数据信息应用手段与现代信息技术针对不同需求提供高水平、多层次的服务，为促进我国文化事业的发展以及国家软实力的提高提供支持。目前，国家数字图书馆开展网络资源服务的方式主要有以下五种。

4.1 制定元数据规范，加强资源整合与揭示能力

元数据是描述资源关键信息的一种结构化数据。制定规范化的元数据标准能实现对数字资源的查找揭示和有效管理，对用户发现资源、利用资源具有重要

作用。与传统数字资源相比，网络信息资源无论是在结构、分布、种类，还是在传播范围、载体形态、对象格式等方面都显示出新的特点。为对网络资源进行客观深度描述，实现对资源内容最大限度的揭示，同时便于网络资源的组织整合与检索，国家数字图书馆在研究现有数字资源元数据规范的基础上，根据实际需求建立了一套较完善的网络资源元数据著录规范，实现对国内外网站资源、网络专题资源等的规范化著录。同时，为促进网络资源联建与共享，还制定“数字图书馆推广工程”网络资源元数据规范，并成为各地方馆资源建设的参考标准。基于网络资源元数据著录规范，国家数字图书馆对国家图书馆及各地方馆存档的网络资源进行有序整合与统一发布服务，实现资源的一站式检索，提高资源发现与服务效益。

4.2 建设专题资源库，开展多样化、特色化网络资源服务

国家数字图书馆十分重视优质网络文化资源的建设与服务，每年都会围绕国家重要领域和热点议题，遴选并采集相当数量的网络资源，并通过门户网站发布。目前提供服务的网络专题资源涵盖治国理政、文化与民俗、科学技术、环境保护等14个分类，专题数量超过175个，每个专题汇集约200个从国内各大门户网站精选的优质网络资源。此外，国家数字图书馆紧密结合国家重大发展战略，重点建设“一带一路”和“大众创业，万众创新”资源库，通过多种方式和手段对相关领导讲

话、政策文件、重要事件、新闻报道、地方践行、背景知识等内容作出集中整合与展示,为相关科学研究和信息获取提供支持。此外,在“数字图书馆推广工程”的支持下,各地方图书馆积极开展具有当地文化特色的网络专题资源库建设与服务,进一步加强国家数字图书馆在公共互联网文化领域发挥的重要作用和影响力。

4.3 应用可视化技术手段,提升资源展示与服务效果

国家数字图书馆一直重视现代信息技术在文化资源建设和服务中的重要作用,通过技术创新构建基于互联网、移动互联网、广播电视网等渠道的文化服务网络,不断变革网络信息资源的组织和服务方式,较大程度提升了网络信息服务效果,取得良好社会效益。国家数字图书馆建有专门网络资源发布与服务网站,通过资源推荐、快照保存、信息检索、资源分类浏览等实现资源的有序组织与呈现、信息发现与检索,通过知识库向用户介绍网络信息保存相关知识和内容,并提供代存档服务。“一带一路”和“大众创业,万众创新”重大专题资源库中大量使用现代信息技术,通过时间轴、地域轴、地图、知识图谱等可视化手段进行资源展示与用户交互,以热门关键词实现资源推荐检索。可视化技术的应用改善资源组织与呈现方式,极大提升用户体验,较大程度提高网络资源服务效果,取得良好社会效益。

4.4 开展定制化资源服务,满足用户个性化内容需求

为用户提供准确的信息资源,使其享受到个性化的服务体验是信息时代国家数字图书馆追求的主要目标。当前,我国数字图书馆服务人数急剧增多,年龄、学历、地域、民族情况复杂,若文化服务缺乏针对性,文化内容和服务手段千篇一律,不但会造成资源浪费,也容易降低数字图书馆的用户黏性,从而降低在文化信息传播和服务领域的影响力。鉴于此,国家数字图书馆主动寻求社会需求,积极与社会机构开展合作。特别是在与全国人大图书馆的合作中,国家数字图书馆针对其业务与服务需求,对其网络资源建设与发布进行定制化开发,建设“人民代表大会网站典藏”资源库,旨在通过对各地级以上人大官方网站资源进行采集、保存与整合,提供在线资源展示与信息检索,从而促进全国人

大网站资源长期保存和服务体系构建,为政府立法决策和服务提供支持。该资源库是国家数字图书馆联合政府机构开展资源和服务合作的重要示范项目,对于掌握我国人大组织和工作情况、研究我国人大事业发展历程,为全国人大事业提供资源保障有重要意义。

