国家数字科技文献资源长期保存体系的战略与实践

张晓林 吴振新 赵 艳 付鸿鹄 张智雄 郑建程
（中国科学院文献情报中心）

摘 要 为解决社会对数字文献资源日益增长的需求和数字文献资源处于多重威胁、“濒于损毁”的使用现状之间的突出矛盾，国家科技图书文献中心牵头组织了国家数字科技文献长期保存体系的建设，在中国大陆本土长期保存对我国具有重要价值的国内外数字文献资源，确立了合理平衡的长期保存权益管理制度，形成了国家主导、联合参与、可靠管理和公共服务的长期保存体系，建立了可靠高效的长期保存技术流程和技术系统，提出了公共保存审计和保存机构认证机制，已经保存了一大批重要的科技数字文献资源，为我国数字知识内容资源的长期保存奠定了坚实基础。

关键词 数字文献资源 长期保存 国家科技图书文献中心 NDPP

DOI: 10.13663/j.cnki.lj.2017.12.002

Strategies and Practices of National Digital Preservation Program for Scientific Literature

Zhang Xiaolin, Wu Zhenxin, Zhao Yan, Fu Honghu, Zhang Zhixiong, Zheng Jiancheng (National Science Library, Chinese Academy of Sciences)

Abstract In order to meet the challenges presented by the increasing demands for digital literature resources as well as the threats the usage involves, the National Science and Technology Library of China has developed the National Digital Preservation Program for Scientific Literature (NDPP). NDPP aims to preserve the domestic and foreign digital literature resources of great values to China's research, innovation, and development in the long run. It has devised a balanced preservation rights management mechanism, formed a state-led preservation network of joint participation and reliable management that is based on public service, established a trustworthy and efficient preservation tech workflow and system platform, put forward public preservation audit and certification processes, and already preserved a large number of important scientific digital literature resources, all of which laya solid foundation for the long-term preservation of digital knowledge content resources in China.

Keywords Digital literature resources, Long-term preservation, NSTL, NDPP

1 数字文献资源持续利用的需求与挑战

1.1 数字文献资源已成为科研教育基础设施

数字文献已经成为科技领域学术信息创作与出版的主要形态。据国际出版社 Springer 2013 年介绍，自 1996 年它的所有期刊首先以数字形式出版，自 2004 年它将所有期刊的回溯卷期全部数字化；自 2006 年它的所有图书首先以数字形式出版，自 2011 年它将以前出版的所有图书数字化，因此它的所有出版物已经全部数字化、以 e-First 形式出版[1]。这实际上是所有国际科技出版社的基本出版形态。

数字文献已经成为我国科技信息用户的主流信息资源[2]。2016 年，中科院电子期刊占外文期刊订购费 85% 以上和使用率 99% 以上、占中文期刊订购费 90% 以上和占使用率 99% 以上。据高校数字资源采购联盟（DRAA）介绍，主要高校采购了 7 824 馆次 136 个数据库，中外文期刊数据库已成为默认基础资源。实际

上，数字文献资源为主的信息保障体系已成为各国科技教育不可或缺的基础设施之一。

1.2　数字文献资源实质上已处于“濒于损毁”的状态

数字文献资源的普及在极大提高用户获取信息能力的同时也带来了可持续利用的严峻挑战，技术角度的挑战包括：存储介质的长期可靠性，数据格式的长期可用性，存储系统的长期可靠性以及在变化格式及其提供机制下的内容完整性和内容的可使用性。

而且，当前数字文献资源采购获得的只是采购网络使用权，采购方一般不能对所采购资源进行本地长期保存，这带来新的严重挑战：网络链接可能中断，出版社可能停止出版某些出版物，出版社可能被并购或者倒闭，检索获取服务系统可能因为网络攻击、技术故障、人为失误等无法使用，出版社可能因为地缘政治因素中止对特定市场服务，更不用说频发的自然灾害甚至战争等带来的危害。

