国家图书馆网络资源采集与保存平台的技术实现

赵丹阳
（国家图书馆，北京 100081）

摘要：国家图书馆网络信息的保存与服务工作开展多年，积累了丰富的实践经验，并开发了网络资源采集与保存平台，联合全国图书馆共同开展网络保存业务。本文在对国家图书馆网络资源保存体系、网络资源采集与保存平台分析的基础上，详细介绍平台的构建思路、技术路线和关键技术，以期为业界提供有益的参考和借鉴。

关键词：国家图书馆；分布式；网络资源采集与保存平台；Web存档
中图分类号：G255　　DOI：10.3772/j.issn.1673-2286.2020.09.006
引文格式：赵丹阳. 国家图书馆网络资源采集与保存平台的技术实现[J]. 数字图书馆论坛，2020 (9)：41-47.

1 国家图书馆网络资源保存目标

网络信息时代，数字信息逐渐成为人类文明记忆的载体。中国互联网络信息中心发布第45次《中国互联网络发展状况统计报告》[1]显示，截至2019年12月，我国网站数量为497万个，网页数量2 978亿个。然而，网络信息更新快、易流失，要更好地传承数字时代的文化、文明，及时、完整地保存网络信息资源非常重要，这也是国家图书馆履行文献保存与保护的职责所在。

1.1 保存项目开展历程

作为国家公共文化服务体系的重要组成部分，国家图书馆一直重视网络信息资源的保存与服务。最早于2003年开展的网络信息资源采集与保存试验项目（Web Information Collection and Preservation，WICP）[2]，开始实验性地对中国境内的互联网资源进行采集与保存。2007年，国家图书馆正式加入并成为国际互联网保存联盟（International Internet Preservation Consortium，IIPC）成员单位。在联盟框架下，国家图书馆广泛地进行交流与合作，并基于国际通用的标准和技术体系，开展了国内网络资源保存，进而促进了该项工作的国际化和标准化进程。2009年成立了国家图书馆互联网信息保存保护中心，致力于中国互联网信息资源长期保存和保护；2012年开通网站服务，对采集到的互联网资源进行组织与展示；2014年依托“网事典藏”项目，联合全国公共图书馆共同开展互联网资源的保存和服务；2018年研发并推广部署网络资源采集与保存平台，实现互联网资源高效和规范化地采集、编目、回放、发布和服务；截至2018年底，全国各级公共图书馆累计采集网站2.3万余个，实现了涵盖政府公开信息、国内外重要网站网页等互联网资源的保存与保护。

1.2 网络资源采集和保存策略

国家图书馆网络资源采集模式分为全域采集[3]和专题性采集两种。

全域采集也是整站采集，是以域名为单位对网站全部对象数据进行采集与保存。在采集网站与网络信息的选取方面，人工介入与定量、定性的考量成为必不可少的要素。国家图书馆的网络资源保存工作以保存中华数字记忆为主要职责，遴选具有重要保存价值的优质网络文化资源进行采集。此外，在采集过程中充分考虑采集的延续性和采集后数据的有效利用。其中，政府

网站是采集的主体，特别是在《政府信息公开条例》和数字图书馆推广工程的推动和支持下，国家图书馆联合全国各级公共图书馆，对我国党政机关网站进行全面采集，政府级别涵盖中央级、省部级、地市级、县级、乡镇及以下等，其中国家图书馆负责中央级网站采集，各级公共图书馆负责采集各自区域及相对应级别的政府网站。多年来，采集的资源为民众查阅政府信息和学者研究提供了良好的资源保障，同时也记录了中国政府机构改革历程，为国家决策提供参考。

专题性采集是指围绕国家重要领域和社会热点，对相关主题的网络资源进行采集和保存，并提供给社会大众更有深度、更有挖掘价值的整合性信息。采集的资源包括各网站的专题栏目、相关网页及其他类型的网络资源。网络专题资源建设可以说是数字图书馆在新环境下提供资源服务的重要方式，不仅在于数字遗产的保存与传承，还可以为社会提供数据整合后的知识服务。国家图书馆目前每年会建设60个左右的专题资源库，并构建若干重点专题，通过资源聚类、可视化、关联分析等技术创新服务模式，提升服务体验效果。

