国家社科基金
GUOJIA SHEKE JIJIN HOUQI ZIZHU XIANGMU
后期资助项目

数字资源长期保存技术的研究与实践

STUDY ON DIGITAL PRESERVATION TECHNOLOGIES AND PRACTICES

张智雄 等 著

國家圖書館出版社
National Library of China Publishing House

国家社会科学基金后期资助项目研究成果

数字资源长期保存技术的研究与实践

张智雄　等　著

國家圖書館出版社
National Library of China Publishing House

图书在版编目(CIP)数据

数字资源长期保存技术的研究与实践/张智雄等著. --北京：国家图书馆出版社，2015.9

ISBN 978-7-5013-5641-6

Ⅰ.①数… Ⅱ.①张… Ⅲ.①数字技术—信息管理—研究 Ⅳ.①G203

中国版本图书馆 CIP 数据核字(2015)第 166603 号

书　　名	数字资源长期保存技术的研究与实践
著　　者	张智雄等　著
责任编辑	高　爽　王炳乾

出　　版	国家图书馆出版社(100034　北京市西城区文津街 7 号) (原书目文献出版社　北京图书馆出版社)
发　　行	010-66114536　66126153　66151313　66175620 66121706(传真),66126156(门市部)
E-mail	btsfxb@nlc.gov.cn(邮购)
Website	www.nlcpress.com ——→投稿中心
经　　销	新华书店
印　　装	北京华艺斋古籍印务有限责任公司
版　　次	2015 年 9 月第 1 版　2015 年 9 月第 1 次印刷

开　　本	710×1000(毫米)　1/16
印　　张	36.25
字　　数	650千字

书　　号	ISBN 978-7-5013-5641-6
定　　价	80.00 元

国家社科基金后期资助项目
出版说明

后期资助项目是国家社科基金项目主要类别之一，旨在鼓励广大人文社会科学工作者潜心治学，扎实研究，多出优秀成果，进一步发挥国家社科基金在繁荣发展哲学社会科学中的示范引导作用。后期资助项目主要资助已基本完成且尚未出版的人文社会科学基础研究的优秀学术成果，以资助学术专著为主，也资助少量学术价值较高的资料汇编和学术含量较高的工具书。为扩大后期资助项目的学术影响，促进成果转化，全国哲学社会科学规划办公室按照“统一设计、统一标识、统一版式、形成系列”的总体要求，组织出版国家社科基金后期资助项目成果。

全国哲学社会科学规划办公室

2014 年 7 月

国家社会科学基金后期资助项目
“数字资源长期保存技术的研究与实践”(09FTQ005)
项目成员及本书作者

张智雄　吴振新　付鸿鹄　刘建华
曲云鹏　王玉菊　刘　振　郭红梅
廖建军　高广尚　王敬东　马　娜
刘家益　李春旺　向　菁　郭家义

序

作为人类的一种重要社会记忆机构，保存人类历史文化遗产是图书馆永恒不变的重要职责。在当今数字时代，迎接数字技术挑战，开展数字保存工作，实现重要数字信息的长期保存是数字图书馆的一个工作重点。

自21世纪以来，国外数字保存的研究和实践取得了长足进步，逐步从理论研究、技术试验、小规模实验向更加深入的数字保存实践方向发展。欧美等主要发达国家以国家战略的方式有机组织数字保存研究和实践的开展，出现了较为成熟的、产品化的数字保存系统，构建起了可靠的数字保存服务，形成了较大规模的数字保存网络（联盟）。