因此，一方面，数字科技文献资源成为科学研究、技术创新、教育、经济与社会发展的基础战略资源；另一方面，我国用户对国际数字科技文献的获取处在多种复杂因素的高度威胁下，原来基于图书馆分散馆藏的保存机制已无法应对这些挑战，数字科技文献资源实质上处于“濒于损毁”的状态，它们在我国本土的长期可靠保存和可持续利用已经成为国家的重大战略需求。

1.3　各国持续努力应对数字文献资源长期保存挑战

世界各国都采取了多种机制来应对这种挑战。例如，荷兰国家图书馆基于法定存缴制度要求所有在本国的出版社（包括了 Elsever、Springer 等）将其数字期刊存储到该馆的 e-Depot 存储系统[3]，大英图书馆推动国家立法保证自己作为国家图书馆能长期保存在英国出版的数字化学术出版物[4]，美国国会图书馆也建立了对本国出版电子学术期刊的长期保存机制。同时，图书馆界和出版界也通过多种联合机制支持数字学术文献的长期保存，例如准商业化的 Portico[5] 和联盟化的 CLOCKSS[6]。像 PubMed Central[7] 和 HathiTrust[8] 这样的专业知识库和数字图书馆系统也对数字资源可持续服务提供了重要的支持。

2　中国行动与 NDPP 总体目标

2.1　中国已在数字文献资源长期保存方面持续努力

中国图书馆界在数字资源长期保存方面进行了长久的努力。2003 年中科院与斯坦福大学进行了 LOCKSS 项目合作实验；2004 年中科院发起在北京召开了 iPres2004 会议（International Conference on Digital Preservation），邀请欧洲多名专家来华共同讨论数字文献资源长期保存的挑战与措施，促成了 iPres 系列会议逐步成为全球数字信息资源长期保存的权威学术会议；2004 年 9 月，国家科技图书文献中心（NSTL）委托中科院文献情报中心开展“数字化科技信息资源长期保存体系与政策机制”研究；自 2007 年开始，中科院启动“数字科技文献长期保存服务系统建设”项目，研究长期保存的权益、技术、系统、管理、可信赖审计等，并与国内外重要科技出版社合作进行长期保存实践，2009 年首次与 Springer 签署长期保存协议并开始存储其电子期刊，随后又与 NPG、Wiley 等开始了长期保存合作。

2.2　NSTL 启动国家长期保存示范体系建设

2013 年科技部正式同意由 NSTL 牵头组织数字科技文献资源长期保存，2014 年 NSTL 开始国家数字科技文献长期保存示范系统（National Digital Preservation Program，NDPP）建设[9]，履行国家平台职责，深化资源保障能力，创新合作服务机制。

NDPP 总体目标是，作为国家科技信息资源的战略保障，长期保存我国科技创新用户所需的主要数字科技文献资源，同时积极参与国家教育文化社会各领域所需的其他数字文献资源的长期保存。

NDPP 的总体任务框架如图 1 所示：

NDPP 的运行原则为：①国家主导：作为国家科技文献保障平台的有机部分，由国家投资，由 NSTL 管理；②联合参与：动员国内图书馆合作参与，并选择若干图书馆作为“合规保存机构”联合承担保存任务；③可靠管理：

