



科学数据库 与信息技术 论 文 集

中国科学院科学数据库中心 编

中国科学技术出版社

科学数据库与信息技术 论文集

中国科学院科学数据库中心 编

中国科学技术出版社

(京)新登字 175 号

内 容 提 要

这本由 36 篇论文组成的论文集，从总论、数值数据库、文献数据库、信息技术四个方面介绍了我国科学数据库的现状，深入探讨了科学数据库发展的理论、趋势与规划、数据评估与质量控制、各类型数据库建设的方法及数据的开发利用，对从事数据库及信息系统研究的科技人员，以及相关学科的研究人员、大专院校师生都有较高的参考价值。

* * *

科学数据库与信息技术论文集

中国科学院科学数据库中心 编

责任编辑 初炳英

*

中国科学技术出版社出版(北京海淀区白石桥路 32 号)

新华书店北京发行所发行 各地新华书店经售

北京市密云县印刷厂印刷

*

开本: 787×1092 毫米 1/16 印张: 15.125 插页: 字数: 340 千字

1993 年 2 月第一版 1993 年 2 月第一次印刷

印数: 1—1,000 册 定价: 11.80 元

ISBN 7-5046-0904-8 / TP · 41

我院科學數據信息確
庫是一项基礎工程，必須準確
而又有全面地為我事業提
和我國科研事業提供服務！

衷心祝賀我院科學數據庫首屆學術討論會
論文集出版！

盧嘉錫
一九九二年春月

加強科技數據的積累
而分析利用，當好科技
研究與發展的耳目

王秉琳 一九九二年
五月廿一日

《科学数据库与信息技术论文集》 编辑委员会

主任：崔俊芝

编委：罗百昌 许志宏 龚义台 罗晓沛
苏振泽 郑崇直 翁志成 孙九林
许 禄 张俊逸 鲍信炯 史忠植
王 源 吴光义 徐克学

第一册

主 编：吴光义

副主编：施慧中

编 辑：赵诗生

张春娥

目 录

前 言	(1)
总 论	(1)
促进科技发展的现代化信息系统	罗百昌 (3)
现代数据库技术与科学数据库建设	罗晓沛 (6)
跨国服务的科技信息系统	苏振泽 (10)
地球科学数据库发展设想	施慧中、孙九林 (16)
应重视科学数据库服务工作	吴光义、赵诗生、李望平 (23)
数据库的发展和问题探讨	陆长旭、王能琴 (26)
数值数据库	(31)
集成化化学化工信息系统的发展	许志宏 (31)
稀土综合信息系统	许 禄、李国权、王淑云等 (40)
一种简便的树形(子)结构编码方法及应用	彭 琛、袁身刚、郑崇直 (47)
化学数据库—重要的化学信息资源	袁身刚、王 源、郑崇直 (51)
中国光学镜头数据库系统设计	翁志成、陈志勇、丛小杰等 (58)
用混合数据库实现科学工程数据库	翁志成、任 涛、丛小杰等 (63)
中国微生物资源数据库	马俊才、赵玉峰、王大耜等 (68)
微生物学数据源的评估与质量监控	赵玉峰、马俊才、刘澎涛等 (73)
中国珍稀濒危植物数据库服务系统的开发与研究	李 奕 (79)
建立中国高等植物形态特征数据库的研究	袁文林 (87)
兼容型数据库的结构设计初探	李国权 (93)
计算机辅助农药筛选和分子设计进展	周家驹、严新建、曹凌霄等 (97)
化学物质毒性分级、评价知识库系统	白乃彬 (103)
生物活性与结构数据库	谢桂荣、谢 前、周家驹 (108)
计算机辅助评审化工物性数据及估算方法	李晓霞、严新建、许志宏 (114)
专家系统在腐蚀与防护科学中的应用	李洪锡、张淑勤 (120)
中国国家能源数据库系统分析和总体设计	蒋镇平、庄 幸、李京京等 (124)
中国药用植物数据库系统—数值、文献 兼容技术探讨	厉森梁、肖念华、张春娥 (137)
有关数值数据库标准的探讨	罗东川、许慧芳、向卫平 (143)
文献数据库	(149)

