

云南省教育厅科学研究基金项目 (06Y350G)

数字信息资源 长期保存研究

张绍武 著

SHUZI XINXI ZIYUAN CHANGQI BAOCUN YANJIU



云南出版集团公司
云南科技出版社

云南省教育厅科学研究基金项目 (06Y350G)

SHUZI XINXI ZIYUAN CHANGQI BAOCUN YANJIU

数字信息资源长期保存研究

张绍武 著

云南出版集团公司

云南科技出版社

· 昆 明 ·

图书在版编目 (C I P) 数据

数字信息资源长期保存研究/张绍武著. —昆明: 云南科技出版社, 2009. 6

ISBN 978 - 7 - 5416 - 3304 - 1

I. 数… II. 张… III. 信息管理—研究 IV. G203

中国版本图书馆 CIP 数据核字(2009)第 100425 号

云南出版集团公司

云南科技出版社出版发行

(昆明市环城西路 609 号云南新闻出版大楼 邮政编码:650034)

昆明市五华区教育委员会印刷厂印刷 全国新华书店经销

开本: 889mm × 1194mm 1/32 印张: 8.25 字数: 200 千字

2009 年 7 月第 1 版 2009 年 7 月第 1 次印刷

定价: 35.00 元

序 言

一次偶然的机会读到了张绍武先生的文章——关于数字信息资源长期保存的文章。因为是首次接触，初读时不以为意，但越读越为文所吸引，一气读完可谓是印象深刻。与张先生之交往，实为先见其文，后会其人，人文相照，如影随形。之后又读了不少张先生的文章，虽彼此相会不多，却知之不少。其对学问、学术研究之执着，著文之认真、扎实，做人之真诚，甚为钦佩。可以说，多年来张先生致力于“数字信息长期保存”的研究，孜孜不倦、义无反顾，今之能以“云南省教育厅科学研究基金项目”、以专著的科研成果形式收获果实，也可说是水到渠成。

数字信息资源长期保存问题，是网络技术迅速发展应用、信息资源数字化及其社会化应用之必然。

图书馆在其自身发展的过程中，以保存利用文献信息资源为其基本职能。但随着文献资料载体形式的不断变化，从远古时代的甲骨、泥版、竹木简、帛等，到纸质印刷品、到当今的胶片、磁带、光盘，直至数字化信息资源等，这其中尤以纸质文献流行最为漫长、广泛，而以数字资源来势最为迅猛。对于纸质文献，人类所付与的保存模式，可以说是已臻完善。而有史以来增长速度更快、含量更大、更复杂的数字信息资源的保存，却给了我们这个现代化的信息社会，特别是收集及提供文献信息利用的直接工作者以巨大的冲击与挑战，可以说已到了时不我待之当务之急。

数字信息资源长期保存涉及的问题繁多，诸如理论、技术、资源、法律、管理及机构、设备、人员等等。对于这一命题，国外业界已开展了近20年的广泛研究，取得了相当的成果与进展。

而国内图书情报界，从初始提出至今，不过十余年，从事研究的专家学者也十分有限，要达到系统化的认知，科学化的操作、实施、管理还有相当大的差距，特别是要做到从观念到社会化的管理实施相一致，还需要付出极大的努力。

在数字信息资源长期保存研究领域，张绍武先生 2000 年前夕便开始关注并倾力研究，张先生不断地分析国内外大量的有关文献资料、最新学术动态及实践活动，于 2000 年 6 月初发表《数字信息长期保存的策略取向》，至今已发表本专题论文近十篇。与此同时，张先生对数字信息资源长期保存的诸多因素、问题进行系统化、全面性的归纳、总结，因此于 2006 年申报获准云南省教育厅科学研究基金项目——《数字信息长期保存研究》。多年来的辛勤研究与笔耕，今终得硕果面世，实为可喜。

该著作全面、系统、逻辑、条理地阐述了本专题的诸多主要问题，如数字信息资源的产生与发展、长期保存总框架、长期保存模型、长期保存的先决条件等共九个方面的问题。本书的著述过程，作者力图贯穿知识性、系统性、新颖性于始终，其中以系统性尤为突出与可贵。作者把学术界长期分散的研究作了系统化、条理性地归纳，从而使“数字信息资源的长期保存”理性地组织成为一个有机的整体。

当前我国对于本专题的研究，学术专著成果还为数甚少，而今出版此一力作，实为难能可贵。它为我们这一领域的研究及实践工作的开展作出了应有的贡献，同时也填补了云南省本专题研究领域的空白。

