

国家社会科学基金项目
GUOJIA SHEHUI KEXUE JIJIN XIANGMU

数字资源整合： 理论、方法与应用

马文峰 杜小勇 著



北京图书馆出版社

G250.76/18

2007

国家社会科学基金项目

数字资源整合:理论、方法与应用

马文峰 杜小勇 著

北京图书馆出版社

图书在版编目(CIP)数据

数字资源整合:理论、方法与应用/马文峰,杜小勇著. —北京:北京图书馆出版社,2007. 12

ISBN 978 - 7 - 5013 - 3535 - 0

I. 数… II. ①马…②杜… III. 数字图书馆—研究 IV. G250.76

中国版本图书馆 CIP 数据核字(2007)第 149327 号

书名 数字资源整合:理论、方法与应用

著者 马文峰,杜小勇 著

出版 北京图书馆出版社 (100034 北京西城区文津街7号)

发行 010-66139745 66151313 66175620 66126153

66174391(传真) 66126156(门市部)

E-mail cbs@nlc.gov.cn(投稿) btsfxb@nlc.gov.cn(邮购)

Website www.nlcpress.com

经销 新华书店

印刷 北京集惠印刷有限责任公司

开本 880×1230 毫米 1/32

印张 10.5

版次 2007 年 12 月第 1 版 2007 年 12 月第 1 次印刷

字数 270(千字)

书号 ISBN 978 - 7 - 5013 - 3535 - 0/G · 722

定价 28.00 元

前 言

进入 21 世纪以来,数字资源整合作为信息资源管理学科的一个新兴领域得到广泛的重视和迅速的发展,主要体现在:第一,整合数字资源,提高其利用效率,已成为信息资源管理实践的必然选择,多种形式的数字资源整合活动在各个领域如火如荼地开展起来;第二,数字资源整合实践的发展推动了整合技术的成熟与完善,支撑数字资源整合的新技术不断应用于实践,又进一步促进了实践的深入;第三,在实践的基础上,学术界也积极进行了理论探讨,数字资源整合的理论研究空前繁荣,取得一系列有意义的研究成果,丰富了对数字资源整合的理论认识。

数字资源整合的研究与实践在这样短的时间内取得如此迅速的发展,无疑是一件令人振奋的事。但从整体上看,也存在一些亟待解决的问题,集中表现为:在理论研究上,虽然成果丰硕,但多数研究缺乏深度和广度,尚未形成公认的核心概念和完整的理论体系,理论研究的滞后制约了数字资源整合的可持续发展;在实践层面上,由于缺少理论的指导,各领域数字资源整合活动发展迅速的同时,也带有不同程度的盲目性和随意性,在具体整合过程中涌现出了诸多问题,需要从理论与实践的结合上给出解释和回答。

所以,在国内外学者研究的基础上,结合数字资源整合实践,通过必要的综合、凝练和新的研究以达到更高的发展水平,是数字资源整合研究领域的发展路径。

本书即是对这一发展路向的一个努力。概括地说,本书立足于数字资源整合的实践,在现有研究成果的基础上,对数字资源整合进行全面系统的梳理,提炼出既具普适性又有针对性的原理与方法,从整体上构建起数字资源整合的研究框架。

需要说明的是,数字资源整合的研究内容非常丰富,在有限的空间中不可能面面俱到,为兼顾全书的体系结构,突出重点,避免过于臃肿,有些内容只能从略。

数字资源整合是一个正在迅速发展的学科领域,虽然在本书中我们力图反映最新的研究动向,但还是难以跟上数字资源整合研究快速发展的步伐,一些新的内容没有能够在书中得到充分反映,有些重要的理论、方法与技术问题的探讨还要继续下去;特别是语义网、知识网格以及新型技术方面的突破性进展,揭开了数字资源整合研究的新篇章,本书仅在十一章予以揭示,没有进行更为深入的研究,但我们热切地期待着同行们(当然也包括本书作者)能在后续的著作中予以全面而充分的反映。

数字资源整合是一个全新的研究领域,许多方面的研究都还刚刚开始。本书的宗旨是以我们对数字资源整合的概念和研究内容的理解为基础,着重讨论其理论、方法和基本模式,揭示数字资源整合的本质及主流发展趋势。因此,书中包含了许多作者个人的看法,我们将这些可能并非全部得到公认的观点写出来,目的在于抛砖引玉,引起思考。真诚地希望得到学术界同仁的赐教和匡正,共同推进数字资源整合研究更加全面、系统并走向深入。