文献数据库集中服务系统的建立	王 源 (149)
中国生物学文献数据库的检索系统	龚义台 (155)
中国生物学文献数据库主题标引实践	张建清 (160)
文献的预处理是文献数据库质量控制的关键	赵桂云 (165)
资源科学主题词典编制研究	施慧中 (171)
中国稀土文献数据库特点剖析	张志升 (179)
中国计算机科学技术文献库的主题标引和服务工作	王能琴、刘 琪 (183)
RIS 系统的设计与实现	李百俊、张海泉 (188)
信息技术	(195)
数据库中的语音、图形和图象技术	厉森梁、肖念华、张春娥等 (195)
中国 1:400 万地图数据库系统研制	岳燕珍、李光荣、温凤艳 (201)
建在天文星表数据库上的天体图形处理系统	林钢华、柯大荣 (208)
附 录	(214)
ENGLISH ABSTRACT (英文摘要)	(214)
关键词索引	(230)

前　　言

中国科学院从 70 年代起, 就开始在化学领域研制数据库。由于历史原因, 化学领域数据库的数据多数来源于国外。1986 年受国家委托, 中国科学院承担了建立“科学数据库及其信息系统”的研制任务。该项任务要求数据库以面向科研、生产、教学服务为目的, 组织、协调科学院各类科学数据库和计算机化信息服务系统的建设, 其中包括数据的采集、整理、评价和加工, 制定有关标准, 统一建库规范等, 提高服务质量, 适应社会对现代化信息服务的需求。建库类型包括科技数值型、科技文献型和管理型数据库。建库内容以基础科学、应用科学和新兴技术科学为主, 注意发展其中应用性、技术性较强的数据库。

从 1987 年开始, 中国科学院已投资兴建了 23 个专业数据库。这些库主要涉及化学、天文、地学、生物和新技术等学科领域。参加建库任务的有我院所属 20 多个研究所的三百多名科研人员。

这些数据库的数据源, 大部分是我国科研、生产近几十年, 甚至上百年所积累的数据。一部分数据直接来源于中科院各研究所, 还有些数据来源于这些研究所与国内外高等院校和企业合作开发的项目, 因此数据更新与扩充的来源有保证。入库的数据是经过严格筛选的, 完整可靠, 它们都经过提供数据的研究所的学术委员会审定。重大发明中的数据, 还由发明者所在研究所报请国家有关专门机构审定。部分数据库是以引进数据磁带为基础建立的, 但是其中也引入了我国科研、生产所获得的一些数据。这些数据与国内外同行已开始进行交流。

建成的 23 个专业数据库在“七·五”期间均已通过了专家鉴定。为了进行学术交流, 提高建库队伍的水准, 科学数据库中心于 1991 年 9 月在西安临潼召开了首届“科学数据库与信息技术”学术研讨会。会议收到学术论文 63 篇, 这些论文涉及了科学数据库与信息技术发展的理论与规划、数据评估与质量控制、各类型数据库建设的方法, 范围广、领域宽; 又由于论文均出自于建库第一线的实际科研人员, 因此有一定深度。这些论文分别在大会和专业分组会上进行了充分的交流和讨论。为了在更大范围内作学术交流, 科学数据库中心特邀各学科专家组织成立了《科学数据库与信息技术论文集》编辑委员会, 请专家们对这些论文进行了仔细、认真的评审, 从中挑选了 36 篇论文汇编成本集。同时, 经科学数据库中心专家委员会讨论商定, 今后每两年举行一次“科学数据库与信息技术学术研讨会”, 届时出版一册论文集, 原《科学数据库通迅》不再出版发行。我们拟通过《科学数据库与信息技术论文集》的形式, 与国内外同行广泛进行学术交流, 以提高学术与技术水平; 集各方之良策, 为推动我国的信息产业发展尽微薄之力; 并使科技信息对我国经济繁荣和科技发展能有所裨益。

为了有效地向更大范围的用户提供服务, 充分发挥这些信息资源的效益, 在“八·五”期间建设科学数据库中心实体, 建立以综合性多学科数据库群体(包含数值型、文献型及事实型和多媒体)为核心、功能较强的信息系统, 该信息系统将运行在 IBM 计算机和 DEC 计算机系统上, 通过中关村地区科教示范网(NCFC), 借助先进的通讯手段, 使科学数据库作为网上的重要资源, 投入运行服务。同时, 还要装置各种服务设施, 为用户提供多种服务项目。

