数字信息资源分布式协作保存网络构建研究

吴振新，付鸿鹄
（中国科学院文献情报中心，北京 100190）

摘要：本文基于对国家保存体系中分布式协作保存网络的需求分析，明确保存网络建设应遵循中心注册管理、独立节点管理、多种类型节点分类、松耦合、异构、网络架构灵活、可扩展、参与机构角色可转换和扩展的设计思路。介绍国家协作保存网络整体框架及节点功能，并对协作保存网络建设和运行的关键问题包括基于注册的管理机制、主动推送的数据提交模式、不同类型节点间的协作模式、唯一持久标识及系统的扩展性进行分析。

关键词：长期保存；数字信息资源；协作保存；保存网络
中图分类号：G250
DOI：10.3772/j.issn.1673-2286.2016.9.007

随着社会数字化的不断发展，数字内容复杂的类型及飞速扩展的体量，使保存机构面临更艰巨的保存任务与更为复杂的保存环境。依靠任何单一机构进行长期保存本身就是一种风险，而数字资源长期保存作为一种风险防范机制，需要通过合作保存，分摊责任和风险，从而提高长期保存本身的可信赖性。构建数字资源协作保存网络，协调和调度足够的社会资源，共同分担保存风险和责任，合作进行保存活动，避免资源重复保存及遗漏，已经成为各国保存机构的必然选择。数字资源长期保存已经是一个关系国家信息安全的战略问题。

对我国而言，通过建立国家数字科技文献资源长期保存体系，从国家层面整体实施长期保存战略，能有效解决个体机构实施长期保存普遍面临的经费、技术等难题，有利于全面形成可持续和可靠的长期保存服务体系，确保长期保存服务的共建共享。因此，构建高效可行的分布式协作保存网络，协调各机构的参与，合作开展数字资源保存，成为正在启动的国家数字资源长期保存体系示范系统建设项目最为紧迫的一项任务。

中国科学院文献情报中心在多年开展保存实践的基础上，结合国家保存体系示范系统的建设需要，深入开展协作保存网络的研究[1-2]，初步构建分布式协作保存网络，本文详细介绍该协作保存网络的设计思路以及关键问题的解决方案。

1 分布式协作保存网络的需求分析

“国家主导，联合参与，责任分担，协同保障”是国家保存体系的基本原则，国家保存体系需要吸纳全国相关领域机构积极参与，既要兼顾各方利益和责任，又要实现共建共享，因此，在进行分布式协作保存网络框架设计中，需要充分考虑六方面的需求。

（1）能够充分利用现有的工具和成果。中国科学院文献情报中心经过多年保存实践，已形成一个遵循OAIS标准，具有摄入、保存管理、公共服务和合作服务功能的可靠保存平台，并已稳定运行多年。在保存网络设计中，要考虑在该平台的基础上，以多个示范保存机构为结点，通过注册管理机制，实现分布式协作保存，构建分布式保存示范网络，为进一步发展多机构参与的分布式合作保存体系奠定基础。

（2）考虑参与机构所具有的不同职责、角色、能力。国家保存体系是由国家主导且长期稳定支持，兼具体采购和使用资源图书馆共同参与的一个面向全国的公共服务体系，需要在国家的统筹规划下，参与机构分工合作完成。除NSTL本身作为核心管理机构外，还要从参与国家保存体系的机构中，遴选一批符合条件的机构作为合规保存机构，分工合作负责数字科技文献资源长期保存。因此，要考虑多种角色定位，使参与机构

能根据自身特点，在国家保存体系中发挥不同的作用。

（3）具备一定包容性，支持个性化保存实现。能够允许各机构根据自身的特殊需求，灵活配置工作流程和数据存储管理策略，以满足在协作保存协议规定下的个性化保存实现。

（4）具有灵活、可扩展的体系结构。对新技术具有良好的适应能力和扩展能力，能够方便地集成其他软件和功能模块；具备弹性存储能力，能够满足协作保存网络规模不断扩大产生的存储空间持续增长的需求；能够快速部署新的存储系统，具备灵活扩展保存服务的能力。