中国也在积极应对数字资源长期保存的挑战，国内一些重要机构已经开始逐步开展数字保存的实验和实践。如中国科学院文献情报中心自2005年以来，一直致力于可信赖数字资源长期保存系统的建设，通过与国内外的重要科技文献出版商合作，目前已经实现了十余个重要电子科技期刊数据库和科技图书数据库的本土化长期保存。而国家图书馆则积极推动中文资源的数字保存，目前已经实现了中文图书、数字家谱、敦煌IDP数据、数字方志、数字善本、电子报纸、网页资源等十余种数字资源的保存工作，资源量达1140TB。2013年，国家科技图书文献中心（NSTL）在国家科技部的支持下，启动了国家数字科技文献资源长期保存体系建设专项，经过两年多的工作，目前已经初步建成了这一体系的两个示范保存节点，即中国科学院文献情报中心节点和中国科学技术信息研究所节点。这些工作对于促进我国数字保存的研究和实践起到了重要推动作用。

然而，与欧美主要发达国家相比，我国数字资源长期保存的研究和实践在广度和深度上都有着较大的差距，这些差距的产生有着多方面的原因。技术因素是导致我国数字保存工作相对落后的一个重要原因。众所周知，数字资源长期保存工作开展必须基于相关技术系统和技术工具来进行。但由于数字资源长期保存所涉及的技术问题复杂，实现数字资源长期保存的技术系统的门槛较高，所需的技术研发投入较大，并且数字资源长期保存技术乃至数字资源长期保存本身的价值不能在一个短期的时间内产生效果，因此，在国内图书馆界，数字资源长期保存技术的研究和实践尚未广泛而深入地展开。技术方法的缺乏，严重制约着我国数字资源长期保存的实践。我国数字资源长期保存实践的开展，迫切需要在数字保存的技术方法上率先取得

突破。

如何有效在数字保存的技术方法上快速取得突破？鉴于国外数字资源长期保存相关技术较为成熟的事实，在引进、吸收和消化的基础上，进一步面向未来进行创新实践是有效提高我国数字保存技术能力和水平的主要途径。

在国家社会科学基金后期资助项目“数字资源长期保存技术的研究与实践”（09FTQ005）的支持下，中国科学院文献情报中心的数字保存技术研究团队以欧美国家数字资源长期保存的技术研究和实践为基础，并结合本单位数字资源长期保存技术研究和实践的实际情况，梳理了当前国际上数字保存研究和实践的基本现状，分别从数字保存的研究内容、数字保存的技术基础、数字保存的功能实体、数字保存的技术专题和数字保存的实践案例五个方面，较为系统地研究分析了数字资源长期保存在技术研究和实践方面所涉及的关键技术问题、这些问题的基本解决思路、主要解决方案、相关的技术工具以及这些技术方法的具体应用实例。全书基本可以反映当前数字资源长期保存技术的主要研究问题和相关研究现状，能够对我国数字资源长期保存的研究和实践有所裨益。

本书是“数字资源长期保存技术的研究与实践”（09FTQ005）项目组众多成员努力的结果，很多章节是多人协作共同完成的。全书的组织策划、章节编排、内容组织、统稿审阅由张智雄完成。具体各章节的作者分别如下：第1章由张智雄完成；第2章由张智雄、郭红梅、刘振完成；第3章由张智雄、郭红梅完成；第4章由张智雄、马娜完成，第5章由吴振新、张智雄、刘建华完成；第6章由吴振新、张智雄、马娜完成；第7章由张智雄、付鸿鹄、曲云鹏完成；第8章由曲云鹏、张智雄完成；第9章由张智雄、吴振新、曲云鹏完成；第10章由廖建军、张智雄完成；第11章由廖建军、张智雄完成；第12章由张智雄、吴振新完成；第13章由刘振、张智雄完成；第14章由刘建华、张智雄完成；第15章由刘振、张智雄、郭家义完成；第16章由刘家益、张智雄完成；第17章由曲云鹏、张智雄、马娜完成；第18章由高广尚、张智雄完成；第19章由高广尚、张智雄完成；第20章由王敬东、张智雄完成；第21章由吴振新、张智雄、付鸿鹄、王玉菊完成。除了张智雄之外，郭红梅、王敬东、马娜承担了较多的编辑校对工作。