截至2018年底，国家图书馆已采集政府网站超过2万个/次，国内网站2 319个/次，国外网站5 583个/次，专题超过276个，数据量总计超过210TB。网络资源已成为图书馆数字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。

2 国家图书馆网络资源保存体系

为了应对海量网络资源的采集与保存挑战，国家图书馆已经初步形成较为完整的网络资源保存体系和有效服务机制，通过标准规范、技术应用、平台建设、数据挖掘和分析，以及用户服务体验、分工合作等方式，实现网络资源保存工作的长期、科学、可持续发展。

2.1 制定网络采集元数据规范，加强资源整合与揭示力度

网络资源在分类、结构以及表现形式、数据特点等方面与传统的数字资源相比，有很多独特的性质。当完成大规模的资源采集和归档后，要用客观性、完整性、统一性的规范化语言进行描述，进而才能为进一步的长期保存、有效揭示以及深度的数据挖掘奠定基础。国家图书馆在研究现有数字资源元数据规范的基础上，结合网络资源的数据特点和传播特点，建立了一套较为完善的网络资源元数据著录规范，实现对国内外网站资源、网络专题资源等规范化著录。同时，为促进网络资源联建与共享，制定了“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。

中国政府公开信息整合服务平台，是国家图书馆联合全国各级公共图书馆共同建设的。目前已经230余家公共图书馆参与到项目建设之中，所有参与馆采用统一的元数据标准和规范，各自采集整合本行政区的政府公开信息，通过“分层建设、共建共享”模式，将采集到的网络资源有序整合与统一发布服务，实现公共图书馆对政府公开信息的收集、整理、保存并服务于公众。

2.2 基于数字图书馆推广工程，联合全国公共图书馆开展保存业务

作为国家公共文化服务体系的重要组成部分，国家图书馆重视并积极推动中国网络资源保存与服务的知识普及、能力提升、业务推广。从2014年开始，依托国家“数字图书馆推广工程”，国家图书馆联合全国各地市级图书馆，基于“统一调度、分工采集、集中索引、分散保存”模式，逐步构建了覆盖全国的分级分布的网络信息保存体系。

在网络资源保存体系中，创立了“网事典藏”项目，广泛号召各个图书馆参与到该项工作中。在“网事典藏”项目中，参与图书馆采集并保存反映当地政治、经济、文化发展的重要网站和热点专题资源。其中网站采集主要以当地人民政府及下属机构网站为主，整站采集；专题资源采集围绕当地重大文化事件、地方民风民俗、地方文化保护等主题，采集本地区完整的、延续性较好的重点专题的网络资源，采集对象为网站专题频道和网页。参与该项目的图书馆逐年递增，2014年首都图书馆、湖北省图书馆、浙江图书馆、吉林图书馆以及新疆建设兵团图书馆5家省级公共图书馆成为首批联建成员。2015年有78家省市级图书馆积极申报网络信息保存工作，此后逐年递增，到2018年申报的图书馆数量已经达到115家。

2.3 建设分布式云存储管理平台，实现多机构业务协同发展

国家图书馆的网络资源保存业务从2005年开始基

于开源软件Heritrix[4-5]进行采集、编目和保存。基于网络资源的增多、业务效率的提升以及存储空间的有效利用等因素，国家图书馆在网络资源采集与保存的工作中，一直跟踪业界技术发展，尤其关注存储技术、分布式结构以及云服务架构，在业务实践中通过不断自行开发软件程序实现网络资源采集与保存平台的功能改进。随着全国范围内多个图书馆参与到网络资源保存工作中，更加需要一个规范性、开放性、共享性的软件平台，用以适应不同基础硬件环境的图书馆网络采集业务需求，并可以支撑多个图书馆基于同一软件平台共同开展网络资源采集和保存工作，共同促进中国的网络资源保存事业的发展。