（5）支持保存体系能够循序渐进地发展和扩展。允许成员机构方便地加入或者撤销，并能够支持灵活的角色转换。当保存机构不再具备合规保存机构资质时，能够方便地进行保存服务转移。

（6）具备协同工作的能力。允许多个机构跨地域、多层面的协作，能够实施任务分担计划；能够解决多个机构间数据同步问题；能够从整体上配置存储空间，计算资源。

2 协作保存网络设计思路

基于上述需求，结合前期对分布式长期保存网络的调研和分析，本文对如何构建国家保存体系分布式协作保存网络提出设计思路。

（1）中心注册管理机制。保存网络由多个节点组成，其中包括一个中心管理节点执行日常的管理和监控，多个保存节点独立运行，可以具备不同的保存功能，也可以互为备份和补充。

（2）独立节点管理机制。独立节点管理即节点自治。除中心节点外的其他保存节点，是独立运行的一套保存系统，独立执行保存功能，独立运维保存系统，包括本地PID分配、本地权限管理机制。

（3）多种类型节点分类。令保存网络的节点具备不同功能，以完成不同的分工，执行不同的任务，除中心管理节点外，保存节点亦因所具备的功能不同而加以区分，以保证参与机构能以不同角色参与保存网络的建设与运营。

（4）松耦合。各节点主动向中心节点推送信息，各节点关闭、停止、撤销，不影响整个保存网络的正常运行。

（5）异构。允许参与保存的机构采用任何保存系统，只要遵循相关标准，向中心节点推送相关信息，即可成为保存网络的节点。这种设计使保存体系的参与机构能够采用不同的系统，保存不同类型的数字对象，同时使保存网络易于扩展。

（6）分布式协作保存网络需架构灵活支持扩展。要求分布式协作保存网络不仅能够随着规模的扩大不断增加节点，也能够根据参与合作保存的机构不断增加节点。

（7）参与机构的角色可灵活转换和扩展。每个保存机构独立维护运营自身网络节点，但无法确保每个机构都能对保存系统提供长期有效的支持。从长远发展的角度考虑，保存网络应支持每个机构在保存网络中的角色和功能的转换、扩展，这就要求保存节点的软件系统的功能可方便扩展和转换。

3 国家协作保存网络整体框架

国家协作保存网络是一个分布式协作保存网络，主要由两层结构组成：保存管理层和保存执行层。保存管理层通常是一个中心节点，接收来自其他节点的信息；保存执行层的每个节点分别负责不同资源的保存管理（见图1）。

3.1 中心管理节点

保存网络中心管理节点能实现对所有保存节点的统一管理，通过提供注册功能，对各保存节点的相关管理信息进行存储，同时借助推送功能，获得各保存节点的存档信息。管理节点通过接收的保存节点数据，对整个保存网络进行监督和管理。

中心管理节点的主要功能包括保存协议信息管理、各节点信息管理、公共服务管理、命名空间管理、存档数据管理、各种报告管理、保存规划管理、备份管理、硬件管理、人员权限管理等。

3.2 保存（执行）节点及功能分类

保存（执行）节点根据部署平台功能的不同，可执行不同的保存任务。目前可以划分为四种类型。

（1）F型（Full functions），提供具备完整保存功能的节点。部署包括DPS服务器、Fedora服务器在内的一套完整系统，执行全部数据的保存任务，包括资源摄入、存储、管理、备份、公共服务等，是一个独立执行完整保存功能的节点。

（2）P型（Preservation），提供资源摄入与保存管理功能的节点。部署接收与摄入平台以及保存管理平

台，包括DPS服务器、Fedora服务器，执行全部的数据保存任务，包括资源的摄入、存储、管理等，不能提供公共访问服务。

（3）B型节点（Backup），提供备份功能的节点（仅提供备份的功能）。

（4）A型（Access），提供公共访问服务的节点。A型节点利用其存档的资源为终端用户提供对存档资源的访问服务，通常与其他保存节点分隔开，以保证保存网络的安全性。存档节点通过推送功能向访问服务节点单方向推送数据。