感谢国家社会科学基金后期资助项目“数字资源长期保存技术的研究与实践”（09FTQ005）支持本书的完善和出版，感谢项目评审专家对本书书稿所提出的中肯意见和建议。感谢北京大学李广建教授、南开大学柯平教授、中国科学院文献情报中心孙坦研究员、清华大学姜爱蓉研究员、国家图书馆申晓娟研究员对本书书稿所提出的很多建设性意见。感谢张晓林馆长和中

国科学院文献情报中心对本项目团队长期一贯的支持。国家图书馆出版社的王欢、金丽萍等老师为本书的出版付出了辛勤劳动，在此一并致谢。国内外的数字资源长期保存研究为本书的写作提供了坚实的知识基础，书中参考了很多学者的相关研究成果，都在相应章节的参考文献中进行了标注说明，在此也对这些参考文献的作者表示感谢。

由于数字资源长期保存技术所涉及的问题复杂、内容众多，要系统深入地对数字资源长期保存相关的关键技术问题进行有见地的研究剖析是相当困难的一件事件，虽然本书作者为之倾注了大量的辛勤劳动，但由于作者的能力和水平有限，书中不妥和疏漏之处还在所难免，恳请各位专家学者不吝赐教。

张智雄

2015 年 7 月于中国科学院文献情报中心

目　录

绪　言 …… (1)

第一篇　数字保存的研究内容

1　数字保存的基本概念 …… (7)
2　数字保存的主要研究内容 …… (30)
3　数字保存研究和实践的主要发展历程 …… (50)

第二篇　数字保存的技术基础研究

4　数字保存的技术体系研究 …… (83)
5　数字保存的技术策略研究 …… (112)
6　数字保存的信息模型研究 …… (131)

第三篇　数字保存的功能实体研究

7　摄入功能实体研究 …… (159)
8　档案存储功能实体研究 …… (199)
9　数据管理功能实体研究 …… (231)
10　行政管理功能实体研究 …… (255)
11　保存规划功能实体研究 …… (275)
12　访问功能实体研究 …… (304)

第四篇　数字保存的技术专题研究

13　数字保存的格式管理 …… (333)
14　数字保存元数据体系 …… (365)
15　数字保存的标准体系 …… (408)
16　数字保存的成本问题 …… (425)
17　数字保存系统研究 …… (445)

18 仿真的保存技术方法 …… (477)
19 迁移的保存技术方法 …… (494)

第五篇 数字保存的实践案例研究

20 LOCKSS 系统的实践案例研究 …… (513)
21 一个数字保存系统建设的实践案例研究 …… (539)

绪　言

1. 数字保存是数字时代社会记忆机构的一项重要职责

数字化、网络化、信息化改变着人们的生存方式，数字信息已经成为我们生活工作中接收处理、交流应用、组织管理的主要信息。尼葛洛庞帝在1996年描述的“数字化生存”已经成为当今我们这个时代的重要特征，我们工作生活的方方面面都已经离不开数字信息。

然而，与传统的文献信息相比，数字信息的生存能力却极其脆弱。记载于甲骨、金属、竹木、石头、丝绸、纸等介质之上的信息，如果没有人为的故意毁坏，可能存活上千年还仍然能够被后人所理解。但以比特为单位进行存储的数字信息，如果得不到有效的维护和管理，很容易被盗取、篡改和破坏，或因数字技术的飞速进步、存储设备的过时、读取设备的淘汰而变得不能被读取、被理解和被应用。数字时代，我们业已离不开的数字信息面临着严重的生存挑战。记载于比特位上的任何数字信息随时都面临着来自技术、经济、组织、人为因素、自然灾难等各方面的威胁。这些威胁对数字信息的可存活性、真实完整性和可理解性形成了重要挑战。

在这种情况下，数字资源长期保存成为对重要数字信息进行维护管理，确保重要数字信息资源可以长期存活下去的一种解决方案。在本书作者看来，数字资源长期保存是一系列对数字信息进行持续管理和维护的活动，其目标是为了确保数字信息长期存活，保证数字信息真实可信，能够被未来的使用者所理解和应用。需要明确指出的是，在本书中“数字资源长期保存”与“数字保存”二者表示的是同一个概念。实际上，在本书的各个章节，为了与国外通行的“Digital Preservation”和“Digital Curation”两个术语相对应，“数字资源长期保存”也被称为“数字保存”。