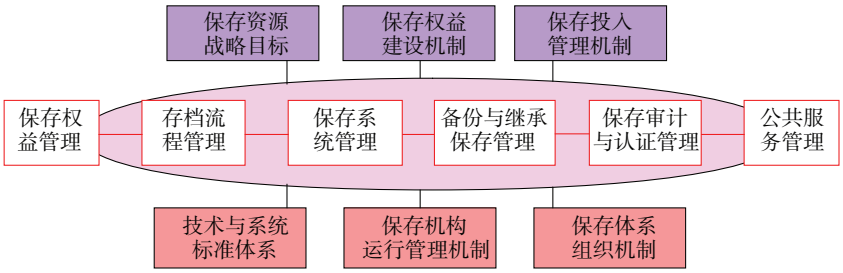

图 1 NDPP 数字科技文献资源长期保存工作框架

保证重要数字科技资源在大陆本地可靠长期保存，保证各个长期保存系统的可靠运行和可靠监管；④公共服务：长期保存体系本身作为公益性公共服务由国家支持。

2.3 NDPP 长期保存资源的选择标准

（1）NDPP 保存资源选择标准：选择使用价值高、学术价值高、风险高、保存可操作性强的数字科技文献资源，优先部署长期保存，并创造条件保存其他具有保存价值和消除风险作用的数字文献资源。使用价值体现在当前及长期需求量、包括采购机构范围和实际使用量；学术价值主要考虑文献资源在科学研究和国际发展中的长期学术价值、且保证系统性历史地覆盖重要资源；风险程度主要考虑资源依赖度、资源内容或使用渠道不可替代性及其风险程度；可操作性主要考虑权益安排可接受度、内容格式规范化程度、所需保存技术的可掌握程度、保存机制的可负担性、保存合作协调能力等。

（2）NDPP 保存资源范围和类型：首要目标资源范围包括自然科学、工程技术领域，其他范围包括与自然科学、工程技术密切相关的管理、经济、行业、市场、金融等领域。目标资源的类型包括：①上述学科范围（下同）的主要国内科技期刊数据库、主要国际综合出版商的科技期刊数据库和科技会议录数据库及这类数据库的回溯数据内容；②主要专业学会/协会和专门出版商的科技期刊数据库和科技会议录数据库；③主要的开放获取期刊、会议录、专业知识库等数据库；④重要数字科技专著数据库和重要数字工具书资源；⑤其他重要数字知识资源（例如学位论文、科技报告等）。

3 NDPP 的运行机制

3.1 NDPP 的运行目标

NDPP 在 NSTL 领导下，通过一组内在分工合作的合规保存机构来具体实施期保存。合规机构的条件包括：公共事业法人，本身采购数字资源并提供服务；支持 NDPP 长期保存战略目标和运行目标，参加或承担被保存资源的保存谈判；承担长期保存的法律责任、接受公共认证审计；提供被保存资源在触发事件下的公共的获取服务，支持特殊情况下的继承保存；遵守长期保存体系的法律、技术和管理规范，提供长期保存系统的基础设施、人员和管理支持。

NDPP 建设分为三个阶段：①示范系统建设，2014 年，建设中科院文献情报中心和中国科技信息研究所示范节点，保存一批重要资源，形成符合可信赖保存国际认证要求的长期保存示范系统，形成权益、技术、运行、服务和审计等规范。②基本保存系统建设：2015–2016 年，扩展建设若干个国家级长期保存中心，保存相当规模的权威的国际数字科技文献资源，建立比较完善的长期保存运行、管理和服务规范。③体系完善和持续运行：2017–2020 年，保存多数重要国际科技期刊库和主要的国际会议论文、学位论文、专利库，保存若干国际科技图书数据库，形成巩固的国家级数字科技文献长期保存体系，完善长期保存运行、管理、服务和公共审计机制。

3.2 NDPP 的长期保存权益管理权利

全国数百家图书馆在 2015 年签署的《数字文献资源长期保存共同声明》[10] 指出，图书馆对所采购数字文献资源拥有长期保存的权利。图书馆采购数字文献资源实质上是采购知识内容为用户提供当前和未来服务，因此资源

的长期可靠利用是这种采购行为的内在含义和基础条件，没有可靠的长期保存，图书馆对所采购知识内容的长期利用权就不完整和不可靠，因此长期保存是图书馆持续进行资源采购的必要基础，其实是图书馆对出版社销售资源的支持条件。

图书馆对数字文献资源的长期保存权包括数据存档权、数据处理权和数据服务权。数据存档权，即对相关数字内容完整、可靠、及时地进行摄入和保存的权利；数据处理权，即因长期保存需要对存档数据进行准确的检验、转换、提取或加载元数据、迁移、重新格式化等处理的权利；数据服务权，即在出版社无法提供服务的情况下利用所保存资源向相应用户群提供检索获取服务的权利，直至出版社恢复正常服务。而且，为提高长期保存的效益，图书馆拥有委托国家保存体系或其他图书馆进行长期保存的权利，受委托进行保存的图书馆有在自己不能继续履行保存责任时选择能合理保护各方合法权益的合作保存机构继承保存的权利。合作保存权和继承保存权是保存权的自然延伸，也是长期保存权利平等实现的基础。