科学数据库及其信息系统工程的建设与发展,受到了中国科学院院长周光召、副院长胡启恒、孙鸿烈的亲切关照;在中国科学院计划局和技术科学局的直接领导下进行的。本书出版时,中国科学院老院长、学部委员卢嘉锡教授、中国科学院学部委员、技术科学部主任王大珩教授非常关心科技信息产业的发展,并题词祝贺,还有各建库单位的热情支持,我们在此一并表示感谢。

编 者

总 论

促进科技发展的现代化信息系统

罗百昌

中国科学院科学数据库首届学术会议论文集在各方面的支持下,终于和大家见面了。这在我们的建库工作中是件大事。它表明我们已迈出了艰难的第一步。我们不仅很好地完成了规定的建库任务,而且在工作中有所进步,有所提高,能从自己的心得体会中总结出规律性的东西来,这对科学数据库的发展是极为有益的事情。

科技信息是巨大的社会财富,科学数据库是将科技信息转化为生产力的重要手段。由于各种科技数据和文献的数量急剧增长,利用传统的工具和方法已不能有效地处理和传播,于是不得不求助于计算机和先进的通讯技术。从而大大推进了科学数据库及信息技术的发展和应用。

科学技术是第一生产力。因此科技信息理所当然地受到人们,尤其是科技人员的重视。主要的科技发达的大国都建立了自己的大型综合科技数据库。高效方便地向科技人员提供了大量可靠的科技数据和文献服务(通过文字、数据、图形、图像,甚至声音等不同方式)。由于这些数据库对科技的发展确实起到了很大的促进作用,科技数据库在这些国家得到大力支持和迅速发展。

我国的科技工作者早在 70 年代就已感到了建立科学数据库的必要性并开始筹建专业数据库(如工程化学库,红外光谱库等)。尽管当时还都是小规模的,分散的个别行动,但是已经在新产品试制,新工艺设计及相关领域的科学的研究中发挥了作用,个别情况下甚至起到了决定性的作用。它们带来的效益远大于它们所消耗的资源,已显示出了自己巨大的潜力。但是由于技术差,数据量少和其它一些问题严重地影响了这些库发挥更大的作用,也限制了它们的进一步发展。我国在这方面的落后状况势必对科学技术的发展造成不良影响。中国科学院院长周光召在国际科技数据协会中国分会的一次会议上就曾指出“当今科技发展非常迅速,竞争日益激烈,如果不能知己知彼,如果不能把科技界以前的工作成果整理出来,让每一个研究人员尽快知道并加以利用,那么我们将处于十分不利的地位”。这种紧迫感正在逐渐成为广大科技界的共识。今年 5 月 29 日《科技日报》的评论员文章,呼吁加强科学数据积累工作,建立全国科学数据网络和信息系统,并视之为支持科学技术发展的重要条件。

实际上科学数据库也是四化中科学技术现代化的重要环节,它应是国家科学研究事业的基础工程之一。

由于国家计委和中国科学院领导的关心和支持,中国科学院科学数据库在经费、设备等各方面的问题都有望得到了相应的解决,为建立有我国特色的大型综合科学数据库提供了良好的基础。但是有了好的条件还不等于有了好的数据库,我们还必须有数量大、质量高的各种科技数据和一个功能齐全的高效信息管理系统。由于科学数据库处理的信息不仅

是文字和数字,还有图形,图像和声音,而服务对象又是科技人员的研究和发展活动,因而和以往的以企业管理为主要用途的数据库有很多不同之处,它不得不使用多媒体(multimedia)和超文本(hypertext)等新技术,而且对数据库管理系统也提出了新的要求。这些不仅是科学数据库建设过程中需要解决的技术问题,也为信息科学的发展提供了很有价值的研究课题。国家科技事业需要和国际上信息处理技术的迅速发展都对我们这些从事科技数据工作的人员提出了新的课题。这不仅仅是一种要求,也是一种激励和鞭策。我们要努力提高自己的水平,更好地完成自己的任务。我希望我们的论文集能对此有所裨益。当然,我们的水平还有待进一步提高,但这却是一个可贵的开端,以后隔一定时间就要召开一次学术会议,总结经验,交流心得。我们的论文集将一集一集地出下去,每一集都将有所进步,有所提高,为我国的科技发展做出自己的贡献。