国家图书馆基于近十年的软件开发基础和平台功能改善积累，在2018年研发出一套分布式云存储架构的网络资源采集与保存平台。该平台基于开源软件Heritrix进行定制开发，通过模块化和流程化，实现网络资源采集和保存全流程业务的规范化管理和可视化操作，提升了工作效率。平台采集的资源以国际通用标准WARC[6]格式存储在分布式文件系统中，经质检合格后的数据即可作为发布级数据用于网站前端展示。同时，质检合格的所有WARC数据经过封装后提交国家图书馆长期保存系统进行长期保存。平台的数据处理能力能支持至少100万条元数据数据的管理和使用。根据实际业务需要，平台共设计有8个功能区，即采集、编目、回放、内容管理与发布、数据保存、统计、用户管理、维护备份，每个功能区采用模块化设计，便于进一步的扩展和改进，也可实现模块化轻量级部署和服务。

网络资源采集与保存平台采用了分布式云架构，可以实现国家图书馆与多个图书馆（机构）共同开展网络采集业务。网络资源采集与保存平台的中心节点部署在国家图书馆，由国家图书馆进行统一操作和管理。其他图书馆作为并行节点，可以按照自身的业务需求，以定制化的方式安装业务需要的功能模块。在云架构下的每个图书馆，可以依靠本地所有的服务器设备和网络进行资源采集，并实现本地存储；也可以统一使用平台中心节点所包含的服务器设备和网络进行资源采集，并实现中心节点统一存储。云架构下的所有采集节点采集到的网络资源元数据集中存储于中心节点上，进而实现多机构网络采集数据的集中管理和使用。

网络资源采集与保存平台有较为完善的用户和权限控制机制，既做到分工明确，又可以确保平台上各个图书馆自有的资源和操作不受干扰。位于国家图书馆的平台中心，统一管理全国各个节点的网络信息保存任务，同时也是所有图书馆节点的信息聚合点，平台中心可实时了解各个时间段、各个节点网络信息保存任务的进行情况，对信息保存相关业务进行调度与管理。

3　国家图书馆网络资源采集与保存平台构建思路

国家图书馆根据自有的网络资源采集和保存的业务特点及业务管理需求，同时考虑到面向全国多机构共同开展业务，设计了网络资源采集与保存平台的系统架构及功能，使其具有自动化、分布式、云管理的个性化特点。

3.1　业务流程管理要自动化、流程化和模块化

以往在IIPC框架下实现的网络资源采集和保存的整个业务流程，类似于种子链接的部署、采集结果的汇总、索引文件的建立以及发布链接的质检等操作，均需要业务人员手工进行操作和干预，并且这类工作经常需要重复操作。随着业务的持续发展，在各个业务环节中产生及需要处理的数据量大幅增长，手工操作已经无法满足业务发展的需求。此外，由于全国不同图书馆的业务人员的计算机知识水平不等，在进行网络资源采集和保存的工作中存在很大的困难，严重阻碍各地图书馆的网络资源采集和保存业务的推进和发展。

新构建的平台要以用户为中心，降低业务操作难度，规范业务流程，提高业务工作效率；此外，还要通过模块化的形式，使得网络资源采集和保存的完整流程切分成多个合理的、彼此有关联、个体相对独立的业务模块。平台具有可视化的操作界面，让不具备深入的网络资源采集或计算机知识的业务人员也可以操作和完成工作，进而让更多的图书馆加入到网络资源采集与保存的业务中，以共同实现网络资源的及时保存和有效服务。

3.2　采集硬件架构要高可用、高效率和可扩展

以往的网络资源采集与保存业务，使用多台服务器，每台服务器部署一个Heritrix实例，由这台服务器单独运行Heritrix的采集进程，完成指定采集任务。这种模式依靠服务器的处理性能以及运行的Heritrix实例

数量来决定采集的效率和采集的频度。当承担不同的采集任务时，服务器的工作负荷会有较大区别；采集网络资源需要的网络带宽负荷、存储空间负荷均会有较大区别。经过一段时间的监控，发现在实际网络资源采集业务中，会出现部分服务器内存、CPU等占用较满，网络带宽占用较多，而与此同时，还有一些服务器处于闲置状态。这种模式不能充分利用所有硬件资源，对网络带宽、存储空间利用均有或多或少的资源浪费。