NDPP 受 NSTL 委托、代表国家和联合签署上述共同声明的图书馆行使上述长期保存权。NDPP 充分承认和维护出版社的合法权益，承担诚信尽责保护出版商合法权益的义务，承诺建立可信赖的法律、管理、技术等措施保证相关利益方遵守在长期保存中的权利与义务。

NDPP 建立行使长期保存权的制度化机制，包括推动国家确立公共资金采购数字文献资源的长期保存责任原则，所有使用公共资金采购数字文献资源的图书馆应要求所采购资源在中国本土得到可靠长期保存，并将长期保存权利纳入与出版社签署的采购合同中。

3.3　NDPP 的可信赖管理机制

数字文献资源长期保存是一种风险防范机制，需要确保被保存资源在任何技术、经济、市场和管理状况下的长期可用性。但长期保存本身又是一项涉及复杂的权利、技术、过程和大量经济与管理投入的长期工作，可信赖性是长期保存服务的运行基础和核心能力。

长期保存服务的可信赖性应按照国际通行的长期保存服务体系模型（如 OAIS 参考模型[11]）和长期保存服务可信赖性标准体系来设计和检验，按照可靠公共服务所遵循的最佳实践和可持续市场服务所遵循的最佳保障机制予以管理。

为此，国家保存体系将：①建立明确的长期保存规划与政策，清楚阐述长期保存目标、权利、服务机制、可持续性保障机制等，作为指导长期保存服务运行的基本准则；②建立可靠的长期保存权益管理机制，建立符合法律要求和覆盖长期保存全过程的权利与义务体系，建立具有法律约束力和可操作性的权益管理执行流程；③建立覆盖长期保存全生命周期的业务流程管理机制，保证整个流程及其所有部分都得到可靠和高效的技术方法与系统的支持、都得到可操作和可检验的规范与制度的支持；④建立可靠的长期保存技术系统，全面支持长期保存全生命周期的所有任务、遵循可信赖保存技术系统国际标准、能与各利益相关方系统有效兼容和与未来技术变化有效兼容；⑤建立权责体系明晰、具有高水平知识和能力、以业务流程各阶段有效契合的长期保存团队；⑥建立稳定健康的长期保存经济投入机制，保证经济上合理高效、成本核算科学清晰、预算有效执行、运行得到持续评估；⑦建立可靠的继承保存机制，通过事先建立的具有约束力的关于继承责任、继承条件、继承流程、权益转移规则、数据迁移标准、经济与管理责任转移规范等的规定以及可靠的先期测试，保障在必要时能顺利无损地实现继承保存。

数字文献长期保存作为一种公共服务，其可信赖性依赖这个服务及其管理过程的公开、透明和可检验上。国家保存体系自觉建立公开的自我保存审计、第三方保存审计、保存机构可信赖性公共认证等机制，纳入公共力量来监督长期保存系统的运行，防止因自身人员疏忽、管理懈怠、技术失误、经费缺失等原因造成保存内容的损害。

4　NDPP 长期保存流程与系统

4.1　NDPP 长期保存的宏观流程

长期保存基于一系列研究、技术和管理流

程，以保证长期保存活动的科学高效可靠。这个流程至少应包括以下步骤：

长期保存需求分析：明确所服务的目标社区及其需求，把握相关的数据内容生产者及其产品状况，跟踪相关的信息技术发展状况分析，开展数据内容产品的风险评估等。

长期保存责任体系组织：协助各个机构确定自己在国家保存体系中的合适角色，包括承担保存任务的合规保存机构和其他参与合作保存的机构，后者将通过协调采购政策、参与保存谈判、参与保存审计和合规保存机构公共认证等，协助和监督国家保存体系的可靠运行。