科学数据库及其信息系统工程一览表(七五期间)

序号	专业数据库名称		完成数据量	序号	专业数据库名称		完成数据量
1	工程化学数据库		97MB	11	中国微生物资源数据库		9MB
2	谱图数据库	有机化合物质谱数据库	80MB	12	中国经济植物资源数据库		7.34MB
		红外光谱数据库	60MB	13	中国药用植物数据库		7.1MB
		C-13 核磁共振图数据库	1MB	14	中国自然资源数据库		300MB
3	化合物命名与结构数据库		120MB	15	中国能源数据库		18.5MB
4	化学物质毒性数据库		120MB	16	天文数据库		1050MB
5	晶体结构数据库		40MB	17	中国光学镜头数据库		84MB
6	中国稀土综合数据库	稀土物性数据库	2.75MB	18	中国光学文献数据库		24MB
		稀土萃取数据库	3.60MB	19	中国计算机科学文献数据库		85MB
		稀土文献数据库	16MB	20	计算机市场信息数据库		4.8MB
7	X 射线光电子能谱数据库		8MB	21	中国腐蚀数据库		6.0MB
8	中国化学文献数据库		102MB	22	中科院科技经济信息数据库		40MB
9	核酸顺序与蛋白数据库		21MB	23	中科院科学家数据库		4MB
10	中国生物文献数据库		27MB				

现代数据库技术与科学数据库建设

罗晓沛

(中国科学院研究生院)

摘要：本文简介了当前数据库新技术中的分布式、多媒体和知识库技术以及它们对科学数据库建设的意义和作用。

关键词：数据库技术 科学数据库

科学数据库系统是一类收集、贮存与科学研究、技术发展相关数据（一般包括科技数据和科技文献）的数据库系统。科学数据库所贮存的数据及建造其上的应用软件系统，构成一个支持科学研究与技术应用的科技信息系统。因此，一个较大规模的科学数据库系统应该是由多种学科和技术领域中所收集到的正确和有价值的数据，存储和处理这些数据的设备以及处理这些数据的方法和相应软件所组成的一个完整的系统。

科学数据库虽然是一类不同于传统应用（它们多用于企事业管理）的数据库系统，但其建造方法和技术却可以继承和借鉴已经成熟的传统数据库的建造方法和技术，它们在诸多方面为科学数据库的建造提供了方法论、经验和技术支持。

尽管传统数据库系统主要面向企事业的数据管理，其管理系统的设计也主要是针对企事业的数据管理需求，但当前基于传统数据库技术的发展却又有了很多新的内容。由于各种应用需求的促进，多种技术的交互作用，计算机技术和外部设备的发展，使得当前的数据库已经朝着面向多种应用，管理多种复杂数据，支持多种存储媒体和提供具有一定智能的软件的方向发展，而这些成果又大多可为科学数据库及其信息系统的建造提供技术手段和支持，并对科学数据库及其信息系统的建造水平和质量起到促进作用。

当前数据库技术的发展，在下列三方面会对科学数据库的建设产生直接的影响。它们是：分布式数据库技术、多媒体数据库技术和知识库技术。

一、分布式数据库技术

80年代以来，数据库的系统结构有了新的变化。早期的数据库是集中式系统，即数据集中存储在一台计算机上，多个用户共享同一设备资源和数据资源。由一个单一的数据库管理软件，即数据库管理系统来进行集中管理和操作，由集中的控制来保证数据的一致性和完整性。但以下原因促进了数据库系统结构从集中式向分布式发展：

1.计算机网络技术的发展和应用，使得分散在不同地理位置上的计算机能够实现数据通信并达到数据库的资源共享。

2.由于社会和经济的发展，部门扩大并分散，传统的集中式数据库系统无法支持分散的应用。

- 3.数据库应用水平和需求的提高和发展，人们已不满足于对单一数据库的使用要求。
- 4.需要将现存的多种实用的数据库互联起来，以达到进一步的数据共享。
- 5.集中式的远距离的数据交换，将使通信费用增加，加重用户的经济负担。
- 6.集中式系统在系统的安全性方面存在较大的问题。