新构建的平台要适应国家图书馆及全国多个图书馆共同进行网络资源采集的业务模式，因此要整合所有服务器的资源，根据采集业务的实际情况动态调整服务器配置和采集业务的任务，进而支持持续性、不间断、大数据量、多机构的网络资源采集和保存业务；同时，还要较大程度地发挥服务器集群的整体性能，以虚拟化和分布式的模式整体调配服务器硬件、网络及存储资源，让其为网络资源采集和保存发挥最大的效能。

3.3 存储模式要分布式、可共享和可扩展

在以往的网络资源采集业务流程中，以网络资源采集任务为保存目标，将其保存为warc格式的文件，并压缩处理成为gz格式进行最终的保存，所有采集到的资源保存于存域网的在线存储空间中。这种存储模式，不但限制了Heritrix软件对于网络资源在单次采集中的最大抓取数量，而且会产生存储空间的失衡。因为在Heritrix实例运行过程中，其所在服务器连接的存域网空间是独享给该服务器使用的，当Heritrix实例采集的网络资源容量接近或者超出其服务器所支配的存储空间容量时，Heritrix软件的采集进程会受到影响甚至中断。此外，由于单台服务器的存储模式是独享其连接的存域网空间，不支持服务器虚拟化以及Heritrix多线程的虚拟化采集模式，不能为虚拟化后的分布式集群提供存储服务。

新构建的平台所采用的存储模式要适应虚拟化和分布式的硬件架构，形成可以共享的云存储池为国家图书馆以及平台上的其他机构提供可靠的数据存储服务；此外，云存储池可以实现根据业务需求，扩充存储空间、调度存储分配策略等，进而实现对采集业务的有效支撑。

4 国家图书馆网络资源采集与保存平台技术路线

网络采集与保存平台需最大化地提高原采集方式的自动化程度，且需持续性地进行海量的网络资源的采集、保存和服务。因此，基于平台要建设的分布式硬件架构、集群化批量采集模式以及共享式的存储空间管理，平台采用IIPC的Heritrix（采集）、Wayback[7]（索引）和回放服务的基本采集流程框架（见图1），保障网络资源采集和保存的整体流程完整和规范；此外，在基础框架上进行了多个个性化功能的改造和研发，充分发挥了开源工具的优势。

为了有效、统一管理整个网络采集与保存平台，在基础框架之上研发了一个B/S模式的管理端，不但可以实现本地采集任务的分发配置、采集资源的自动调度等管理，还可以实现平台上不同机构的采集任务的监控和管理。在扩展采集服务器的同时，使用虚拟化技术，形成服务器集群，并且部署多个Heritrix采集节点，运用Heritrix多线程的采集特点，形成大规模的、分布式采集结构；优化管理平台程序，调用Heritrix接口，实现批量采集任务的部署。摒弃以往直连式、独享式存储空间管理模式，采用GlusterFS分布式文件系统，作为Heritrix采集结果的统一组织和存储容器。采用Openwayback和NutchWAX[8]组件，对分布式文件系统中保存的采集数据进行URL索引和全文索引，解决以往传统架构中需要人工参与索引操作的问题。

5 国家图书馆网络资源采集与保存平台关键技术

网络采集与保存平台整体架构为分层、分级的云架构（见图2），对于普通用户来说，可以通过互联网发布平台来获取本平台的资源服务；对于业务人员来说，可以通过管理平台来控制和管理网络资源的采集和保存全流程。互联网发布平台与管理平台之间通过标准接口来实现数据和需求的互通，彼此之间都是透明的。管理平台完整地管理接入到本平台上所有机构的网络资源采集与保存的全流程，基于分步式云存储架构实现了分层分级式资源共享和集成管理。

