本书介绍了数字资源长期保存的理论和关键技术

董晓莉 著

# 数字资源长期保存关键技术研究

中国书籍出版社
China Book Press

本书介绍了数字资源长期保存的理论和

董晓莉 著

# 数字资源长期保存关键技术研究

中国书籍出版社
China Book Press

**图书在版编目（CIP）数据**

数字资源长期保存关键技术研究 / 董晓莉著. -- 北京 : 中国书籍出版社, 2013.8
ISBN 978-7-5068-3604-3

Ⅰ. ①数… Ⅱ. ①董… 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（2013）第164803号

**数字资源长期保存关键技术研究**

董晓莉 著

**策划编辑** 安玉霞
**责任编辑** 安玉霞
**责任印制** 孙马飞 张智勇
**封面设计** 展 华
**出版发行** 中国书籍出版社
**地 址** 北京市丰台区三路居路97号（邮编：100073）
**电 话** (010)52257143总编室）(010)52257153（发行部）
**电子邮箱** chinabp@vip.sina.com
**经 销** 全国新华书店
**印 刷** 北京京海印刷厂
**开 本** 880毫米×1230 毫米 1/32
**印 张** 6.5
**字 数** 170千字
**版 次** 2013年10月第1版 2013年10月第1次印刷
**书 号** ISBN 978-7-5068-3604-3
**定 价** 25.00元

# 序言

人类进入信息社会，信息资源已经成为重要的战略资源。随着知识经济时代的到来，人类阅读形式的变化，数字出版业的蓬勃发展，数字信息资源在社会生产中的地位越来越重要。然而，数字资源有其特殊性，一旦销毁或者更改，将无法恢复，信息的传递也将无法继续下去。数字信息资源的保存是对信息时代提出的新的挑战，数字资源长期保存的重要性和紧迫性更显突出。

在当今的数字化环境中，长期保存数字资源是保存文化遗产的迫切要求，是一项重大的历史责任。为了我国的数字化发展战略，各保存机构、研究机构和信息机构应结为最广泛的联盟，开展最深入的合作研究，最大限度地保存有价值的数字信息资源，为人类文明历史的传承做出自己的贡献。

按照《现代汉语词典》的解释，保存是“使事物、性质、意义、作用等继续存在，不受损失或不发生变化”。英国国家图书馆、英国国家人文艺术数字资源服务中心、美国研究图书馆组织RLG和联机计算机图书馆OCLC给出了几组定义，虽然表述上有一定的差异，但其内涵都趋于一致，基本都认为“数字资源长期保存”包含两层意思，即长期存储和长期可获取。

联合国教科文组织提出的《数字遗产保护章程》中将数字遗产定义为：数字遗产是特有的人类知识及表达方式，它包含文化、教育、科学、管理信息、技术、法律、医学以及其他以数字形式生成的信息，或从现有的类似的模式转换成数字形式的信息。具

体形式包括文本、数据库、静止及动画影像、音带、照片、软件、网页等；其价值主要通过数字信息来表现，主要有存贮于特定载体（如光盘、磁盘、DV 带等）的信息资源、存贮于计算机数据库中的信息资源、以及通过网络传播的资源信息[1]。此外，以保护文化遗产为己任的联合国教科文组织于 2001 年制定了《数字遗产保存草案》，2003 年制定了《数字遗产保存指南》，为数字遗产的长期保存提供指导。由此可见，数字资源的长期保存和保护已成为保存人类文化遗产的重要部分，并逐步发展成为一个新的研究领域。

数字资源长期保存是通过对数字信息生命周期的管理，实现对数字对象的长期可生存能力、可呈现能力和可理解能力的维护，但由于其涉及诸多方面，不仅包括了设计机构、技术、方法、设备等方面的技术性问题，还涉及到很多非技术性问题，如费用问题、版权问题、安全性等问题，目前我国还没有建立起一个真实可信的数字资源长期保存体系。

本书介绍了数字资源长期保存的理论和关键技术，旨在为读者提供将数字资源长期保存应用于实际问题所必需的知识。本书涵盖五个主题：长期保存体系、软件系统架构、存储系统、技术策略、可信性，从概念、模型、系统实现、工作流程等方面深入浅出地进行了论述和分析。目的是使读者在透彻地理解数字资源长期保存基础的同时，还能了解更多重要的高级主题。

本书首先从数字资源长期保存的概念入手，分析了数字资源长期保存的体系构成，进而对数字资源长期保存系统的实现进行了深入探讨。结合我国数字资源长期保存的现状和存储技术的最新进展，对底层存储系统模型进行了多角度、多层次的深入分析，