分布式数据库系统不是简单地将多个数据库分散设立，而是一个在物理上分散，在逻辑上统一的数据库系统。它根据分散在网络上各个节点的数据库所具有的特性，利用逻辑关系把各数据库结合起来，形成一个基于网络的数据库系统。数据存储、检索和管理都由一个称为分布式数据库管理系统(DDBMS)的软件来统一进行。

分布式数据库是分布在计算机网络的不同节点上的数据集合。网络上每个节点都具有独立的处理能力，即可执行局部应用。同时，每个节点借助于系统的通信子系统至少执行一个全局应用。

与集中式相比，分布式数据库具有本身的特殊性质，包括：节点自治、数据独立性和分布透明度、数据冗余以及并发控制等。

节点自治即节点具有独立处理能力。传统的集中式数据库强调对数据库的集中控制，分布式数据库却不再强调集中控制，而更多的强调整节点的独立处理能力。

数据独立性指的是对于应用而言，实际数据的组织是无需用户知道的。应用只考虑数据的逻辑组成，而不考虑数据的物理存放，即应用独立于数据的物理组织。分布式系统还增加了分布透明度的新概念，它指应用面对分布存在的数据，无需考虑数据的分布情况，就象在使用集中式数据库一样。

数据冗余最小，是对集中式数据库的基本要求，它有利于节省存储空间和保证数据的一致性。而对于分布式数据库来讲，由于节点自治概念的提出，它强调局部应用，因而在许多情况下，不但不消除冗余，而且有意识地保留一定的冗余数据，以达到提高应用效率，减少通信费用，并增加系统的安全保证等目的。

完整性、可恢复性和并发控制都是集中式和分布式数据库应解决的问题，它们是相互不同而又相互联系的问题，它们都与数据库运行中的事务不可分性有关，而处理上述问题时，分布式数据库环境下要比集中式数据库环境下困难和复杂。

分布式数据库是应用的需求和发展的方向。而科学数据库作为一类涉及多学科、广地域的大型数据库系统，从总体要求和最终目标看应该是一个具有分布式数据库特征的系统。多种学科的数据不可能，也不应该完全集中存放，虽然不同学科间的数据关联、用户对各种数据库的查询、应用需求可能不如面向管理的数据库的查询来得紧密，但其总体结构也应该是分布的，因而分布式数据库的技术成果在科学数据库系统的建设中会起到参考和借鉴作用。

二、多媒体数据库技术

80年代以来，支持多种媒体数据的存储和处理的数据库技术得到发展。多媒体数据库(MDBS)是相对于传统的仅支持单一媒体的数据库而言，是将图象、图形、文字、声音等多种媒体数据结合为一体并统一地进行存取、管理和应用。与传统的数据库系统相比，

MDBS 不仅在所处理的数据及其存贮形式上有很大的差别，而且它所面临的对数据的处理要求也比传统应用领域的事务管理环境复杂得多，因而多媒体数据库系统对 DBMS 提出了更多的功能要求，可以概括为以下几个方面：

1. 要求 DBMS 能方便地描述、存取和处理多媒体数据。在多媒体信息管理中，应用所涉及的数据除了传统的字符、数值数据外，还有图象、图形、语音等非常规数据。而这类数据在实际应用中数据量大、操作复杂，因而希望 DBMS 能提供对这类数据的描述、存取和处理机制。

2. 要求 DBMS 具有媒体独立性。这是指 DBMS 对多媒体数据的管理机制应与具体的多媒体数据无关。即无论多媒体数据库中数据的媒体形式作什么改变，都不应改变 DBMS。

3. 要求 DBMS 具有定义和处理动态数据模式的能力。在多媒体信息管理中，复杂的多媒体数据已难以用一个单一的、静态的数据库模式来描述，在多媒体实体之间，它们在构造上也不完全相同。这就要求 DBMS 具有描述动态可变的数据库模式的能力，以满足应用的需要。

4. 要求 DBMS 能刻划事物的时间属性并进而承担版本控制。在多媒体信息管理中，往往需要记录状态变化的时间、事件发生和某种变化或状态开始生效等，反映事物的时间属性并向应用者提供描述时间特性的手段，这是对 DBMS 的一个新要求。