马文峰设计全书的内容结构,撰写并整理本书的全部内容;杜小勇主持本书第九章、第十章内容——“经济学领域本体”、“经济学学科知识检索系统”和“语义标注系统”的研发,整理相关资料,并参与全书主要内容的讨论。

马文峰

2007年6月于中国人民大学

目 录

前 言	(1)
第一章 绪论	(1)
第一节 研究背景	(1)
第二节 研究意义	(6)
第三节 研究核心	(10)
第四节 研究方案	(15)
第二章 数字资源整合研究现状综述	(19)
第一节 国外研究	(19)
第二节 国内研究	(26)
第三章 数字资源整合概述	(44)
第一节 数字资源整合界说	(44)
第二节 数字资源整合的主要特征	(47)
第三节 数字资源整合的基本原则	(49)
第四节 数字资源整合的方式	(53)
第四章 数字资源整合的理论基础	(56)
第一节 知识论	(56)
第二节 哲学本体论	(64)
第三节 系统科学理论	(69)
第五章 数字资源整合的方法基础	(76)
第一节 哲学方法	(76)
第二节 系统科学方法	(80)
第三节 知识组织方法	(82)
第四节 内容分析法	(88)
第六章 基于数据的资源整合	(96)
第一节 数据整合概述	(96)
第二节 基于数据仓库的整合方式	(101)

第三节 基于中介模式的整合方式	(113)
第七章 基于信息的资源整合	(130)
第一节 信息整合概述	(130)
第二节 基于信息链接的整合	(135)
第三节 基于信息门户的整合	(149)
第八章 基于知识的资源整合	(163)
第一节 知识整合界说	(163)
第二节 基于主题图的资源整合	(172)
第三节 基于本体的资源整合	(182)
第九章 学科领域本体构建的技术平台	(208)
第一节 学科领域本体构建的基本流程与方法	(208)
第二节 领域本体描述语言	(218)
第三节 本体库管理系统	(233)
第四节 基于本体的语义标注系统	(241)
第十章 学科领域本体的构建与应用	(263)
第一节 学科领域本体构建的关键:领域本体进化	(263)
第二节 学科领域本体构建案例:“经济学学科领域本体”及 “经济学学科知识检索系统”	(270)
第三节 学科领域本体评价	(290)
第十一章 数字资源整合的趋势与展望	(304)
第一节 理论研究趋势	(304)
第二节 应用研究趋势	(309)
第三节 技术研究与应用趋势	(314)
第十二章 结语	(324)
第一节 本项研究的主要特色	(324)
第二节 继续研究的主要问题	(326)
后 记	(328)

第一章 绪论

第一节 研究背景

数字资源整合的研究与应用有其深刻的发展背景。

我们正面临这样一种信息环境的变革时期:数字信息日渐成为信息资源的主流,整序数字信息的异构资源系统急剧增加,数字资源局部有序但整体呈无序的状况日趋严重,优化数字资源利用环境,有效获取信息、知识成为人们关注的重点。

如何利用新型的技术和方法,对局部有序而整体无序的数字资源进行整合,将分布于异构系统中的数字资源进行类聚、融合和科学的重组,最终形成以知识组织体系为支撑的数字资源知识整合的大环境,提高数字资源传播与利用效率,是信息资源管理所面临的紧迫课题。数字资源整合即是在这样的背景下发展起来的一个研究领域。

一、数字信息成为主流媒体

在数字载体兴起之前的几百年中,信息载体体系是以纸型载体为主导的,由纸型载体与缩微载体、磁载体和激光载体等多种载体相互兼容而形成的一个稳定的体系结构^[1]。千百年来,在信息载体的家族中,纸型载体始终一统天下,占据主体地位。数字化、网络化环境催生下的数字信息载体,以其特有的性能优势对传统载体带来了巨大的冲击,使得纸型载体的主流地位受到重大挑战。虽然纸型载体永远不会退出信息传递的历史舞台,但数字信息将逐步成为信息传递的主流媒体;特别是在一些对信息获取实效性要求高的社会活