长期保存协议权益获取：合规保存机构根据分工向出版商提供提出长期保存要约，并结合资源订购谈判进行长期保存协议谈判；长期保存协议签署后在NDPP登记，保存机构对长期保存协议的执行接受NDPP的保存审计。

长期保存技术系统建设：长期保存系统的技术架构坚持OAIS框架，系统功能和技术流程设计遵循OAIS和ISO16363标准，支持长期保存生命周期的各个功能环节，确保对主流技术标准的长久兼容，并建立技术方法审计与更新机制。

公共服务管理：在NDPP与出版社共同认可的触发条件下，被保存资源将被用于向原采购用户范围提供公共访问服务，并建立用户接入管理、公共服务监管、各方权益保护、服务效果评价等机制。

长期保存审计与认证管理：NDPP将安排对合规保存机构所保存资源的年度保存审计，保证每个机构每年有被保存资源接受保存审计，每三年对所有被保存资源都进行保存审计，保存审计结果向各个合规保存机构和所有参与保存机构通报。

保存方案包括《保存管理方案》和《保存技术方案》两个部分具体流程和要求、说明特色功能和需求等。

4.2 长期保存的技术流程

NDPP和保存机构针对每种被保存资源制定接收、检验、摄入流程。数据接收周期性进行，根据长期保存协议规定按周、月或季度接收并处理存档数据。

数据接收管理：保存机构在长期保存协议规定的时间下载出版社按照协议确定的格式提供的被保存内容（SIP），并对SIP自动进行病毒检查、恶意代码检测、完整性检查、一致性检验等，并形成数据接收检查报告。

数据摄入管理：保存系统对检查合格的数据建立符合保存系统统一标准的存档包（AIP），包括描述元数据抽取、保存元数据抽取、文档格式识别与技术元数据抽取等。系统将AIP摄入到存储与管理模块，同时生成存档信息统计清单以支持自动保存审计。

数据存档管理：为确保被保存内容在长期保存全生命周期都能够保持完整性、真实性、可理解性，数据摄入后要对数据内容进行长期有效管理，包括内容更新、完整性审计、不变性检查等。以期刊数据库为例，完整性审计包括对接受的各批次存档数据、具体期刊、具体文章的完整性审计。

数据保存管理：包括存档文件不变性检查、存档文件格式检查、备份有效性检查、存储介质有效性检查、保存策略检查与更新、保存设备管理等。

灾备管理：建立数据备份制度，避免因为灾害发生造成保存数据全部或部分丢失，并能在灾难发生后以最快速度恢复数据与服务。具体工作包括：数据安全分析、建立数据备份策略、建立安全性检查制度、确定灾后恢复策略、进行灾备测试和灾备制定评价更新。

5 NDPP保存资源及后续努力

5.1 NDPP已保存资源

截至本文发稿，NDPP已对大批数字文献资源长期保存，其中：在中科院文献情报中心节点，保存了Springer期刊库、NPG期刊库、Wiley期刊库、RSC期刊库、PNAS期刊库、IOP期刊库、BMC期刊库、Springer RLOS期刊库、AGU期刊库、维普科技期刊库、Springer实验室手册数据库、Springer电子图书数据库、Wiley电子图书数据库、IOP电子图书数据库、RSC电子图书数据库等资源。

在中国科技信息研究所节点，将对NSTL采购并已获得长期保存权利的45种文献数据

库进行长期保存，包括美国冷泉港实验室期刊库、美国气象学会期刊库、美国动物学会期刊库、澳大利亚科学院出版社期刊库等学协会和大学出版社期刊库。目前已有 19 家出版社的内容得到保存，并将在 2018 年完成所有数据库的长期保存。

5.2　NDPP 后续长期保存努力

NDPP 中科院节点、中信所节点和北京大学节点已经与万方科技文献数据库、Emerald 期刊数据库、ProQuest 学位论文数据库、CUP 期刊数据库的出版方签署了长期保存协议，将在 2018 年实现这些资源的长期保存。