数字资源长期保存已经成为当前国内外图书馆、档案馆和博物馆越来

越关注的一个重要问题。一些有志之士还提出数字资源长期保存是当前纷繁复杂的数字时代下,图书馆、档案馆和博物馆等记忆机构承担历史文件遗产保存、维护数字社会真实可信的一项重要社会职责。

2. 本书的主要目的

自20世纪90年代中期,随着“美国记忆”等项目的开展,国外数字资源的长期保存逐步开始成为一个研究领域。经过近20年的飞速发展,国外的数字资源长期保存研究已然成为当前数字图书馆研究中的一门“显学”。特别是近年来,随着理论的不断成熟和实践的不断丰富,国外数字资源长期保存的研究和建设项目越来越多、研究越来越深入、成果越来越显著,出现了数字资源长期保存研究和实践产品化、规模化、标准规范化的趋势。例如,出现了很多成熟的、产品化的数字保存系统和服务;欧洲、美国等主要发达国家以国家战略的方式组织数字资源长期保存实践的开展;而可信赖数字仓储审计和认证规范已经成熟,很多前期建设的数字保存仓储都纷纷根据标准进行可信赖数字仓储的认证。

而在国内,数字资源长期保存的研究和实践还处于较为初级的阶段。数字资源长期保存的意识尚未在国内得到高度关注,数字资源长期保存的国家战略尚未出台,数字资源长期保存尚未在法律政策层面得到支持,也鲜有大规模的数字资源长期保存研究和实践项目。当然也尚无对数字资源长期保存的理论、技术、方法、实践等较为深入的研究。

本书的作者认为数字资源长期保存实践的开展必须要面对一系列技术上的困难和挑战。而从数字资源长期保存工作开展的情况来看,技术因素已经严重制约了我国数字资源长期保存工作的开展。因而有必要对数字资源长期保存的技术问题和技术实践问题进行深入的研究。

自2005年以来,在中国科学院、国家科技图书文献中心(NSTL)的支持下,本书的作者团队开展了一些数字资源长期保存相关的技术、系统、实践方面的研究,在2008年初步形成了名为“数字资源长期保存技术”的研究和实践报告。

2009年,本书的作者团队得到了国家社会科学基金项目“数字资源长期保存技术的研究与实践”项目(09FTQ005)的资助,希望在前期研究成果的基础之上进行系统化的提炼和升华,深入对数字资源长期保存的各个关键技术进行了研究分析,能够对国内数字资源长期保存技术实践的开展提

供支持和参考。

本书的作者团队在原有申请材料的基础上，对前期研究成果的主要内容进行大幅的调整和优化，结合当前数字保存的研究和实践现状，补充了大量资料，并从整体上对相关内容进行了重新组织和深化。形成了当前五篇21章的主要结构。希望能够反映当前数字资源长期保存技术的主要研究问题和相关研究现状，能够对我国数字资源长期保存的技术研究和技术实践提供有益参考。

3. 本书的基本结构

本书共21章，分为五篇来组织，分别为数字保存的研究内容、数字保存的技术基础研究、数字保存的功能实体研究、数字保存的技术专题研究和数字保存的实践案例研究。

第一篇是“数字保存的研究内容”，共分三章，分别为“第1章　数字保存的基本概念”“第2章　数字保存的主要研究内容”和“第3章　数字保存研究和实践的主要发展历程”。在这三章中，研究了数字保存的基本概念，论述了数字保存的主要研究内容，回顾了数字保存研究和实践的主要发展历程，并重点对当前国际上数字保存研究和实践的现状进行了总结分析。