1 联合国教科文组织．数字遗产保护章程 (2002). http://www.saac.gov.cn

提出基于战略联盟保存的分布式存储系统的多种实现方式。数字资源长期保存过程涉及的相关技术分布在计算机、网络、通信、标准化等多个领域，本书对数字资源长期保存关键技术策略进行了介绍和分析，并分析了每种策略的可行性和适用性。目前，迁移技术已被认为是一种有效的长期保存策略，笔者以图书馆为例，详细分析了数字资源长期保存迁移技术，对迁移过程的风险、迁移对象模型，迁移模型、迁移流程和目前主流迁移工具进行了深入的探讨。可信性是数字资源长期保存实践的核心和基础保障，笔者从技术角度着重分析了数字资源长期保存可信性的流程和模型，并根据流程推出具体的实施方法。最后，以图书馆数字图像长期保存格式为例，对不同常用长期保存格式进行对比分析，提出未来数字图像保存格式的发展趋势，并通过格式迁移测试，对长期保存格式迁移的风险进行分析。

本书力图在以下几个方面有所突破：

第一、可靠的存储环境是开展数字资源长期保存实践的核心和基础保障，目前尚无成熟、完整的长期保存存储体系架构可供参考，仍属于探索阶段。本书从系统出发综合把握，通过对“数字资源长期保存存储系统”的需求、设计原则等方面的论述，在此基础上对“数字资源长期保存存储系统模型”进行了层次化的分析和设计。

第二、迁移技术已被认为是一种有效的长期保存策略。但由于其涉及诸多方面，除迁移策略选择、迁移实施、迁移途径设计等技术性问题外，还涉及到很多非技术性问题，如费用问题、版权问题、安全性等问题，目前尚无大规模的实施案例。笔者在介绍数字资源长期保存迁移概念的基础上，通过对迁移过程的风险和迁移对象模型进行分析，进而提出迁移模型和迁移流程，并对

目前主流迁移工具进行了分析。

第三、数字资源长期保存的可信性是开展数字资源长期保存实践的核心和基础保障，目前的研究主要是基于数字仓储系统的可信性研究，尚未建立一套基于长期保存体系的可信保障机制。本书在介绍数字资源长期保存可信性基本概念的基础上，从技术角度着重分析了数字资源长期保存可信性的流程和模型，并根据流程推出具体的实施方法。

第四、图像资源是数字图书馆数字资源的主要组成部分，因此对于需要长期保存的图像资源来说，确定合适的长期保存格式和压缩算法是开展长期保存工作前必须制定的保存目标之一。图像资源长期保存格式的确定涉及诸多方面，除需考虑开放性、保真度、无损压缩、独立性等格式自身特性外，还涉及到原始格式的迁移更新等问题以及一些非技术性问题，如费用、版权、安全性等问题，目前尚无统一、明确的保存格式约定。本书在介绍数字图像资源长期保存主流格式概念的基础上，通过对不同格式的对比分析，提出未来数字图像保存格式的发展趋势，并通过格式迁移测试，对长期保存格式迁移的风险进行分析。最后，提出风险控制的方法。

本书具有一定的创造性、实践性和广泛性，相信“数字资源长期保存关键技术研究”将为从事数字资源长期保存工作的人员提供理论和经验参考；同时也可使社会公众全面了解数字资源长期保存领域的技术特点。本书不仅具有重要的实践指导意义，也具有良好的社会效益。

由于笔者能力有限，书中难免存在不足之处，衷心希望读者批评指正。

# 目录

# 1. 为什么要长期保存

步入信息时代，信息资源已成为影响国家安全和长远发展的重要资源。信息资源长久的可持续性存取对未来学术研究、教育、文化、企业发展是必不可少的。人类文明的传承将依赖于大量记载着人类文化遗产的数字资源，数字信息资源的长期保存已成为信息时代必须面临的新的挑战，数字资源长期保存的重要性和紧迫性更显突出。

## 1.1 一个新的研究领域

数字资源长期保存是近年来图书馆界和档案界讨论与研究的热门话题，并已发展成为一个新的研究领域。英国不列颠图书馆（The British Library）琳内·布林德丽（Lynne Brindley）馆长在2005年就做过这样的预测：到2020年，英国的研究著作40%将以电子形式出版，50%将以印刷和电子两种形式同时出版，仅有10%的出版物纯粹以印刷形式出版[1]。由此可见，随着信息时代的来临，印刷媒体占据市场份额的下降和电子媒体所占比重的大幅上升，使得数字资源的长期保存工作成为保存人类文化遗产的重要部分。