NDPP 及其各节点将在全国参与数字文献资源合作保存的所有单位的协助下，推进与其他重要文献出版机构的长期保存谈判，在 2020 年实现所有重要数字文献资源在我国本土实现长期保存。

NDPP 将进一步完善长期保存的可信赖可检验机制，完善 NDPP 管理与服务体系，推动面向参与数字内容采购与服务的所有机构乃至社会的长期保存宣传培训，完善长期保存的技术体系和运行基础设施，并与国际图书馆界和数字文献资源长期保存机构开展合作。

与此同时，NDPP 还将探索和试验其他形态的数字内容资源的长期保存，包括数字音像资源、科学数据资源、社交媒体资源、数字人文资源、交互式多媒体数字艺术资源等的长期保存。

经过多年努力，我国数字文献资源长期保存的战略和工作框架已经建立，大批资源已经得到保存，统筹协同开展长期保存的态势和机制已经形成。但是，数字内容资源规模迅速增长、形态不断变化、服务持续创新、市场日益复杂，我国建设创新型国家中日益增长的数字内容资源需求和这些资源长期可靠利用存在的严峻风险之间的矛盾依然存在，全国学术界、图书馆界和出版界仍需艰苦努力来确保数字内容资源作为战略基础资源能可信赖地长期保存和永续利用。

参考文献

[1] Kwong, Maurice. STMe Book Publishing. High level forum between international publishers and Chinese libraries, Beijing, 2013-08-29.
[2] 张晓林. 数字文献资源长期保存的战略意义、主要任务与主要机制. 数字文献长期保存研修班. 上海, 2017.8.
[3] KB. Long-termusability of digitalresources. https://www.kb.nl/en/organisation/research-expertise/long-term-usability-of-digital-resources.
[4] Maureen Pennock, Digital Preservation at British Library. 数字文献资源长期保存香山科学会议, 2016.11.
[5] Portico. http://www.portico.org/digital-preservation/.
[6] CLOCKSS Archive. https://www.clockss.org/clockss/Home.
[7] Pub Med Central. https://www.ncbi.nlm.nih.gov/pubmed.
[8] Hathi Trust Digital Library. https://www.hathitrust.org/.
[9] 国家科技数字文献资源长期保存体系.http://www.ndpp.ac.cn/.
[10] 国家科技图书文献中心. 数字文献资源长期保存共同声明, 2015.9. http://www.nstl.gov.cn/NSTL/facade/news/newsInfo.do?act=toNewsContent&id=125164.
[11] Reference Model for an Open Archival Information System (OAIS). https://public.ccsds.org/Pubs/650x0m2.pdf.

张晓林　中国科学院文献情报中心，研究员，教授。研究方向：数字知识系统的理论、技术与实践。作者贡献：体系设计、项目组织、权益管理设计，参与公共审计与认证设计，论文撰写。E-mail：zhangxl@mail.las.ac.cn　北京　100190
吴振新　女，中国科学院文献情报中心，研究馆员。研究方向：信息技术与信息系统。作者贡献：技术流程与技术系统设计开发，公共审计与认证组织，协助项目组织。　北京　100190
赵　艳　女，中国科学院文献情报中心，副研究馆员,博士研究生。研究方向：信息资源组织与管理。作者贡献：协助权益管理设计，权益管理谈判。　北京　100190
付鸿鹄　女，中国科学院文献情报中心，副研究馆员。研究方向：信息技术与信息系统。作者贡献：参与技术流程与技术系统设计开发，协助项目组织。　北京　100190
张智雄　中国科学院文献情报中心,研究员。研究方向：信息系统和智能信息处理。作者贡献：协助技术流程与技术系统开发。　北京　100190
郑建程　中国科学院文献情报中心,研究馆员。研究方向：信息资源组织与管理。作者贡献：合作机制设计、参与公共审计与认证组织。　北京　100190

（收稿日期：2017-11-27）