●马文峰　杜小勇

论数字资源整合的方法与技术基础*

摘　要　数字资源整合在数字图书馆建设领域的迅速发展不仅有深刻的实践动因,也有深厚的方法和技术的支撑。哲学方法、系统论方法和知识组织方法是数字资源整合的重要方法基础。支持知识整合的主要技术包括知识表示、知识获取、知识推理和知识共享等。参考文献9。

关键词　数字资源整合　方法基础　技术基础　数字图书馆

分类号　G253

ABSTRACT　The rapid development of the integration of digital resources has its practical impetus and methodological and technological supports. The basis of the integration of digital resources consists of the methodologies of philosophy, systems theory and knowledge organization, and the technological basis consists of knowledge representation, knowledge acquisition, knowledge inference and knowledge sharing. 9 refs.

KEY WORDS　Integration of digital resources. Methodological basis. Technologial basis. Digital library.

CLASS NUMBER　G253

数字资源整合是依据一定的需要,对分散无序、相对独立的数字对象进行类聚、融合和重组,重新组织为一个新的有机整体,形成一个效能更好、效率更高的新的数字资源体系。数字资源整合是一种组织、管理数字资源的理念,是一种优化、重构数字资源的过程,也是一种有效利用数字资源的环境。

数字资源整合在数字图书馆建设领域的迅速发展不仅有深刻的实践动因,也有深厚的方法和技术的支撑。数字资源整合是以科学的方法为指导,通过一定的技术手段实现的,它必须建立在一定的方法与技术基础之上。在此,笔者无力对数字资源整合的方法与技术进行方方面面的详细阐述,只是择其要者对其方法与技术基础进行分析和说明。

1　数字资源整合的方法基础

数字资源整合有3个重要的方法基础,即哲学方法、系统论方法和知识组织方法。哲学方法是在一定哲学原理的指导下,在社会实践中逐步总结出来的关于认识事物和解决问题的根本方法的理论,对人类活动具有普遍的指导意义。哲学方法是数字资源整合所遵循的基本规范和准则。系统论方法是一般科学方法论,是哲学方法论的具体化,对数字资源整合活动起着直接的指导作用,是数字资源整合领域应用最多和最有效的方法之一。知识组织是数字资源整合领域重要的方法论基础,是系统论方法在该领域的深化和具体化。哲学方法、系统论方法和知识组织方法既互相区别、互相联系,又相互影响、相互补充。

1.1　哲学方法

客体(客观对象)的多样性、完整性和系统性是数字资源整合的本体论基础。主体认识和改造客体的方法论,是哲学方法论的首要和重要的内容之一。客体是主体活动指向的对象,其运动形式和现象复杂多样,千差万别,这种多样性表现为多种层次、多种结构和多种功能。客体的多样性和差异性,决定了数字资源纷繁复杂的表现形式,决定了人们对数字资源必须进行分门别类地研究,对各种类型的数字资源进行不同层次和功能地组织,由此而形成数字资源系统的多样性和差异性。

虽然客体是多样的,但客体间却存在各种各样的联系。关系是客体的根本属性之一,是指客体之间相互作用、相互关联的状态,包括纵向和横向两种方式[1]。纵向关系是指客体的自我联系,反映了客体发生发展的历史过程;横向关系是指客体之间的相互联系,一客体如果不同其他客体相联系,就不能规定自身的存在。没有关系,客体就失去存在的基础。客体的关系属性决定了客体联系的普遍性,决定了整个客体世界是一个互相联系的统一整体。整合数字资源的活动必须建立在对客体关系属性认识的基础之

* 本文系国家社会科学基金项目(项目编号:04BTQ003)和国家自然科学基金(项目编号:60496325)的研究成果。

上。客体的关系属性决定了数字资源联系的普遍性，因而在数字资源整合的活动中，既要认识作为客体的数字资源对象，也要把握、研究数字资源对象的关系属性。对数字资源的类聚、融合和重组，实际上就是对多样性和差异性的数字资源系统进行整体性研究；对数字资源不同层面的组织与整合，正是客体联系的普遍性原理在资源整合领域应用的结果。