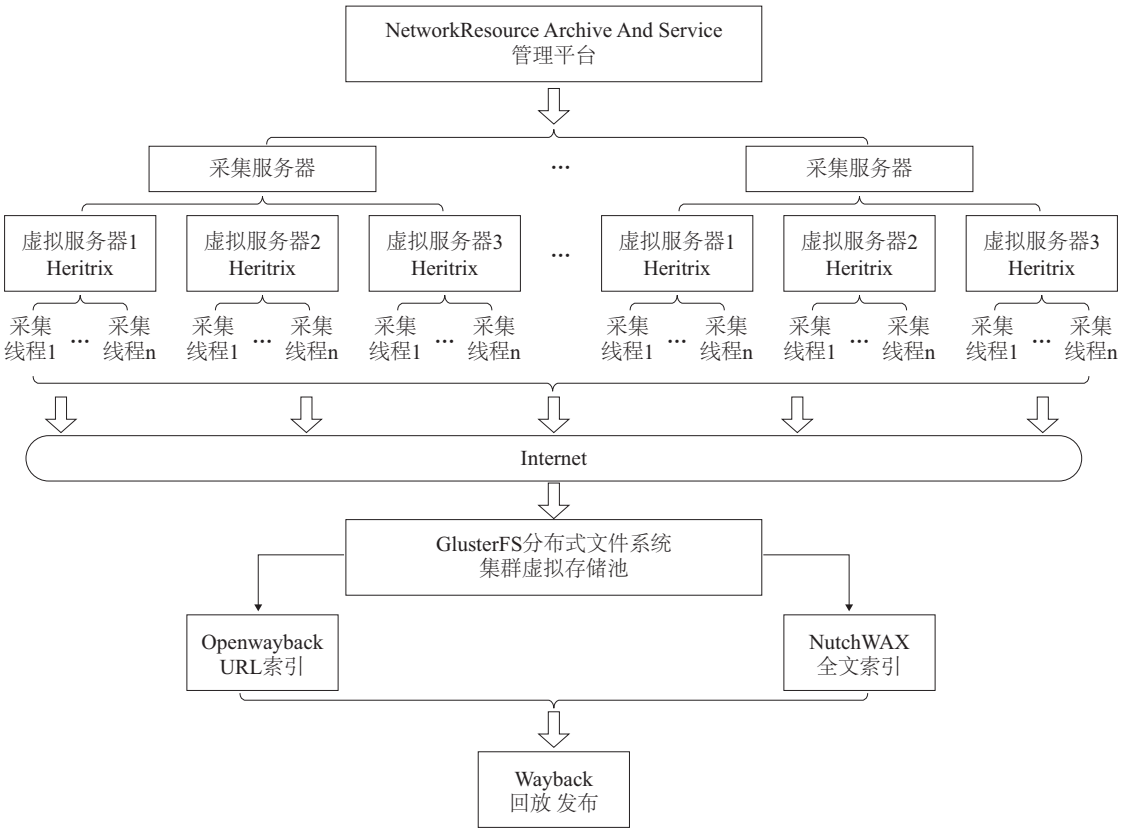

图1　基于IIPC工具的采集存档流程框架

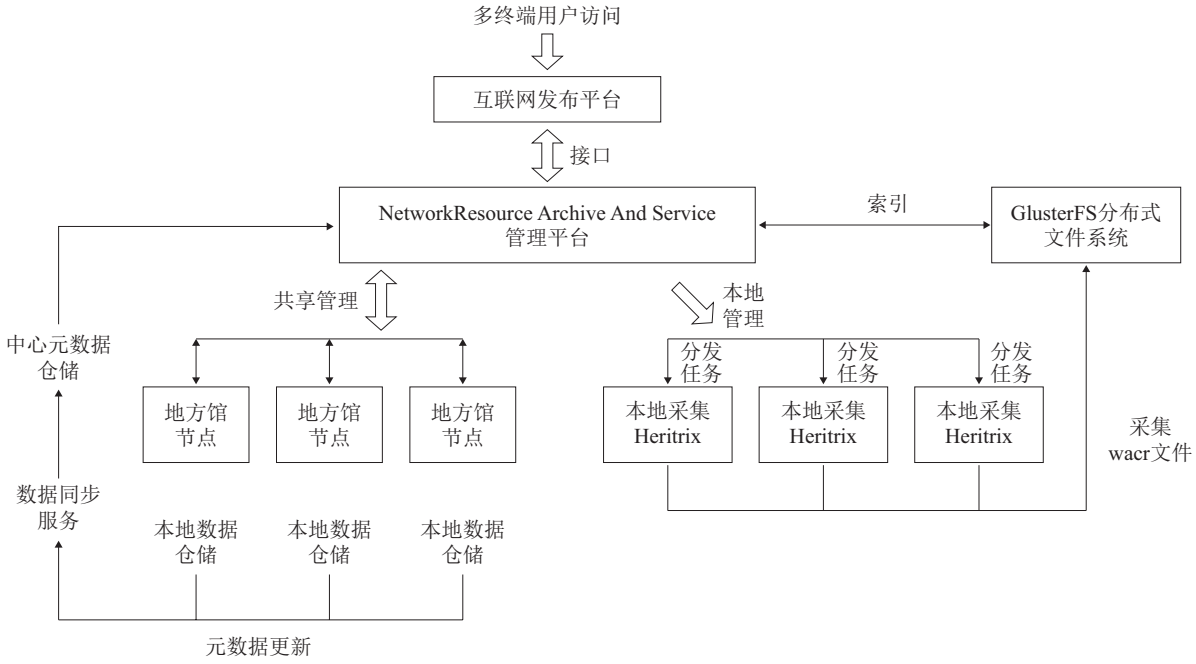

图2　网络资源采集与保存平台整体架构

5.1　分布式采集和存储的技术实现

（1）统一调度的实现。网络采集与保存平台的“管理平台”部署在一台主服务器上，采集节点部署在多台服务器上。管理平台与采集服务器节点之间建立通信反馈，对采集节点及其节点上的多个采集线程之间的采集

任务进行统一配置和调度，并负责所有部署完成的待采集任务组成的任务队列的调度管理，分配采集节点的空闲资源，调度待采集任务进行爬虫和抓取操作。

（2）分布式采集的实现。网络采集与保存平台采用分布式存储架构，可平行扩展增加多台物理服务器，每台物理服务器进行虚拟划分部署作为Heritrix采集节点，利用Heritrix多线程的特点，一个或者多个采集URL在多个线程中并行运行，进而提升采集的效率，充分利用采集服务器的性能。为了保障分布式采集的高效和统一，平台还要监控所有采集任务的执行情况，平衡每个采集节点网络及硬件资源效率；针对部署的任务队列采集和完成情况，调度Heritrix开展任务的抓取。

（3）分布式存储和服务的实现。为了实现分布式采集多个节点的采集数据统一归档、有效索引，平台采用高可用性、高性能、扩展性强以及对硬件性能要求低的GlusterFS分布式文件系统[9]实现数据的存储管理。