国内外已出现多个将珍贵历史、文献等资源数字化并提供使用的项目。最著名的是美国国会图书馆（Library of Congress, LC）于1990—1995年间实施的试验性计划“美国记忆（American Memory）”项目。该计划的数字馆藏对象主要为美国的历史文献，包括历史照片、手稿、历史档案及其他文献等。2012年，在进行了初步调研、酝酿和策划的基础上，国家图书馆也启动了“中国记忆”项目，以实现记录历史、保存文献、传承民族记忆、服务终身学习的目的。

然而，与传统资源相比，数字资源有着很多独有的特性如数字资源是二进制位流，必须存放在某种载体上，“保质期”短；具有其自身的格式，需要有理解这种格式的软件及能运行这些软件的硬件，才能解析和呈现；复制成本非常低廉；更新频度高；易受到来自网络的侵害；标引和著录要求往往比传统资源更为复杂；检索和利用相比传统资源极为方便、便利和高效等特性[2]。数

1 Lynne_Brindley. http://en.wikipedia.org/wiki/Lynne_Brindley

2 孟广均. 国内外图书馆学与情报学最新理论与实践研究[M]. 北京：科学出版社，2009.

字资源的这些固有特性造成了数字资源易于改变、容易消失。就物理存储方面而言，数字资源及其所依赖的存储载体非常不稳定，容易受环境影响而损坏，从而使信息本身无法读取甚至消失；信息技术的快速变化使得硬件和软件都在不可预测地老化；在使用方面，多数情况下信息机构仅购买了数字资源的使用权，而由于多种原因，数据库商或其委托的镜像服务商有可能终止经营或者停止向用户提供服务；此外，数字信息容易被未经授权操作而改变甚至删除；新网页的出现实际上就意味着旧网页的消失和大量信息的消失。所有这些特性都显示了数字资源的脆弱性，而数字资源一旦无法使用将会带来很多问题，使人类文明无法延续，因此如不采取有效的长期保存措施，很多数字形式存在的文化遗产将永远消失。鉴于此，以保护文化遗产为己任的联合国教科文组织提出“数字遗产”的概念，并于 2001 年制定了《数字遗产保存草案》，2003 年又制定了《数字遗产保存指南》，为数字遗产的长期保存提供指导。

由此可见，数字资源长期保存是近年来社会各界讨论与研究的热门话题，并已发展成为一个新的研究领域。数字资源长期保存关系国家的文献保护和文化传承，做好数字资源长期保存工作具有重大而深远的意义。

## 1.2 长期保存的背景

随着信息资源数量的不断增加，各国对信息资源的依赖性不断增强，数字信息已经成为当今世界上最重要的信息之一，随之而来的是对数字信息和数字遗产 (Digital Heritage) 的保护问题。

联合国教科文组织提出的《数字遗产保护章程》中将数字遗

产定义为：数字遗产是特有的人类知识及表达方式，它包含文化、教育、科学、管理信息、技术、法律、医学以及其他以数字形式生成的信息，或从现有的类似的模式转换成数字形式的信息。具体形式包括文本、数据库、静止及动画影像、音带、照片、软件、网页等；其价值主要通过数字信息来表现，主要有存贮于特定载体（如光盘、磁盘、DV 带等）的信息资源、存贮于计算机数据库中的信息资源、以及通过网络传播的资源信息[3]。在武汉大学研究项目“我国情报学学科建设、发展与前瞻性研究（70373038）”的阶段性成果《情报学前沿领域的确定与讨论》[4]中明确提出，数字资源长期保存是目前情报学专业的前沿领域之一，在学科建设中发挥着越来越重要的作用，这一方向已经成为以后重点发展领域[5]。

1991 年五个北欧国家的国家档案馆（瑞士、挪威、丹麦、芬兰和冰岛）就电子文件保护与存取问题进行了调研，并最终颁布《电子文件的存取与保护》研究报告。1994 年 12 月美国保护与存取委员会（Commission on Preservation and Access, CPA）与美国研究图书馆（Research Library Group, RLG）共同组建了数字存档特别工作组。1996 年，该工作组完成了《保护数字信息：数字信息归档特别工作组报告》（Preservation Digital Information: Report of the Task Force on Archiving of Digital Information）。该报告对数字资源长期保存中存在的问题做了比较全面的分析，并提出了建设性意见，引起了世界有关国家政府的重视。

1996 年，澳大利亚图书馆制定了“澳大利亚电子出版物的国

3 联合国教科文组织．数字遗产保护章程 (2002). http://www.saac.gov.cn

4 赖茂生等．情报学前沿领域的确定与讨论 [J]．图书情报工作．2008(3):15-18

5 http://blog.xmulib.org/2007sz/Wuzx.pdf