主体具体认识的有限性与人类认识的无限性之间的矛盾是数字资源整合的认识论基础。实现主体与客体的统一是哲学方法论的根本原则，主体与客体的统一是通过认识活动和实践活动来实现的。但一方面，由于受客体条件的制约，使得主体对客体的改造、反映只能达到一定的程度或层次；另一方面，人类对客体的认识能力是无限的，通过不同历史时期和不同认识过程的积累，从而达到有条件的、近似的把握客体。数字资源整合即是主体与客体间不断统一的活动过程。人们对数字资源的组织与整合受制于两方面局限：一是主体认识能力的局限，人们对数字资源整合理念的认识有一个逐步发展完善的过程；二是实践条件特别是技术水平的局限，数字资源的整合程度依赖于整合技术的成熟程度。主体具体认识的有限性与人类认识的无限性之间的矛盾既决定了数字资源整合是一个逐步深化的过程，是一个复杂曲折的发展过程，同时又推动人们不断地向纵深和广阔方向探索数字资源整合的本质及其发展规律。

1.2 系统论方法

现代系统论是一种研究部分与整体之间相互作用的科学理论。它以抽象的系统为研究对象，着重考察系统中整体与部分、结构与功能之间的关系，并运用数学手段和计算工具，确定适用于所有客体系统的一般原则和方法。系统方法就是按照事物本身的系统性把对象放在系统的形式中加以考察的方法，它是认识、调控、改造、优化和创造系统的有效手段。其重要的方法论原则是整体性、结构性和功能性[2]。

整体性原则是系统方法的核心和基本出发点，也是哲学方法论关于普遍联系原理的具体体现。整体性是指系统都具有若干要素，这些要素相互依存、相互关联、相互制约，构成具有特定功能的有机综合体；同时，系统是整体与环境的统一，任何系统都在一定的环境中存在，因而也与其环境发生相互作用，在这个相互作用的过程中，系统表现出相应的整体功能。这种整体功能具有一种“非加和性”，即整体不是各要素简单的组合，而是由于系统要素的有机关联，使系统的整体功能发生了质的飞跃，远远超出各单个要素的功能总和。结构性原则要求把事物看成不仅是一个整体，而且视为是按一定的层次和等级组成的层次性框架，整个系统就是由各种不同层次和等级的系统交织起来的网络结构。功能性原则即是运用各种有效方法，促进系统的组织、结构和功能的整体改进，使得系统整体对环境产生最佳的作用和影响，具有较大的灵活性和发展性。

系统论方法为人们研究和进行数字资源整合提供了有效工具。根据系统论的整体性方法原则，应该将数字资源整合看成是一个与环境相互作用的整体系统，将其置于更广阔的整个社会大环境中去考察，从整体的观念出发去研究探讨数字资源整合的本质和规律。根据关联性、结构性和功能性的方法原则，在对数字资源的整合中，要注意数字资源的关系、层次和功能的整合，既要反映数字资源对象间的结构和语义的内在关系，保持数字资源对象学科的完整性，也要对数字资源进行多维整合，体现资源整合的结构性和层次性，同时要运用一定的技术手段和方法，使数字资源得到优化组合，取得最好的组织结构和组织功能。

1.3 知识组织方法

所谓知识组织，是在信息组织的基础上，研究知识的获取、描述、整理、表达、控制、共享等整个知识组织过程的理论与方法。知识组织的重点在于对知识和知识间的关联进行揭示和组织，知识获取、知识处理、知识表达和知识共享是知识组织的重要内容。

知识组织方法是在系统论方法的指导下，依靠专门的技术，按照知识的本质属性组织知识、建立知识系统的方法和手段。它有以下几个显著特征：(1)整体性。是系统方法整体性原则的具体体现。整体性包含两层含义：一是指知识系统内部的不可分割性，二是指知识系统内部的有机关联性。知识系统整体性的主要表现形式是系统内部要素空间的整体性、时间的整体性和逻辑的整体性。整体性是建立知识组织系统的重要依据。(2)综合性。主要指知识组织系统不是由单一要素、单一层次、单一结构、单一功能构成的整体，而是由多类知识载体、不同层次的结构和系统多种功能要素构成的有机总体。(3)层次性。是指知识系统内部各要素的构成关系及所形成的纵向上不同质态的排列次序。知识系统的层次性大致包括时间关系层次、空间关系层次、逻辑关系层次和数量关系层次。知识系统层次的数量、质量、顺序和