动领域和学科领域,数字媒体正扮演着越来越重要的角色;在中青年这一庞大的信息用户群体中,数字信息已经成为重要的信息来源。

据 Tgdaily 报道,Internet 继续以惊人的速度发展,预计到 2006 年底全球将出现超过 1 亿个网站^[2]。根据美国加州大学对全球信息量发展的调查测算,2003 年,全球的浅层网页(Surface Web)数量已达到 40 亿,存储信息量达 167TB。与 2002 年相比,深层网页(Deep Web)数量增长了 400—450 倍,达到 5 500 亿,并且存储了多达 90 000 多 TB 的数量。美国研究图书馆协会 ARL (Association of Research Libraries)的统计数据表明,2002 年至 2003 年度,该协会的 110 家成员图书馆在采购和建设电子资源方面共投入两亿多美元,占图书馆总经费的 25% 左右。而在 1991—1993 年度,这方面的花费在图书馆总经费中仅占 3.6%^[3]。

据国际数据公司(International Data Corporation,简称 IDC)2007 年 3 月发布的最新研究报告《数字宇宙膨胀:到 2010 年全球信息增长预测》^[4],2006 年全球每年制造、复制出的数字信息量共计 1610 亿 GB,中国数字信息量为 127.1 亿 GB,占全球信息量的 7.9%;受“富媒体”、用户创建内容和 16 亿网民三大因素推动,到 2010 年,全球数字信息量预计为 9880 亿 GB,其中,中国的数字信息量预计为 900.5 亿多 GB,占全球信息量的 9.1%。2006—2010 年,全球信息量增幅为 6 倍,而中国则高达 7 倍。

数据库是数字信息开发利用的最有效的形式,是存储、检索数字信息的主要工具。随着数字信息量的高速增长,数据库建设也迅猛发展。据有关调查统计,1990 年我国具有一定规模的数据库只有 655 个^[5],到 2005 年 12 月 31 日,全国在线数据库约 29.54 万个,拥有在线数据库的网站约为 17.0 万个^[6]。

数字信息的迅速发展,激发了人们通过网络利用数字信息资源的热情。IDC 报告指出,1996 年,仅有 4800 万人使用互联网,2006 年,互联网用户数达到 11 亿。IDC 预测,2010 年将有 5 亿新用户上网。中国互联网络信息中心(CNNIC)2007 年 7 月 18 日发布的《第

20 次中国互联网络发展状况统计 报告》^[7]指出:与 2006 年中相比,主要基础资源呈爆炸式增长。其中,IP 地址增长率接近 40%,域名数增长率达到 211%,网站增长率达到 66%,国际出口带宽增长率达到 46%;我国拥有的 IPv4 地址数达到 1.18 亿个,排名世界第三;内地网民总人数达到 1.62 亿;区域名总数已达到 918 万个,年增长率达到 416.5%;网站数量达到 131 万个,年增长率达到 137.5%。

上述说明,数字信息日渐成为主流的信息资源,网络成为人们寻求信息知识的重要渠道,人们获取信息知识的行为方式发生了根本性的变革。

二、数字资源获取面临困境

数字信息虽然逐渐成为主流信息媒体,但在数字资源的获取上则显得不尽如人意,数字资源孤岛和数字资源超载一直是阻碍数字资源有效获取的两大困境。

“数字资源孤岛”是指由于存储数字信息的系统所具有的多样性和异构性,而导致数字资源获取不便、难以共享的一种信息环境状态。数字资源孤岛最突出的特征是局部有序而整体无序。局部有序即指在单一的资源系统内对数字信息进行了不同程度的描述与组织,整体无序是指众多的异构资源系统间的不兼容性,资源对象和资源内容缺乏有机关联性。

“数字资源孤岛”有其历史成因。不论是图书情报界的数字图书馆建设,还是企业与政务的信息化建设,都经历了从资源建设到资源管理的发展阶段。在资源建设的初始阶段,围绕某项具体应用和特定需求,开发、引进或购置一个个数字资源系统 and 应用系统,从而导致“数字资源孤岛”不断产生,形成资源分散、使用不便、共享困难的局面。

“数字资源超载”是指由于数字资源数量的巨大,超出人们有效接收、利用的能力,导致在海量数字资源中难以准确获取信息知识的一种信息环境状态。数字资源超载最突出的特征是内容交叉重复、