从整体考虑，保存网络中心管理节点也可作为保存网络的一种独特类型，即M型（Management），不执行具体的保存功能，只提供对其他执行保存节点的管理和监督功能。

3.3 保存系统及软件平台功能

协作保存网络各节点采用的长期保存系统，是以中国科学院文献情报中心长期保存系统为基础，经过模块化改造，目前已形成包括接收与摄入管理平台、保存管理平台、公共服务平台和合作保存服务平台四个主要功能组成的可靠保存系统（见图2）。

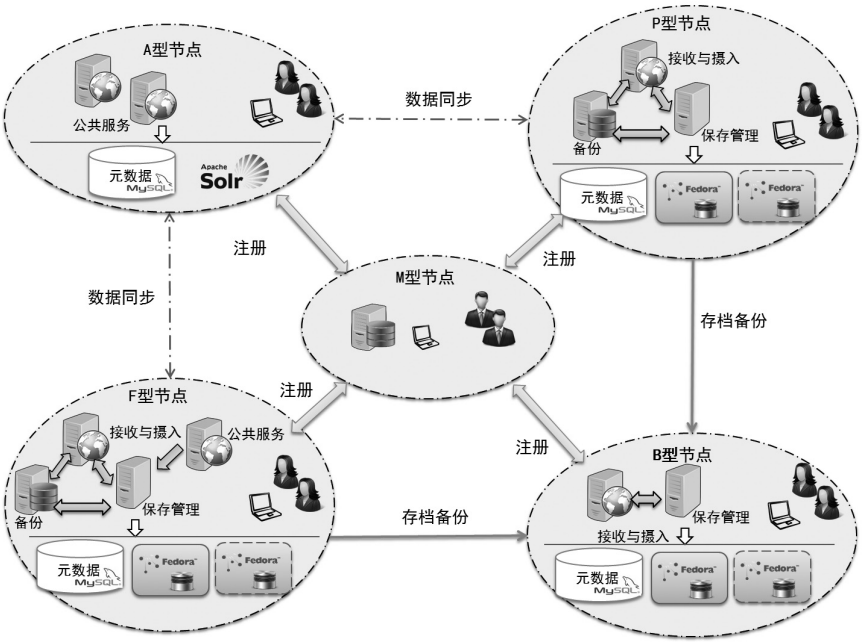

图 1 国家协作保存体系整体架构示意图

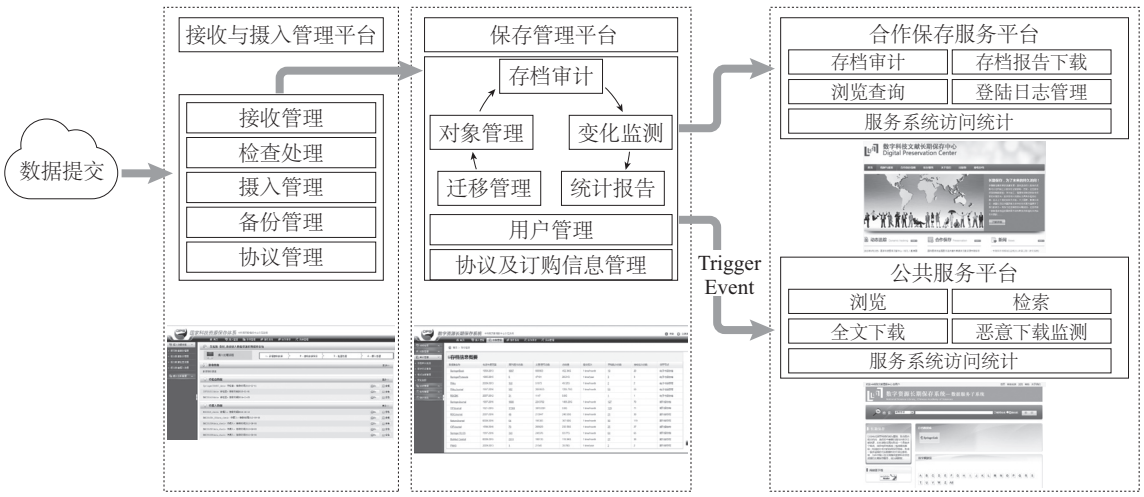

图 2 保存节点系统各组成平台功能示意图

（1）接收与摄入管理平台。对已登记的数据源定时进行数据提交包的检测和下载，并对已下载的数据包进行可用性检查、病毒及恶意代码检测。检测通过的数据认为是可以接收的数据，在系统中进行登记，同时自动通知系统预定义的资源摄入处理人员和备份管理人员进行备份处理和摄入处理。摄入处理人员按照批次进行接收并定制摄入任务，由任务在后台完成数据包清点，进行数据包完整性检验，检查数据包内数据格式以及内容是否完备，抽取描述元数据、保存元数据、技术元数据等，生成符合国家保存体系要求的标准SIP，最后将数字对象存入底层的Fedora仓储系统，同时更新外部Mysql管理数据库和Solr索引。

（2）保存管理平台。保存管理平台提供对存档数据的长期管理和运维，确保数字对象长期可用。审计功能是按照保存协议定期检查存档数据的完整性，如可提供数据集、期刊、文章的完整性检查。数据不变性的检查校验，负责检查数据内容是否未经任何改变；变化追踪功能，可查看数据内容本身变化的历史情况；格式监测功能，可以定期监测数据格式是否过时；统计报告功能，可以提供存档信息统计概要，实现对存档处理过程的追踪和检查，以及生成各种报告；迁移功能，指支持数字对象在保存系统的迁移以及媒体迁移。