同时在工程设计等领域，又要求保存诸如图纸、图片、文档等的不同版本，因而要求 DBMS 能进行版本控制。

5. 要求 DBMS 能提供对多媒体信息管理中特殊事务的支持。在传统的事务管理中，DBMS 处理的大多是几秒钟即能完成的事务，而完整性约束亦很简单，这时对事务并发、完整性、系统恢复的支持机制都相对比较简单。而在多媒体信息管理中，一次查询可能会引发对多媒体数据的一系列处理，这样就使得事务的并发控制、完整性约束、系统恢复的处理十分复杂，需要 DBMS 提供新的支持机制。

多媒体数据库管理系统(MDBMS) 就是介于多媒体数据库与用户之间，完成对多媒体数据库的管理、查询处理、数据处理等功能的系统软件。多媒体信息管理对 DBMS 提出的以上这些新的功能要求，需要我们去建立支持多媒体数据库的 MDBMS。

无疑，科学数据库中有众多的对象将涉及多种媒体数据，如动、植物标本库等，它们对客观世界所存在的对象能更生动地进行保存和查询，除了能浏览相对对象的文字描述，还可以观看对应的图形、图象，甚至听到其声音，典型的鸟类数据库就应该是这样一种类型的数据。

可以设想，提供多媒体数据存储和检索功能的系统将是一个更直观、更真实的科学数据库系统。研究多媒体数据库将涉及到它们的数据模型、存储技术和用户界面等技术领域，它们与传统的数据库技术内容相比有很大的不同，在技术上尚存在着某些难点，目前虽然有一些能支持多媒体技术的计算机系统，但在若干理论问题上（如支持多媒体的统一数据模型）还要再做出努力，当前还缺少商品化的多媒体数据库管理系统。但这些技术问题的存在并不妨碍我们对多媒体数据的应用，它可以从简单到复杂，从初级到高级，渐进发展，并在应用中完善和提高。

三、知识库技术

在 80 年代，作为数据库概念的拓广和衍生，知识库的概念得到发展和完善，知识库系统以及基于知识库的专家系统有了新的进展。知识库是数据库的拓广，指的是在知识库中同样可以存放数据，而且一些组织、管理和维护数据库的方法对于知识库同样可用，可以继承和参考；知识库是数据库的衍生，指的是知识库比数据库有更多新的内容、组织方法和处理类型。

数据处理中数据库起着重要的作用，而在知识运用中知识库起着重要的作用。知识库运用的重要方面是对知识的获取、组织、管理和维护，或统称知识管理。知识的管理包括对知识的分类、组织和存储，知识的检索，知识的增加、删除和更改，知识的复制和转储以及对知识的一致性、完整性和无冗余性的维护等。

知识分类的目的有二：一是对客观世界的更深入了解，便于组织和记忆；二是经分类后组织的知识便于检索和维护。

知识的一致性是保证所获得知识间不允许出现矛盾。一致性维护保证了当加入新知识或删除旧知识时知识集合的无矛盾性。

知识的完整性是指知识应该满足一些预先给定的约束。

知识的冗余性是指一个知识从某知识集合中被删除，既不缩小也不扩大由该知识的集合采用任何方法所能衍生出的知识的集合，即删除前后所能衍生出的知识集合相等。一般当知识集合很大时要保证知识的无冗余性是困难的。

一般认为，知识库的组织方式依赖于知识的逻辑表示形式，但都遵循某些基本原则：

1. 知识库的组织应保持知识库与处理机构的某种独立性，使知识库内部组织方式的改动不引起知识处理机构的大改动。
2. 要尽量便于知识库的扩充、维护和修改，便于功能和性能上对知识库进行改进。
3. 要尽量便于对知识库的运用和输入、输出。
4. 要提供对多种知识表示模式的支持。
5. 要便于进行对知识库内容一致性和完整性检查和维护。
6. 要考虑知识库运行的时、空效率。

从以上对知识库的简介可见，知识库的技术和方法以及它所能提供的应用功能，无疑对科学数据库的深入开发和应用有重要的意义。它可以将已有的科学的研究中所获得的数据（实际是一种知识）加以组织，再加以推理和判断，从而获得更多、更新的知识，专家系统即为应用的一例。将科学数据库与知识库、专家系统的开发结合起来将是一件十分有意义的研究课题和有价值的应用课题。

从长远的要求和应用的前景看，科学数据库系统应该是一个支持分布式应用，提供对多媒体数据的存储、检索和具有一定智能的数据库和知识库系统，当然这一目标是逐步达到的，并要在应用中发展和完善。