衷心祝愿《数字信息资源长期保存研究》对图书馆的信息资源建设、对我国的信息化建设发挥出积极的作用。

李高遠

前 言

自 1993 年美国提出建设信息高速公路以来, Internet 在世界范围迅速普及, 这为数字图书馆的发展提供了重要的基础。20 世纪 90 年代后期, 我国建设数字图书馆的网络基础设施已初具规模, 国内数字图书馆项目的建设与发展, 推动了数字图书馆的研究。十多年来, 我国数字图书馆的研究主要集中在数字图书馆的含义、特征、组织结构与功能, 以及建设数字图书馆的对策和构想等方面。随着数字技术的发展, 人们又进一步对数字图书馆所涉及的问题与技术加以讨论, 其主要有信息资源的数字化、存贮压缩、组织、管理、检索、服务、标准化与知识产权等等。

就在数字图书馆的信息资源建设、存贮与利用成为我国图书情报界研究热点的同时, 数字图书馆的另一个重要职能——数字信息的长期保存, 也开始引起了我国图书情报界有识之士的关注与研究。

我国数字信息长期保存问题最先由李人厚、苗凌两人提出。他们在 1997 年《大学图书馆学报》第六期上发表文章, 文章主要论述了数字图书馆的定义与特征、面临的问题及对策, 其中就谈到: “为了建立可有效利用的数字图书馆, 必须开发一种可长期保存数字化内容的新方法。这种保存方法应该使被保存的内容可经受几代计算机硬件和软件的变化, 也可经受标准的不断变化, 这是十分困难的工作, 但应引起重视”。李、苗二位谈到的这种新方法虽然当时还做不到, 但他们毕竟提出了应该重视数字信息长期保存的问题。汪冰先生在 1998 年《中国图书馆学报》第六期上发文也指出: “图书馆员对于存储和长期保存印刷资料已经建立了一整套方法和操作规范。而数字化信息的内容和位置

都很容易发生变化。在实践中,数字媒介还没有进入图书馆员归档保存的议事日程,更没有为管理这些档案真正地建立恰当的组织程序。因此,如何保存数字化信息以便保存和延续文化,对于熟悉纸本的图书馆员是一个巨大的挑战。”虽然李人厚、苗凌和汪冰提出了数字信息的保存问题,但我们看到,他们只是在谈到数字图书馆建设时顺便提到了它。就当时的条件来看,他们还没有也不可能对数字信息长期保存作具体的论述。而真正对数字信息资源长期保存形成专论的还从司莉开始。

1999年3月,武汉大学图书情报学院研究生司莉在《图书馆杂志》(1999年第3期)上,率先发表了数字信息长期保存的专稿——《数字信息保存面临的挑战与对策》。时隔一月,刘家真教授也发表了《保护数字信息长期存取研究综述》(《情报杂志》1999年第4期)。这两篇文章虽然仅仅只是介绍了国外数字信息长期保存的情况。但其真正的意义还在于,它对图书馆人是一种感召,是一种启示;同时,它标志着我国数字信息资源长期保存研究从此开始。

接着,2000年6月,笔者不揣浅陋,也在《图书情报工作》上发表拙文《数字信息长期保存的策略取向》。随后,笔者一方面陆续发表了近十篇这方面的文章,另一方面则更多地关注国内外数字信息长期保存的研究状况和发展趋势,并搜集了大量的资料。

十年来,我国数字信息资源长期保存越来越受到重视,研究文章、报告和网上资料累计多达200多份。国外的资料更是不计其数。面对国内外众多的文献资料,笔者早在2006年就萌生出要写一部专著的念头。目的就在于要使学术界分散的研究形成系统的理论。2006年11月,我申报的“数字信息长期保存研究”项目被立为云南省教育厅科研项目。本书是该项目的主要预期成果。之后,笔者在借鉴国内外研究成果的基础上,经过两年多的分析、梳理、归纳,终于写成了这本书。

数字信息资源长期保存涉及的领域和范围十分广泛，本书的研究不可能面面俱到。笔者仅从图书情报工作的角度来探讨数字信息资源的长期保存，意在能抛砖引玉，为我国数字信息资源长期保存提供参考。