（3）公共服务平台。公共服务平台执行保存系统的分发功能，采用黑色存档（Dark Archive）模式，即正常情况下不对外提供服务，只有在触发事件的激发下（如因网络中断、战争、公司倒闭等因素、无法获取某一数据库的正常检索服务时）才能够提供公共服务。公共服务平台用于对保存协议规定范围的用户提供存档范围的数据访问服务，包括检索、浏览、全文下载等功能。

（4）合作保存服务平台。合作保存服务平台执行保存系统的另一部分分发功能，用于对参与合作保存的机构（出版商、存档机构）提供检查和审计服务，机构用户可以通过该系统了解保存系统内数据的存档情况、获取存档统计报告、对存档数据进行审计。

目前这四个平台可以联合部署，也可以独立部署。如F型保存节点需要部署运行包括这四个平台的完整保存系统，即可实现存档数据的接收、检查、摄入、审计等相关功能，并可通过公共服务系统对合同范围的用户提供访问服务；而A型保存节点只需要部署运行公共服务平台，仅提供公共访问服务。协作保存网络各节点通过部署和运行具备不同功能的保存平台软件实现不同的功能，同时达到协作的目的。

4 协作保存网络保存关键问题分析

4.1 基于注册的管理机制

基于上述体系架构，协作保存网络采用松耦合机制，通过信息注册方式实现对协作网络的监督和管理，即各节点的相关信息统一汇聚到管理节点，管理节点可及时掌握各节点的情况。

各节点需要注册的数据包括以下10种。（1）保存协议信息注册管理：如期刊清单；（2）节点注册管理：节点、机构、存档资源的基本信息；（3）公共服务注册管理：各存档资源应该由哪些机构的哪些节点提供服务以及具体部署情况；（4）命名空间注册管理：各存档节点PID登记和分配；（5）存档数据管理：汇集各存档节点的数据存档情况，通过各节点实时提交的方式实现；（6）审计与报告管理：各存档节点定期推送审计报告、统计报告；（7）保存规划管理：汇集各存档节点为各资源制定的保存规划；（8）备份管理：各资源备份实施情况的注册管理；（9）硬件管理：汇集各节点参与保存工作的硬件设备信息；（10）人员管理：各节点参与保存工作的人员权限管理。