冗余信息量大,从而严重干扰信息的选择与获取。无论是 Web 搜索引擎还是数据库,当代大多数的检索系统都采用了某种基于关键词的检索技术,这虽然使信息获取具有可能,然而,由信息组织机制直接引发的信息检索的弊端是显而易见的,基于关键词匹配的信息检索技术无法达到从海量数字资源中对信息及知识的高效率获取,难以彻底解决数字资源超载问题。

面对上述两大困境,人们开始反思,需要对数量庞大的异构数字资源进行集成与重组,以消除“数字资源孤岛”和“数字资源超载”现象,提高数字资源的利用效率。数字资源整合正是数字资源建设发展到一定规模与程度的必然要求。

三、用户需求不断深化

数字资源整合具有很强的目的性,它是以用户需求为导向,有针对性地进行的数字资源的管理活动。

数字资源成为主流媒体,改变了人们获取信息知识的行为方式,与此同时,也催生了用户对数字资源获取与利用的新需求。这种新需求主要表现为两方面:一方面,在需求方式上,由分散式获取到集成式获取。用户希望不要遍历所有,在一个入口处即可一次性获得分散在多个异构资源系统中的资源,又能获取具有关联关系的多种资源。另一方面,在需求内容上,由对信息的需求到对知识的需求。用户希望在海量的数字资源中找到需要的信息,更希望在大量的信息中获取蕴涵的知识内容及知识的逻辑关系,帮助自己理解、应用和创造知识。

用户对数字资源需求的深化,集中反映了用户需求所遵循的两个基本原则,即省力原则和价值原则。省力是人类行为的一种普遍法则,用户总是希望以最小的努力(花最少的精力)获得他所需的信息;价值是信息需求满足的主要依据,用户总是希望获得有价值的信息,这是用户信息需求行为的驱动力。在数字化、网络化的环境中,用户需求依然遵循着一些共同的原则和规律,但被赋予了新的内涵。

用户需求是随着信息环境的变化而不断深入的,数字资源整合的研究和应用,也正是适应信息环境和用户需求的变化应运而生的。

四、整合技术的变革与创新

信息资源管理发展的历史表明,人们关于信息资源管理的科学认识和实践,都是建立在特定历史时期信息资源管理技术发展水平之上的。数字资源整合也是如此,它的发展和技术的变革与创新密切相关。可以说,数字资源整合的实质是新兴技术对传统信息资源管理的一场深刻变革。

纵观数字资源整合的实践活动,不同层次、不同类型的整合都必须依赖于一定的技术手段,都是建立在相应的整合技术基础之上的。从分布式构建技术(中间件、数据仓库)、互操作技术(XML、OAI、OpenURL),到新型的知识组织技术(领域本体、概念图、主题图)、知识处理技术(知识发现、知识获取、知识表示、知识展现、知识推理),再到分布式、基于组件的软件技术(Web Services),直至网格技术和语义 Web 技术,每次技术变革,都伴随着新的整合方式的出现,技术变革的每一次跨越,都为数字资源整合创造了一个新的发展空间,都推动着资源整合向更深、更高的层面发展。

“技术和信息体系结构前景逐渐显露出来的发展趋势表明,我们正朝着技术变革时期迈进”^[8],数字资源整合的理论研究和实践应用正是得益于技术的变革与创新。一方面,整合理论与应用的发展,必然推动着整合技术的变革与创新;另一方面,整合技术的不断变革,反过来又为整合理论研究提供思路、方法、证据及材料,促进数字资源整合研究及应用的深入发展。

五、语义网大环境的推动

当前,我们面临一个新的信息大环境,即与人类信息活动密切相关的万维网正跨入一个新的发展进程:由数据网、信息网向语义网演变。语义网的实质是以元数据作为资源的描述标准,以 XML 表示资

源内容和结构,以 RDF 作为描述资源语义的通用框架,通过本体实现对资源语义的描述,使语义网的实现成为可能。语义网的目标是提供具有语义的知识网络。

可以说,语义网是以知识资源为基础的人类知识整合的大环境,是人类知识的宏观网络。语义网的实现,需要以各个领域、各个层面的知识资源整合为基础。语义网的提出及其迅猛的发展态势,为数字资源整合提供了新的思维方式、新的理念和新的技术,从而推动着数字资源整合在数据整合、信息整合的基础上向知识组织与知识整合迈进。在语义网的驱动下,“必将实现知识组织方法‘语法—语义—语用’的新跨越”^[9],也必然对数字资源整合的理论研究与实践探讨产生深远的影响。