本书力图做到以下几点：

1. 知识性。本书以介绍数字信息资源长期保存的知识为主，大凡与本选题相关的知识均作为研究对象。

2. 系统性。从保存体系的大框架入手，详细地论述数字信息资源长期保存所涉及的方方面面的问题，以期使数字信息资源长期保存研究形成一个系统理论。

3. 新颖性。在论述每一个问题时，尽量采用最新的研究资料和发展动态来探讨。尤其是在本书的最后一章，能预示出数字信息资源长期保存研究的发展趋向。

由于笔者水平有限，书中有不妥之处，恳请专家学者和读者批评指教。

张绍武

2009年6月1日

1

第六章 数字信息资源长期保存的技术体系	(133)
第一节 国外数字信息资源长期保存的技术体系	(133)
第二节 国内数字信息资源长期保存的技术体系	(137)
第三节 数字信息资源长期保存的技术策略	(145)
第四节 保存技术方案的选择与评价	(152)
第七章 数字信息资源长期保存的管理机制	(158)
第一节 管理机制的构建原则与内容结构	(158)
第二节 管理机制的权益关系结构	(161)
第三节 数字信息资源长期保存的责任体系	(164)
第四节 管理机制的 implements 措施	(174)
第八章 数字信息资源长期保存的知识产权保护	(191)
第一节 数字信息资源长期保存的知识产权法	(192)
第二节 数字信息资源长期保存涉及的知识产权问题	(199)
第三节 国外数字信息资源长期保存的知识产权保护 及对我国的启示	(211)
第九章 数字信息资源长期保存研究的发展趋向	(221)
第一节 开放存取资源的长期保存	(221)
第二节 未来长期保存的研究趋向	(228)
结束语	(235)
附录 图表索引	(236)
参考文献	(237)
后记	(259)

第一章 数字信息资源长期保存 的产生与发展

数字信息资源长期保存是随着数字信息资源的产生而产生的，因而在论述数字信息资源长期保存之前，有必要对数字信息资源的产生、分类、特点等作一番介绍，以为后面数字信息资源长期保存的研究作铺垫。

第一节 数字信息资源概述

数字信息资源，狭义地讲，可称电子信息资源，也称数字资源。它指一切以数字形式生产和发行的信息资源。所谓数字形式，是以能被计算机识别的，按不同序列的“0”和“1”构成的形式。

数字信息资源就是以数字形式生产和发行的能被计算机识别的可以在网络上交流的信息资源。数字信息资源中的信息，包括文字、图片、声音、动态图像等，都是以数字代码方式存储在磁带、磁盘、光盘等介质上，以电信号、光信号的形式传输，并通过计算机、通讯设备及其他外部设备再现出来，再通过计算机输出设备和网络传送出去最终显示在用户的计算机终端上。^[1]因此，我们可以将数字信息资源界定为：可以通过计算机网络利用的各种信息资源的总和。

一、数字信息资源的起源与发展

数字信息资源是信息资源的重要组成部分，是电子化了的
信息资源。

追溯数字信息资源的起源,应该从 1946 年世界上第一台大型电子计算机——ENIAC 诞生谈起。可以说,数字信息资源是伴随着计算机的诞生而产生,又在计算机技术和现代通信技术的飞速发展中得到了蓬勃发展的。

但最初的计算机只是进行数据计算,并不用于信息处理。最早的数字信息资源应该是 1969 年美国国立医学图书馆用于医学文献检索的 MEDLARS (Medical Literature Analysis and Retrived System) 数据库。这个机编文献目录型数据库可视为数字信息资源的开端。

随后出现了大量的数值型和事实型数据库电子信息。1972 年美国洛克希德公司 (Lockheed) 推出了大型联机检索系统——DIALOG 数据库。同年,美国系统发展公司的 ORBIT 联机服务系统也隆重面世。之后,许多大规模的联机检索服务系统相继出现,如英国的目录检索服务处 (Bibliographic Retrieval Service) 的 BRS 系统和欧洲空间组织 (European Space Agency) 的 ESA—IRS 系统。由于国际联机检索费用高昂,远离数据库生产国本土的国家,除支付数据库使用费外,还要支付高昂的通讯费用,因此,国际联机检索的发展受到一定的制约,普通用户很难望其项背。后来人们开始努力寻求一种存储费用低廉、检索利用方便、更加大众化的电子信息形式。于是,一种新型的电子信息形式——光盘便应运而生。

光盘是 20 世纪 80 年代初出现的一种激光存储技术。由于它具有存储的信息容量大、成本低、体积小、寿命长、检索费用低、可实现随机存取等特点,因此,光盘技术很快被广泛应用于电子信息存贮。80 年代,随着计算机中央处理器速度的提高,外部磁性和光盘等海量存储技术的突破,设备价格的降低和外国信息处理设备性能的大幅度增强,大量的光盘数据库不断涌现,这为电子信息的海量存储和联机检索展现了广阔的发展前景。这一时期一些著名的计算机联机信息服务系统,如美国的 DIA-