4.2 主动推送的数据提交模式

相关数据采用各节点向管理节点主动推送的模式，共三种实现方式。（1）增加功能模块，直接嵌入摄入工作流，每次存档数据摄入完毕都自动推送数据。这种方式更适用于存档数据信息，可以实时提交存档情况，其缺点是每种数据的存档和更新都需要重新调整程序。（2）定制自动调度任务，定期（如每月底）进行1次推送/推荐。缺点是不能实时推送，适用于审计报告、统计报告等提交。（3）手工启动推送数据服务。适用于保存管理平台和公共服务平台间的数据交换，当触发事件发生时，向公共服务平台传输用于公共服务的数据。

4.3 不同类型节点间的协作模式

基于注册机制的协作保存网络拥有可扩展的、灵活的架构。每个参与协作保存的机构可作为一个独立的节点，或每个机构可有几个节点。协作保存网络中多个异地异构节点间可进行多个层面的协作：（1）中心管理节

点对各保存节点进行监督管理；（2）F型保存节点的公共服务平台和备份功能，也可为其他多个保存节点提供服务；（3）访问服务节点可为其他保存节点提供公共服务；（4）备份服务节点可为各类型节点提供备份服务；（5）保存节点间可以作为镜像备份（系统级备份）。

通过对这些异地异构提供不同功能的节点组配，既可以构成只支持多重备份的简单协作保存网络，也可以构成支持多个层面协作的复杂协作保存网络。

4.4 唯一持久标识

协作保存网络制定了命名空间（namespace）管理要求和唯一持久标识符（Persistent Identifier，PID）定义规范。由管理节点进行命名空间管理和统一分配，每个保存结点定义唯一的命名空间，在结点注册时进行分配和验证。每个节点根据所分配的命名空间进行本地PID分配和管理，确保每个数字对象的PID全局唯一。

4.5 扩展性问题

协作保存网络需能随规模的扩大而不断增加节点，也能够支持参与机构的角色可灵活转换和扩展。

目前协作保存网络可从多个层面进行扩展。首先，协作保存网络中的节点数没有上限，可以随需要部署保存系统并在中心节点注册，即可新增一个节点。其次，已存在的节点可以根据自身需要增加底层Fedora仓储系统的部署数量，扩大存储规模；保存节点还可以增加部署不同的平台，即可增加相应功能来完成更多任务。最后，协作保存网络允许参与保存的机构采用任何保存系统，只要遵循协作保存网络的规范，能够向中心节点推送相关信息，即可作为节点加入，协作保存网络也可为其提供公共服务和备份服务。这种异构兼容使保存体系中的参与机构能够采用不同的系统保存不同类型的数字对象，也使保存网络易于扩展。

5 现状与展望

目前协作保存网络的中心管理节点正在试运行，已经建成的中国科学院文献情报中心和中国科学技术信息研究所保存节点以及正在建设的北京大学图书馆保存节点，采用相对简化的部署，3个节点采用统一的架构和平台系统，独立部署和运行完整的保存系统（见图3）。