第二节 研究意义

数字资源整合是针对数字资源无序而难以获取利用的困境提出的。它既是图书情报界、企业界等领域数字资源建设的重要任务,也是信息资源管理领域的一个重要的研究课题和亟待解决的问题。对数字资源整合的研究,不仅对信息资源管理及相关学科理论内容的丰富以及信息组织、知识组织等核心概念的认识和理解上有深远的理论意义,而且在推进信息资源管理领域的文献组织、信息组织到知识组织的转型有着重要的现实意义。

一、理论意义

1. 数字资源整合研究是完善信息资源管理研究内容的需要

信息资源管理(Information Resource Management,简称 IRM)是形成于 20 世纪 70 年代末的新兴的综合性学科研究领域,其历史虽短,但发展迅速,目前信息资源管理的研究内容在不断完善和逐步深化。虽然不同学科领域对信息资源管理具有不同的理解和认识,但

其共识都是将信息看作一种重要的资源,将信息资源的管理看作是以现代信息技术作为手段,对信息资源进行计划、组织、指挥、控制和协调的过程。

信息资源是信息资源管理的研究重心和研究基线,而数字资源是信息资源的重要组成部分。数字资源是经过数字技术处理的,通过网络以数字形式发布、存取、利用的信息资源总和。随着社会各个领域的信息活动全面步入数字化网络化的环境,磁介质正迅速成为普遍的主流存储介质,数字资源已成为人们获取知识、信息的重要载体。但是,与其他载体类型的信息资源相比,数字资源具有分布离散性、复合性、动态开放性、交互性、关联性、嵌套性、集合性、可重组性、软硬件强依赖性等复杂性特征^[10],这种复杂性特征对组织数字资源起着明显的制约作用,直接导致数字资源的组织有着不同于其他类型信息资源组织的典型特征,并引发与此相关的数字资源的收集、检索与利用等一系列问题也具有相应的特征。

但是,从目前信息资源管理研究成果来看,尚未将数字资源作为一种相对独立的信息资源类型来认识,对其基本问题尚未进行深入的探讨。数字资源虽与其他类型的信息资源有一定的交叉和重叠,但从本质上看具有其自身特点和独立性。将数字资源作为一种相对独立的信息资源类型进行研究,全面深入地认识数字资源的内涵、特征及功能,对于信息资源的收集、组织、检索、开发利用与管理的研究与实施具有重要意义。

数字资源的基本理论问题是数字资源整合的研究重点,所以,对数字资源整合的研究有助于发展和完善信息资源管理理论。

2. 数字资源整合研究是深化信息组织理论研究的需要

信息组织是信息采集后的首项工作,是信息检索、信息资源开发与管理的前提。作为一种科学实践,信息组织是信息资源管理过程的重要环节;作为一种学科理论,信息组织是信息资源管理理论的基石。

对于信息组织的理解,研究者依然见仁见智,但其实质都是将信

息组织看作是对信息资源的一种整序活动,是对采集的无序信息进行系统化和有序化的一种过程;信息组织的结果是有序化的集合,信息组织的目的是方便人们利用信息和有效地传递信息。对信息的序化或整序一直是信息组织的中心内容。

随着“知识”作用的日趋凸现以及知识组织技术的发展与创新,“知识”和“知识组织”的研究成为许多相关学科领域共同关注的问题,由文献组织、信息组织向知识组织迈进,成为信息资源组织的一个显著发展趋势。在数字资源日渐成为信息资源主体的环境下,近年关于信息组织的研究成果对数字资源组织给予了充分的重视,也揭示了知识组织是信息组织的发展趋势。

然而,综观这些研究成果,在对数字资源的揭示、加工、序化、存储和构建上,多是运用支撑信息组织的方法与技术,如元数据、主题网关、网络资源指南、文件和数据库及信息构建等,其对数字资源的描述与组织依然停留在信息组织的层面上。通过上述技术方法对数字资源进行整序,虽然可以形成局部有序的信息集合,但难以解决局部有序而整体无序的异构数字资源的利用状态,也无法从整体无序的信息环境中有效获取知识。也就是说,对数字资源知识组织的探讨还是在信息组织的框架下所进行的一种有限研究,没有本质上的突破。