LOG、ORBIT 系统、欧洲的 ESA 系统和德国的 STN 系统等，也都向全世界的联机用户提供光盘产品的信息服务。联机服务业和数据库产品得到了进一步的发展，数字信息资源得到了极大的丰富。^[2]

20 世纪 90 年代，人类社会发展到了网络化时代。1993 年美国政府提出了“信息高速公路计划”，并投入巨资兴建以因特网为依托的信息高速公路。随着因特网的迅速发展，网络信息异军突起、大量涌现，并成为人们进行科学研究、商业活动的重要信息源。由于因特网上的所有主机都采用 TCP/IP 协议连接和通信，网络上的各种计算机也都遵循该协议规定的方式进行数据交换，因而因特网已成为各种数据库的总数据库，它堪称是世界上最大的数字信息资源库。因特网的崛起和发展使网络电子资源如潮水一般涌来。目前这个网络已拥有 186 个国家和地区的 5 万多个注册网络，5 000 多万台计算机，2 500 多个数据库，而且正在以每年高于 20% 的速度激增。^[3]

如今因特网上的信息资源种类繁多，主要有网络信息检索系统、联机数据库信息服务系统、图书馆联机馆藏、电子出版物（网络版报纸、期刊、图书）、软件、官方信息如政府文件、档案和法律法规，电子公告、专题讨论、会议文献、广告（如产品展示、服务介绍、联机订购）、艺术作品等。网络信息成为信息的主体，数字信息的发展步入了以网络信息为主体的、多种载体数字信息并存的数字信息时代。

数字信息资源的发展经历了从早期的书目信息，到后来的全文本信息，以至今天的网络和多媒体信息。数字信息资源在处理技术的发展上，经历了从简单的字段与关键词检索、全文检索，直至今天以网络为主流的超文本与多媒体处理，加之计算机硬件环境和网络的变化，快速、方便、有效地获取数字信息也成了现实。经过近半个世纪的发展，数字信息已成为当今人类和社会不可缺少的重要资源和财富。

二、数字信息资源的类型

数字信息实际上可分为以模拟信息存储的数据信息和以数字存储的数字信息。模拟信息是通过广播、电影、电视等以模拟信号方式传递的信息资源；数字信息是通过计算机等以数字信号来传递的信息资源。本书所讲的数字信息资源主要指数字化的信息资源。

数字信息资源种类繁多，形式多样，按不同的标准可分为不同的类型。

1. 按载体划分

根据载体的不同，数字信息资源可分为磁盘（软磁盘、硬磁盘）、磁带、光盘、半导体存储技术和网络信息等数字信息资源（见表1-1）。

表 1-1 数字信息资源的类型

数字信息类型	数据类型	应用领域
数据集	文字数字	科学调查实验、事务或管理数据、著录数据
格式化文本	格式化文字数字	各种出版物、文本文件、网页等
多媒体文件	格式化文字数字、点阵或矢量图像	各种出版物、多媒体文件、网页、课件等
图表	矢量或点阵图像、文字数字	办公文件、商业文件、各种出版物等
设计数据	矢量或点阵图像、文字数字	计算机辅助设计
图像	点阵图像、文字数字	艺术、医学、建筑、历史与档案管理等

续表 1-1

数字信息类型	数据类型	应用领域
音乐与语音	数字音频数据	语音处理、音乐、数字广播、音乐图书馆
影像	数字视频数据	数字录像或电视、影像图书馆、教育培训
地理空间数据	矢量和点阵数据、文字数字数据	地图、地理信息系统、空间测绘等
交互式多媒体出版物	交互插入的图表、图像、影像、文字与数值数据	交互电子出版物、教育培训系统、商业营销系统、电子游戏等

(1) 磁盘。人们常用的磁存储介质主要有磁盘和磁带两种。磁存储介质中又以磁盘用途最广，磁盘是由一面或两面涂有磁性物质的塑料或金属圆盘构成。磁盘又分软磁盘和硬磁盘。人们日常存储文件使用最多的 3.5 寸软磁盘的容量为 1.44M，而采用最新技术制造的软磁盘容量可达 100M 至 230M。硬磁盘又称硬盘，是个人计算机上不可缺少的存储设备，现在的海量硬盘已达到几十 G、几百 G 的容量，而制作成 USB 接口的便携式硬盘容量已达到几个 G。磁盘采用的是非线性存取，它具有存取速度快，寿命长，可重复使用的特点。

(2) 磁带。磁带是在塑料带基的一面涂上能磁化的氧化铁材料制成，通过电磁脉冲在其表面形成许多肉眼看不见的磁化点，最后将信息记录在磁带上。磁带分卷式和盒式两种。磁带携带方便，容易保管，价格便宜，但它采用线性存取，因此存取的速度较慢，一般适用于信息的保存和交换，主要用于文件存储系统和联机检索系统。