层际关系对系统整体功能具有重要影响。(4)关联性。这是整体性的延续。是指系统的要素之间、要素与系统整体之间和系统与环境间的有机的、多维的关联性。知识系统的关联性体现在三个方面：一是要体现知识概念的关联性，以保持学科知识体系的完整性和系统性；二是要体现不同知识系统间的关联，以保持人类知识体系的整体性；三是要注意知识系统与信息环境的关联，以促进社会大环境中的知识共享和交换。(5)动态性。是指系统处于一种运动、变化和发展的状态。知识是动态变化的，知识系统的要素、要素间的关联、系统的结构、系统的功能、系统与环境等都会随着时间的推移而不断发展，只有处于动态发展中的知识组织系统才能发挥持续的效用，才有生命力。

数字图书馆资源整合是知识组织的实践活动。知识组织是数字资源整合领域最有价值、最适用的具体方法，不论是在资源整合的内容、形式上，还是在资源整合的方式、方法和技术手段上，都应该遵循知识组织的原则和方法。自觉地将知识组织理论与方法运用于数字资源整合中，实现数字图书馆数字资源的整体优化和有效获取与利用，具有非常重要的现实意义。

2　数字资源整合的技术基础

实现数字资源整合不仅需要方法论的指导，还需要依赖相应的技术和工具。数字资源整合大体可分为信息整合(数据整合)和知识整合两种方式。信息整合是知识整合的基础，知识整合是数字资源整合的高级阶段，也是数字图书馆资源整合的最终目标。信息整合的技术主要有信息描述技术、信息集成技术、信息检索技术等，都已较为成熟。支持知识整合的技术包括知识表示、知识获取、知识处理、知识表现、知识推理等与知识组织密切相关的技术，本文着重讨论知识整合的技术与工具。

2.1　知识表示技术

在知识工程领域，所谓知识表示即是对知识进行描述，是知识的形式化和符号化过程，以便于计算机对知识进行存储和处理。简单说，知识表示是考虑“怎样对现实世界建模”。因此，知识的表示同知识的组织和知识使用方式密切相关。可以说，知识表示是知识组织的表现形式。知识表示方法有多种，大致可分为基于符号的表示方法和基于连接机制的表示方法。知识表示方法可以从以下几个方面来衡量：(1)表示能力，即能否简洁有效且准确无歧义地表示领域知识；(2)可利用性，即表示出来的领域知识易于理解，可以有效地被计算机再次获取、分析和重复利用；(3)可操作性，即能使基于知识的推理有效地、符合逻辑地进行；(4)易维护性，即易于对知识进行组织、维护、管理与扩充[3]。

但是，传统的知识表示方法无法同时满足上述四个方面要求。特别是由于因特网的迅猛发展，如何组织、管理和维护海量信息也就成为知识组织领域的研究内容。为此，人们提出了用知识本体来组织和表示知识。知识本体可以看做是领域知识规范的抽象和描述，是共享、重用知识的方法，目前已经成为一种提取、理解和处理领域知识的工具，可以被应用于任何的学科和专业领域。

从形式上说，本体由类或概念、关系、函数、公理和实例5种元素组成。类或概念表示对象的集合，关系表示领域中概念之间的关联，函数是一类特殊的关系，公理代表永真断言，实例即某概念类所指的具体实例。关系在本体中非常重要，从语义上讲，基本的关系包括同义词关系、上下位关系、包含关系和相关关系等。

本体的描述一般都是基于某种逻辑语言的，目前RDF(S)已成为一个能对本体进行初步描述的标准语言。描述逻辑(DL)是一个相当重要的知识表示语言，目前正被积极应用于本体描述，或者作为其他本体描述语言的基础，几个主要的知识本体语言CKML、OIL、DAML + OIL和已成为W3C国际标准的OWL就是建立在描述逻辑的基础上的。

基于知识本体的知识组织和表示的基本流程包括：(1)需求分析：确定本体建立的领域、范围、目的、需求等。(2)本体概念化：确定某一领域知识本体核心概念集，构建知识本体概念关系。(3)本体形式化编码：用选定的本体语言来描述知识本体。(4)本体进化：对知识本体的结构、概念和关系不断进行丰富、完善和改进[4]。

知识本体无疑是数字图书馆资源整合最本质、最重要的技术基础。通过某领域的知识本体可将该领域的知识组织起来，使数字图书馆对知识的表示从信息的集合到知识网络和知识地图，数字图书馆的最终目标——面向用户的知识检索与知识服务才有可能成为现实。基于知识本体技术的数字资源整合是数字图书馆资源整合主流发展模式，应用前景非常广阔。