4.5 依托国家重点文化工程,实现分布式合作模式

目前,国家图书馆作为社会记忆的保存机构,担负着网络信息资源构建的重要职责,依托国家“数字图书馆推广工程”,联合各地市级图书馆,共同构建面向全国的分布式保存机制。各机构合作选择需要保存的网络信息资源,统一实施网络资源评估标准,交流并分享网络信息资源保存的采集策略,形成规模化保存行动,各展所长,实现经验和技术层面的融合。

国家图书馆与各省级、市级及乡镇级图书馆间存在纵向关联关系,通过建立共同联盟机构可实现联合评估、分工采集、联合编目、建立统一管理平台和联机检索。近年来,依托“数字图书馆推广工程”,省级公共图书馆开展了网络信息资源联建工作,随着业务的推进和不断深入,各地图书馆存档的网络信息种类和数据量预计会有显著增长。

国家图书馆与其他数字文化保存机构存在平行关联关系,以国家图书馆为示范,辐射其他数字文化保存机构,形成分布式网络信息采集与保存机制;共同协商,划分各机构保存范围;统一开发和共享保存技术,由国家图书馆主导开发采集软件与网页回放工具,建立统一管理平台;使用统一数据存储格式和元数据保存标准,节省开发成本,共享技术成果。同时,国家图书馆与各机构联合开展人员培训,交流采集、编目、保存经验,与图书馆外机构合作,打破机构壁垒,与更多国家网络信息保存项目合作,借鉴先进经验、技术与标准。

5 结语

国家数字图书馆以保存人类文明为职责,是国内重要的文化知识传播中心。加快推进网络信息资源采集、保存和服务利用,对承载中华数字记忆、弘扬优秀传统文化、促进人类信息发展具有重要战略意义。国家数字图书馆将继续深入推进网络信息的资源联建和共享,构

建全国统一的网络信息保存与服务平台,加强技术研发与创新,通过多种手段和方式面向全国不同用户群体开展知识化、个性化、精准化、一站式信息服务,为扩大国家数字图书馆服务范围和提高服务能力提供支持。

参考文献

- [1] 中国互联网络信息中心.第39次《中国互联网络发展状况统计报告》[EB/OL].(2017-01-22)[2017-05-22].http://www.cnnic.net.cn/hlwfzyj/hlwzxbg/hlwjbg/201701/t20170122_66437.htm.
- [2] 国务院关于印发促进大数据发展行动纲要的通知[EB/OL].(2015-09-05)[2017-05-22].http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.
- [3] 王烁.美国网页归档项目——Internet Archive发展研究[J].兰台世界,2012(6):18-19.
- [4] HAKALA J. Archiving the web:European experiences[J].Program,2004,38(3):176-183.
- [5] 赵丽琴.我国网络信息保存研究述评[J].图书馆学研究,2011(2):5-7.
- [6] 马费成,赖茂生.信息资源管理[M].北京:高等教育出版社,2006.
- [7] KAVOURA A,ANDERSSON T.Applying delphi method for strategic design of social entrepreneurship[J].Library Review,2016,65(3):185-205.

作者简介

张炜,女,1973年生,硕士,研究馆员,研究方向:数字图书馆资源建设与服务,E-mail:zhw@nlc.cn。
敦文杰,男,1982年生,博士,副研究馆员,研究方向:数字资源整合与服务,E-mail:dunwj@nlc.cn。
周笑盈,女,1989年生,硕士,助理馆员,研究方向:数字图书馆新媒体服务,E-mail:zhouxy@nlc.cn。

Network Resource Preservation Practice and Exploration of National Digital Library

ZHANG Wei, DUN WenJie, ZHOU XiaoYing
(National Library of China, Beijing 100081, China)

Abstract: In the internet environment, network resource has become an important carrier of human's civilization. As the important component of our country's public culture service system, the National Digital Library paid attention to the network resource preservation and service. Since developed the web information collection and preservation program in 2003, the National Digital Library has made a big progress on selecting network resource, building manage platform and innovating service, and accumulated a lot of practice experience. This paper summarizes and analyses the work of National Digital Library's network resource preservation, and it is hoped for providing reference resources for future work and relevant practitioner.

Keywords: Network Resource Preservation; National Digital Library; Internet; Culture Service

(收稿日期: 2017-05-24)