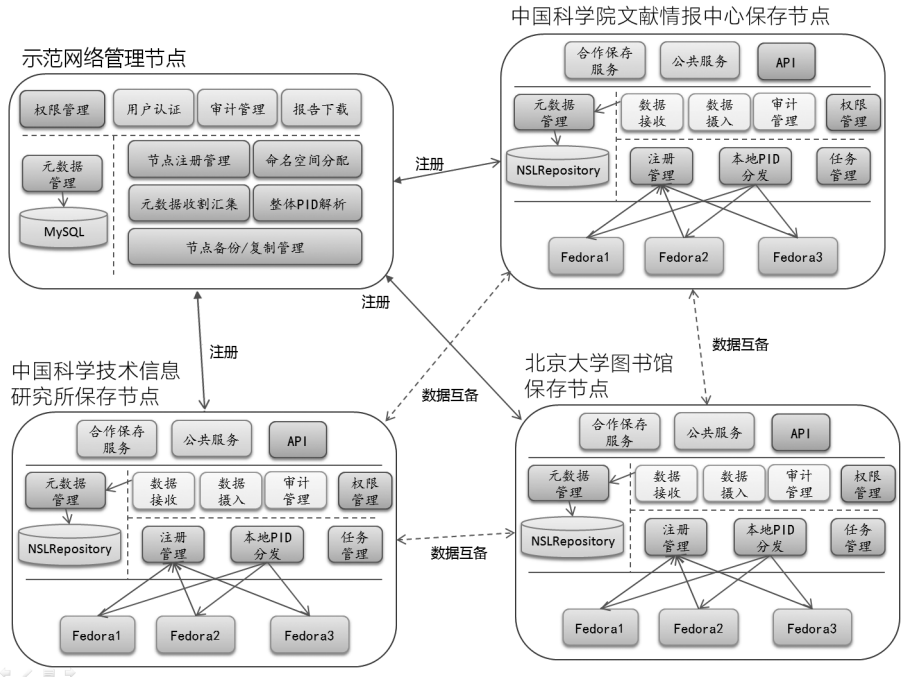

图 3 协作保存网络当前部署示意图

中国科学院文献情报中心节点已保存7家出版社（Springer，Wiley，IOPP，NPG，BMC，RSC，VIP）共13种外文资源，近3 000种外文期刊、1.4万种中文期刊、4 000万篇论文、7.5万种电子图书、3.4万种实验室指

南；中国科学技术信息研究所节点已对43家出版机构的655种现刊期刊进行保存处理。

数字资源长期保存是一项复杂的系统工程，涉及众多利益方与复杂的技术管理，并需要长期经济支持。作为国家科技文献保障体系的组成部分，国家保存体系将站在国家的利益高度和全局的协调角度，带领全国图书馆争取和保持数字文献资源的本土保存权，统筹规划保存目标资源和合规保存机构，支持和组织协助合规保存机构进行长期保存谈判，组织安排对合规保存机构长期保存机制及其保存效果的公共认证和审计，监督和审计必要情况下的公共服务，统筹协调必要的备份和继承保存。

按照发展规划，截至2016年年底，国家保存体系将完成保存网络的初步建设，完成3家国家级长期保存中心建设，保存一定规模的权威国际数字科技文献资源，并建立比较完善的长期保存运行、管理和服务规范。2017—2020年，将持续进行保存体系完善和扩展，选择一批具有较大规模的数字文献资源采购和使用量，具有可靠的经济、技术和管理条件的公共事业单位作为合规保存机构，按照分工，接受委托，承担相应的保存任务。同时保存资源规模也将持续扩展，保存多数重要国际科技期刊及其他重要资源，形成巩固的国家数字科技文献长期保存体系。

国家保存体系致力于在我国本土进行数字资源保存，这项工作面临来自多方面、巨大的困难，需要更多机构的参与和支持，更需要有条件、有资质的机构投身国家保存体系，共同承担国家数字资源长期保存的重任。

参考文献

[1] 高建秀,吴振新.数字资源协作保存网络研究[J].图书馆学研究,2010(23):26-31,25.

[2] 付鸿鹄,吴振新.分布式数字资源保存系统与技术架构研究[J].国家图书馆学刊,2015(2):82-88.

作者简介

吴振新，女，1968年生，研究馆员，硕士生导师，研究方向：数字资源的采集、组织管理、长期保存及再利用，E-mail：wuzx@mail.las.ac.cn。
付鸿鹄，女，1976年生，馆员。

Construction a Distributed Collaborative Network for Digital Information Resource

WU ZhenXin, FU HongHu
(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: In this article, based on previous research and requirements analysis, the author clarified the developing principles of national collaborative preservation network, such as central registry, independent node management, multi-typed nodes, loosely coupled, heterogeneous, flexible network architecture, etc. Then, the author described the overall framework of the collaborative network and functions of each type node. It also provided the key issues solutions including the registration-based management mechanism, data exchange model, cooperative mode between different types of nodes, persistent identifier and system scalability.

Keywords: Long-Term Preservation; Digital Information Resource; Collaborative Preservation; Preservation Network

（收稿日期：2016-08-17）