2.2 知识获取技术

如何获取、揭示数据之间存在的各种关系是知识组织最关键的内容。知识获取也可称为知识发现(Knowledge Discovery in Database,简称KDD),是指从大量数据中提取出有价值的知识,按知识的内容特性聚集,并以特定的方式加以表示的过程与方法。知识发现主要包括数据准备、数据挖掘、结果的表达和解释三个步骤。其中数据挖掘是知识发现过程的核心步骤,是知识发现采用的特定算法或技术。

数据挖掘一般可以分为描述和预测两类[5]。描述性挖掘任务试图刻画数据库中数据的一般特性,而预测性挖掘任务是根据当前数据归纳出一种模型来,以进行预测。如果按照数据挖掘可以发现的模式类型来看,与数字资源的知识组织关系密切的数据挖掘技术大体分为以下几类:类/概念描述(也称特征化和区分,是概念形成的方法之一)、关联分析(用于发现存在于概念/类之间规律性的关联)、分类分析(按照事先定义的标准对数据进行归类分类)、聚类分析(根据最大化类内的相似性、最小化类间的相似性原则对数据对象进行聚类或分组,是分类的逆向方法)等。

知识发现的对象通常是大型数据库或者数据仓库,但从广义上说,知识发现的对象应该是所有的数据集合。目前知识发现的数据源类型向多样化发展,特别是对网络资源的知识挖掘引起普遍关注。但网络资源较之一般的数据库和数据仓库,不仅数据量更大,结构更加复杂,动态性更强,而且具有分布性特点,这就大大增加了知识发现和数据挖掘的难度;同时,新的复杂的数据类型不断出现,需要知识发现和数据挖掘在理论、方法和技术上要有新的突破和发展。

知识发现被认为是在知识组织领域具有重要影响和应用前景的关键技术,将知识发现的机制应用到数字图书馆的资源整合中,无疑将给数字图书馆数字资源的知识组织带来全新变革。

2.3 知识推理技术

领域知识本体是知识组织系统的有机组成部分。在知识组织系统中,除了类(概念)、关系、事实(实例)和规则,还应包括知识推理系统(归纳、演绎等知识推理的方法与机制)。知识推理系统是知识组织系统的重要组成部分,也是重要的技术和方法基础。

所谓知识推理,是指按照某种策略从已知事实或知识中推导出新知识的过程。在人工智能领域,知识推理是最重要的研究课题之一,由此形成了不同的知识推理机制,最有代表性的是演绎推理和归纳推理。计算机系统中的知识推理是由程序实现的,称为推理机或者推理引擎。为使推理过程更加有效,需要设计一系列的控制策略对推理过程实施控制。主要的控制策略包括关于推理方向的控制(如正向推理、反向推理及混合推理),关于搜索策略的控制(如宽度优先搜索、深度优先搜索、分支限界搜索和启发式搜索)和关于冲突解决策略的控制(如规则排序)等。

建立在一阶逻辑基础上的归结反演推理系统是最经典的知识推理系统。但是,一阶逻辑的推理系统并不总是有效的,因此,人们又提出了多种受限的一阶逻辑推理系统。描述逻辑就是这样的一阶逻辑的一个子集,被广泛认为适用于知识本体的表示和推理[6]。描述逻辑之所以在知识组织中备受关注,主要原因就是因为描述逻辑具有足够有效和强大的推理能力,能够提供完备高效的知识推理机制,使蕴涵在知识本体中知识的利用和共享成为可能;描述逻辑的推理系统是可判定的,描述逻辑的推理过程总是能够结束并返回正确的结果。描述逻辑由四部分组成:表示概念和关系的集合,关于问题领域一般性的知识即内涵知识的集合Tbox,与特定问题相关的外延知识集合Abox,基于Tbox和Abox之上的推理机制。不同的描述逻辑系统的表示能力与推理机制由于对这四个组成部分的不同选择而不同。

知识推理的技术与方法是知识组织的核心关键技术,是知识组织系统提供精确、有效知识的基础和前提,需要加以研究和关注。

2.4 知识共享技术

如何有效地使用知识、共享知识资源,是数字图书馆资源整合的最终目标。网格技术为有效使用与共享知识资源提供了技术框架[7]。网格把用通信手段连接起来的资源无缝集成为一个有机的整体,以实现互联网上资源的全面连通,包括计算资源、存储资源、通信资源、软件资源、信息资源、知识资源、专家资源、设备资源等。网格提供的分时共享、资源预约、资源授权、资源组合、数据副本等技术,可以有力地支持广域范围内的资源共享。