第二篇是“数字保存的技术基础研究”，包括三章内容，分别为“第4章　数字保存的技术体系研究”“第5章　数字保存的技术策略研究”和“第6章　数字保存的信息模型研究”。在这三章中，作者研究分析了数字保存技术研究和实践中三个基础性问题，分别为数字保存的技术体系问题、数字保存的技术策略问题和数字保存的信息模型问题，从一个较高的角度分析在数字保存的技术实践过程中，需要关注哪些应用技术，可以采取什么技术策略，如何有效进行数字信息的组织管理。

第三篇是“数字保存的功能实体研究”。在这篇内容中，作者基于“开放存档信息系统参考模型”（OAIS）的功能框架，研究分析了开放存档信息系统的六个功能实体，分别为摄入功能实体、档案存储功能实体、数据管理功能实体、行政管理功能实体、保存规划功能实体和访问功能实体。每章研究分析一个功能实体，共六章，具体为第7章到第12章。各章分别研究了六个功能实体的主要技术功能要求、主要流程规范、关键技术方法、相关系统工具等问题。

第四篇是“数字保存的技术专题研究”。在这篇内容中，作者从专题研究的角度出发，对与数字保存的技术实践密切相关的七个专题进行了较为深入的专题研究。这七个专题分别为数字保存的格式管理、数字保存元数据体系、数字保存的标准体系、数字保存的成本问题、数字保存系统、仿真的保存技术方法、迁移的保存技术方法。七个专题，每个专题自成一章，具体为第 13 章到第 19 章。

第五篇是“数字保存的实践案例研究”。这一篇希望通过具体实践案例的研究，来分析在数字保存中的技术实践需要解决哪些技术实践问题和如何解决这些实践问题。这一篇内容包括第 20、21 章。作者在这两章中，分别以 LOCKSS 保存系统和作者单位数字保存系统建设的实践为案例，研究分析了数字保存系统在实际的建设实现中需要关注的重要问题。

全书较为系统地研究分析了数字资源长期保存技术的相关问题，较为全面地反映了当前数字保存技术的研究进展，希望能够对我国数字资源长期保存技术的研究和实践起到推动作用。

（张智雄）

第一篇 数字保存的研究内容

本篇包括以下内容：

1 数字保存的基本概念

2 数字保存的主要研究内容

3 数字保存研究和实践的主要发展历程

1　数字保存的基本概念

数字技术改变了人类的生活方式，同时也带来了新的挑战。由数字技术支撑的数字信息，与传统的文献信息相比，在信息内容的承载、传输和持久保存方面存在着一系列与生俱来的问题。数字技术的快速变革、数字信息的多种依赖性，导致数字信息即便为了存活一个年代，也需要得到特别的关注、维护和管理。数字保存与传统文献信息的保存在思路方法、技术活动等方面都有着重大差别。本书作者认为，数字保存是一系列对数字信息进行持续管理和维护的活动，其目标是为了确保数字信息长期存活，保证数字信息真实可信，能够被未来的使用者所理解和应用。

本章讲述数字保存的基本概念，在分析了数字技术给现代社会带来的问题和挑战，以及当前社会中数字信息所面临的各种威胁之后，论述了数字保存的概念、数字保存要达到的主要目标、数字保存的意义等内容。

1.1　数字技术带来的问题与挑战

数字化生存是当前我们所处时代的一个重要特性。当前，我们日常生活的各个方面，如文献阅读、信息查找、工作学习、科学研究、文化娱乐、消费购物、社会交往等都已离不开数字技术。然而数字技术却是一把双刃剑，它在给我们带来便利的同时，也带来了数字信息的保存问题。

1.1.1　数字技术：一把双刃剑

数字技术为信息的传播和应用带来了前所未有的便利。基于数字技术的信息与基于印刷技术的信息相比，在信息的获取、复制、传输、存储、携带等各个方面都有了质的飞跃。数字技术让有价值的信息可以被快速方便地复制，让信息可以跨越时空进行传输，让海量信息（如成千上万本图书）可以轻松地携带在身边。数字技术让信息的传播和应用“告别铅与