(3) 光盘。光盘是一种用激光束记录和读取信息的圆形盘。其结构分为三层：基体、存储层和保护层。光盘按读写方式大致可分为只读型光盘、追记型光盘和可擦写光盘三种类型。

①只读型光盘 (CD - ROM)。存储容量为 650 (M) 兆字节, 相当于 30 万汉字的书 1 000 册。其抗污力强, 运行中不受温度、湿度、尘埃、电磁场等影响。只读型 (Read Only) 光盘主要用于一次写入而无需修改的数据、资料的存储和分发, 使用者不能在光盘上追加数据, 光盘上以凹坑的形式存储大量预先录制好的视频、音频或数字信息。激光视盘、激光唱盘, 以及可用于各类多媒体系统、个人 PC 机标准配置和出版各类电子图书的 CD-ROM 均属于这一类光盘。

②追记型光盘 (WORM 或 CD - R)。它是光盘的第二代产品。其特点是可以根据用户需要任意写入数据。一旦写入, 就不能修改或擦除, 但可多次读取数据。它的存贮密度高于硬盘, 并可以像软盘一样随意取下由用户自己保存而不失密。与只读型光盘的不同之处在于, 它可由用户将数据直接写入光盘, 这就大大简化或消除了母盘制作过程, 并可用与传统的磁记录极为类似的方法处理数据。追记型光盘通常采用在灵敏层上“烧孔”的方法记录信息, 光盘上存储的信息有很高的可靠性。它既可用于存储图像、文件、文献和档案资料, 也可用作计算机的外存储器。因其成本大大高于 CD - ROM, 目前追记型光盘主要应用于军事、金融保险、法律和航空等要求较高的领域。

③可擦写光盘。可擦写光盘分为磁光型可擦写光盘 (CD - MO) 和相变型可擦写光盘 (CD - PC), 它们是光盘的第三代产品。它不仅能写入数据, 而且可像磁盘一样任意修改或擦除数据。可擦写型 (Rewritable) 光盘是更为先进的光盘, 因其有可擦写功能, 故可作为 PC 机后援存储器或网络服务器、工作站上的外部存储器, 也可用于存储文献资料。目前, 商品化的可擦写型光盘根据工作原理分为磁光型和相变型两种。磁光盘 (MO)

是利用激光照射改变磁性记录介质磁畴方向的原理进行信息记录的。相变光盘 (PC) 是利用记录介质在激光的照射下产生从晶态到非晶态转变的原理进行信息记录的。

光盘按信息存储的方式可分为：模拟式光盘和数字式光盘两大类。

按存储标准划分，光盘可分为音乐光盘 (CD—DA)、相片光盘 (CD—G)、电影光盘 (VCD)、激光视盘 (LD)、多媒体交互式光盘 (CD—I)、混合光盘以及 DVD 系列盘 [DVD 只读光盘 (DVD—ROM)、DVD 视盘 (DVD—Video)、DVD 音乐盘 (DVD—Audio)、DVD 写一次盘 (DVD—R)、DVD 可擦写光盘 (DVD—ROM)]。

按光盘出版物的收录范围划分，可擦写光盘有纯文字信息光盘、数据库光盘、多媒体光盘。

(4) 导体存储技术。这种存储器一般采用晶体管快闪存储器 (solid-state flash memory)，其优势是适用于宽范围的温度、湿度条件，而且由于没有可运动部件，存取时很安静，具有体积小，防潮、防磁，耐高低温等特点。但价格昂贵，目前容量在 4M 至 128M 之间。它有五种类型：Compact-Flash、MultiMedia 卡、SD (Secure Digital) 卡、SmartMedia、索尼 Memory Stick 等。这些存储器主要用于可移动存储器，现在很热的 MP3 播放器，就是采用晶体管快闪存储器，另外如 NETAC (朗科) 的 OnlyDisk，采用 USB 接口，存储容量在 8~256M，且不需要驱动器，可以直接插在主板的 USB 接口上，使用极为方便。

(5) 网络信息。网络信息是以数据库和网络为基础，以磁盘、光盘为存储介质，通过联机系统及网络向用户提供服务的。网络上的信息资源按所存储的服务器的类型可分为 WWW 信息资源、FTP (File Transfer Protocol) 信息资源、Telnet 信息资源、用户新闻网/电子邮件群/邮件列表信息资源、Gopher 信息资源和 WAIS 信息资源等类型。本节第四部分将作具体介绍。按用户