网格可以分为四个部分:(1)网格资源:指网络所有分布的、可访问的计算资源;(2)网格中件间:包括一系列工具和协议软件,用于屏蔽网络资源的分布、异构,并提供透明一致的接口;(3)网格开发环境和工具:供开发人员开发各种应用,用户代理的环境

和工具，用于在全局资源中调度计算；(4)网格应用层：用于运行网格应用程序，满足用户需求[8]。

欧洲网格项目将网格分为三层结构，即计算(数据)网格、信息网格和知识网格。计算(数据)网格主要解决数据访问的问题；建立在数据网格层次之上的是信息网格层次，其功用主要解决异构信息的统一访问；知识网格处于最上层，是一种提供知识信息的智能化获取和应用服务的技术框架，是一个汇集和共享知识资源的知识平台。在这个知识平台上，可以对多源、异构、海量、复杂、动态的信息进行一体化的智能处理与组织，使用户能够有效地获取、发布、共享和管理知识资源，提供所需的知识服务。

目前知识网格要解决的核心问题有3个：其一是资源的规范组织，即解决如何规范地组织资源空间，使用户能够有效、正确地根据语义操作各种资源，提高资源的使用效率；其二是资源的智能聚合，即解决如何使资源能够相互理解，根据用户的需求，有效动态地聚合各种资源；其三是资源的语义互联，即解决如何使Web资源的语义能够被机器理解[9]。

数字图书馆是以知识资源体系为支撑的一种信息服务与知识服务环境，知识网格则是一个智能互联的大环境，数字图书馆应该融入更广范围的知识网格环境中。通过整合后的数字图书馆资源是知识网格中的重要资源基础。通过网格技术，使全社会得以方便地获得与共享知识资源，是数字图书馆资源整合的发展方向，也是数字图书馆建设义不容辞的责任。

3　结语

数字资源整合是数字图书馆数字资源建设发展到一定阶段的必然要求，也是数字图书馆提供知识服务的重要基础。数字资源整合活动的发展不仅有着深厚宽广的理论与方法基础，也有着坚实的技术基础。在数字资源整合实践中，一方面，我们要自觉地以科学方法为指导，不断地探讨、发现、应用相关技术，促进数字资源整合实践的发展；同时，也要在实践的基础上，不断补充、完善、拓展数字资源整合的方法基础，构建科学的数字资源整合的方法论体系，并以此为指导，推进数字资源整合相关技术的完善和发展。

参考文献

1　金京振. 哲学方法论. 北京：民族出版社，1998
2　常绍舜. 系统科学方法论. 北京：中国政法大学出版社，2005
3　钟义信. 信息科学原理. 北京：北京邮电大学出版社，2002
4　胡运发. 数据与知识工程导论. 北京：清华大学出版社，2003
5　陈京民. 数据仓库原理、设计与应用. 北京：中国水利水电出版社，2004
6　陈文伟，黄金才. 数据仓库与数据挖掘. 北京：人民邮电出版社，2004
7　徐志伟，冯百明，李伟. 网格计算技术. 北京：电子工业出版社，2004
8　刘洁等. 中国织女星知识网格研究进展. 计算机研究与发展，2003(12)
9　饶元，冯博琴. 基于本体的XML知识表示方法研究. 微电子学与计算机，2004(9)

马文峰　中国人民大学图书馆研究馆员。通信地址：北京市。邮编100872。

杜小勇　中国人民大学信息学院教授，博士生导师。通信地址同上。

(来稿时间：2005－02－28)

(上接第48页)问题. 中国图书馆学报，2002(1)

12　陈传夫. 信息社会化过程中若干利益冲突研究. 中国图书馆学报，2002(2)
13　周庆山. 数字时代图书馆权益的保障与著作权法的完善. 国家图书馆学刊，2004(4)
14　刘华.《数字千年版权法》有关版权保护的新规定. 中国图书馆学报，2001(3)
15　张平. 数字图书馆建设中的法律问题及对策研究. 国家图书馆学刊，2004(4)
16　张平. 中国数字图书馆工作中的著作权问题. 科学新闻周刊，1999(28)
17　秦珂. 新传播权概念下图书馆服务的困境及法律调整. 图书情报工作，2002(5)

金胜勇　河北大学管理学院副教授，图书馆学系主任。通信地址：河北省保定市。邮编071002。

白献阳　河北大学图书馆学专业研究生。通信地址同上。

(来稿时间：2005－04－28)