平台在GlusterFS分布式技术的应用上，摒弃了元数据服务器，消除整个系统的单点故障，提高对数据的并行存取速度。引用mysql作为数据库服务，存储平台相关的元数据，通过OMP在精度要求相同的情况下，加快算法的收敛速度。采用弹性HASH算法，数据分布存储在每个文件中，每个存储节点均有相同的目录结构，每个节点存储的数据HASH值为66635/节点数，第一个节点为0-65535/节点数，第二个节点为65535/节点数-2*65535/节点数，以此类推，这种散列分布算法解决了单点故障的问题。存储系统横向扩展、节点增加时，重新计算节点的HASH值，可以将所有数据平衡到各存储节点，实现节点的负载均衡随机可控，提高系统的吞吐量。

平台的采集节点通过InfiniBand和部署GlusterFS分布式文件系统的存储服务器进行通信和数据传输，抓取到的采集数据以warc格式进行存储。采集任务结束后，采集节点将warc格式数据转移至GlusterFS分布式文件系统中，运用其全局统一的命名空间管理[10]将数据资源统一放在GlusterFS集群虚拟存储池中，做统一的组织和存储。利用服务器磁盘中的多个Brick（GlusterFS的基本存储单元），形成存储卷后挂载到GlusterFS客户端，通过挂载点来共享所有的采集存档数据，并在平台前端显示存储的路径。部署的Wayback和NutchWAX组件会对存档数据进行URL索引和全文索引，并进行后续的发布服务。