知识组织与信息组织是信息资源管理在管理内容和管理方式上逐次递进的两个环节和阶段。知识组织虽与信息组织具有一定的重合,但是一种与信息组织具有本质不同的理论、方法与技术,因此在信息组织的框架下是无法容纳丰富的知识组织理论内涵的。为此,创新信息组织的理论方法与技术势在必行。

数字资源整合主要研究数字资源整合的核心概念和理论体系,研究数字资源整合的基本模式与支撑技术。因此,从理论上说,数字资源整合的提出,给 20 世纪中期以来形成的信息组织理论带来根本的变革与挑战,研究数字资源整合有利于拓展信息组织的研究范围,深化信息组织的理论研究内容。

二、实践意义

1. 数字资源整合研究是整合实践发展的需要

作为一种实践活动,数字资源整合是信息资源管理的重要实践领域。随着数字资源日渐成为信息获取的重要载体,由对数字资源的简单“整序”到对异构数字资源系统的“重组”与“整合”,成为人们获取数字资源的迫切需要。

然而,人们逐渐认识到,数字资源存在着不同于非数字资源的复杂特点,直接影响对数字资源的获取与利用,从而导致数字资源的组织活动与非数字资源的组织活动的诸多不同。同时也意识到,数字资源整合活动需要理论的指导,因为在日趋发展的数字资源整合的实践面前,信息资源管理及传统的信息组织的理论与方法逐渐显露出它的不适应性。

信息资源管理是对信息资源进行计划、组织、指挥、控制和协调的理论,但事实上,现有信息资源管理及信息组织理论受特定社会环境的影响,偏重于对非数字资源管理的研究,在其理论的发展中,虽然一定程度容纳了对数字资源管理的研究,但所依据的理论方法与技术还是基于信息组织,较为忽略对数字资源的“重组”,忽略对数字资源基于知识层面的“整合”。因此,在数字资源整合实践活动中所产生的许多具体问题,信息资源管理及传统信息组织的一般原理是无法解释的。

数字资源整合实践活动的特殊性,以及现有信息资源管理理论与方法的不适应性,是数字资源整合研究的重要动因。也就是说,数字资源整合的实践需要理论指导,而已有的理论又不能完全解决实践中的具体问题,这就是数字资源整合研究的存在并要不断发展的必然。

2. 数字资源整合研究来自实践又指导实践

数字资源整合研究是对实践的概括和总结,它应来自于实践,同时又应对实践具有指导作用。近些年来,国内各领域特别是企业和

图书情报领域,数字资源整合实践发展比较迅速,产生了多种数字资源整合方案,形成了不同规模和不同应用的整合系统。然而整合实践的发展也存在不少问题:对数字资源的整合多局限于整合某一方面,如某一整合方式、某种整合技术或某类资源整合,而缺乏对资源整合的宏观的、系统的、全面的把握;多侧重于基于数据的整合与基于信息的整合,对资源整合的主流趋势——知识整合的应用与研究关注不够,应用成果较少,等等。上述问题就是由于缺乏对数字资源整合基本规律的研究和认识所造成的。

数字资源整合应是对数字资源整合实践的科学概括和总结,应反映和揭示数字资源整合活动的本质特征及其客观规律。全面系统地研究数字资源整合,提炼出既具普适性又有针对性的原理与方法,对于解决数字资源整合实践发展中不断出现的新问题,引导和推动整合实践迅速而有序的发展,具有重要现实意义。

第三节 研究核心

一、研究目标

准确把握国内外数字资源整合的主流发展模式,全面梳理数字资源整合的理论、方法与技术,探讨数据整合、信息整合和知识整合这3个层面资源整合的基本原理、体系结构,揭示其研究与应用的现状及趋势;侧重研究知识整合的主要形式——领域本体构建的方法与技术,研制基于经济领域本体的“经济学学科知识检索系统”,初步实现基于经济领域本体的知识整合和知识检索的应用,为国内同仁对数字资源整合的深入研究奠定基础。

二、研究内容

本项研究的核心内容定位在5个层面,即基础理论研究、整合方