5.2 分层分级式资源共享和集成管理的技术实现

网络采集与保存平台将国家图书馆设为平台的中心节点。中心节点作为管理者，统一管理国家图书馆以及连接到平台下的所有机构的网络采集任务，同时也是所有节点的信息聚合点，中心平台可实时了解各个时间段、各个节点网络信息采集任务的进展情况，对网络资源保存相关业务进行管理与调整。平台完整的硬件架构及软件系统均部署在国家图书馆，其他图书馆根据自身的业务需求可以模块化地部署自己需要的业务模块。在云架构的统一管理下，所有图书馆（机构）的硬件资源共同构成了网络资源采集与保存平台的硬件环境。在整体架构下的每个图书馆都可以利用自己本地的服务器、网络和存储空间完成网络资源的采集和保存，不同图书馆之间不会互相干扰。平台架构下的所有图书馆采集到的网络资源的元数据，全部同步和集中到中心节点的数据库中，进而实现元集中的管理模式和服务模式。

6 国家图书馆网络资源采集与保存平台运行效果

网络资源采集与保存平台已在国家图书馆和3家省级图书馆投入使用数月，服务平台访问首页见图3。截至2019年2月，已采集数据总量4.2TB（压缩），URL总数11 150个，整站采集总数为32个，平均采集速度23.75KB/s。目前平台运行效果良好，中心节点和机构节点用户均可通过登录浏览器客户端实现网络资源高效和规范化的采集、编目、回放、发布和服务、数据保存等工作流程，大幅提高了工作自动化程度，在采集任务的批量部署自动分发、采集节点任务调度等方面极大地节省了人力成本。同时，基于系统架构的可扩展性，平台中心节点已经实现由3台物理服务器扩展为7台，21台虚拟机的百余个采集节点，实现分布式采集的同时确保了平台的稳定高效运行，系统扩展性得到充分验证。

7 结语

国家图书馆互联网信息保存保护中心多年来一直致力于网络存档相关领域的研究，积极探索相关技术、政策、发展趋势等，自主研发的网络资源采集与保存平台的平稳运行，为建设国家范围、多机构、分级分步的

图3 服务平台访问首页展示

网络资源采集与保存奠定了基础。未来，国家图书馆还将不断地扩大网络存档规模、加强技术研发和创新、探索数据保存和分析的新模式，以期待满足业务和用户需求的不断改变。

参考文献

[1] 《中国互联网络发展状况统计报告》[EB/OL]. [2020-04-27]. http://www.cac.gov.cn/2020-04/27/c_1589535470378587.htm.

[2] 赵丽琴. 我国网络信息保存研究述评 [J]. 图书馆学研究，2011（4）：5-7.

[3] 孙倩，张炜. 我国图书馆开展网络信息保存的采选策略研究 [J]. 图书馆学研究，2016（17）：28-32.

[4] Heritrix [EB/OL]. [2020-08-05]. https://Webarchive.jira.com/wiki/display/Heritrix/Heritrix.

[5] Heritrix developer documentation [EB/OL]. [2020-08-05]. https://www.docin.com/p-496662679.html.

[6] ISO 28500：2009 Information and Documentation-WARC File Format [EB/OL]. [2020-08-05]. http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=44717.

[7] Web Archive Access Utilities [EB/OL]. [2020-08-05]. http://sourceforge.net/projects/archive-access/files/wayback/.

[8] NutchWAX [EB/OL]. [2020-08-05]. http://archive-access.sourceforge.net/projects/nutch/.

[9] DAVIES A，ORSARIA A. Scale out with GlusterFS [J]. Linux Journal，2013（235）：1.

[10] 王铃惠，李小勇，张轶彬. 海量小文件存储文件系统研究综述 [J]. 计算机应用与软件，2012，29（8）：106-109.

作者简介

赵丹阳，女，1988年生，硕士，馆员，研究方向：网络资源采集、长期保存、数字图书馆服务，E-mail：zhaody@nlc.cn。

Technical Realization on Web Archiving and Preservation Platform of National Library of China

ZHAO DanYang
(National Library of China, Beijing 100081, China)

Abstract: The collection and preservation of web archiving of the National Library of China has been done for many years, accumulated abundant practical experience and developed the web archiving and service platform, and united libraries nationwide to carry out web archiving and service jointly. This paper analyzes the construction ideas, technical routes and key technologies in detail based on the analysis the strategy of web information preservation and the requirements for the system platform.

Keywords: National Library of China; Distributed System; Web Archiving and Preservation Platform; Web Archive

（收稿日期：2020-07-27）