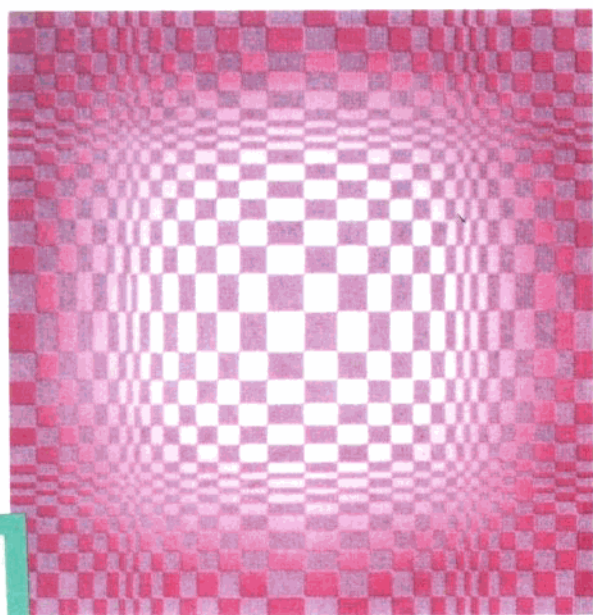


科学数据库 与信息技术论文集

第六集

● 中国科学院科学数据库办公室 编



中国环境科学出版社

科学数据库
与信息技术
PDG

37.21663

144.1

=6

科学数据库与信息技术
论 文 集
第六集

中国科学院科学数据库办公室 编

中国环境科学出版社

• 北京 •

图书在版编目 (CIP) 数据

科学数据库与信息技术论文集. 第6集/中国科学院
科学数据库办公室编. —北京: 中国环境科学出版社,
2002. 7

ISBN 7-80163-327-X

I. 科... II. 中... III. ①科学技术-专用数据-数据
库-文集 ②信息技术-文集 IV. TP392-53

中国版本图书馆 CIP 数据核字 (2002) 第 043157 号

中国环境科学出版社出版发行
(100036 北京海淀区普惠南里 14 号)
长城印刷有限公司印刷
各地新华书店经售

*

2002 年 7 月第一版	开本 787×1092	1/16
2002 年 7 月第一次印刷	印张 19.875	插页 0
印数: 1-1000	字数: 450 千字	

ISBN 7-80163-327-X/Z · 039

定价: 40.00 元

内 容 简 介

本书收入论文 52 篇,主要反映我国近年来科学数据库在建库技术、网络技术、信息服务、总体发展等方面所取得的成果,也反映了围绕着这些问题在学术上取得的进展。集中体现了近年来国内数据库与信息技术的研究和应用水平及发展历程。

本书可供从事数据库技术、网络技术和信息系统研究的科技人员、工程技术人员,以及相关学科的研究人员、大专院校师生参考。

编辑委员会

主任委员：胡启恒

副主任委员：阎保平 孙九林 王 源 马俊才

编 委：（以姓氏笔划为序）

许 禄 纪力强 庄大方 李天宪

李望平 肖 云 陈沈斌 陈维明

杨 锐 孟连生 徐予红 郭 杉

郭红锋 温 浩

第六册编辑部

主 编：李望平

编 辑：单雪明 钱云杉

序 言

中国科学院科学数据库及其信息系统经过近 20 年的建设与发展,至今已经成为国内信息量最大,学科覆盖面最广,综合性最强,服务功能比较完善的科学信息服务系统。至 2001 年,科学数据库共有专业数据库 180 个,总数据量达 7250 亿字节 (725GB)。这批宝贵的信息资源已在国家经济建设、国防建设、规划决策、科学研究、科技攻关、学科发展、国际合作等诸多方面得到应用,取得了显著的社会效益和一定的经济效益,在国内外产生了一定影响。

目前,科学数据库作为中国科学院信息化建设的重要基础设施,已经纳入了中国科学院“十五”期间信息化建设的重大项目。科学数据库在“十五”期间的总体发展目标是在继续扩大数据资源的基础上,促进数据向知识的转化,完善信息化的科学研究环境,初步建成面向科学研究和社会的科技信息服务体系。其中,数据资源建设方面,要继续加强科学数据资源的采集和积累,进一步扩大现有专业数据库学科覆盖范围,实现建库单位达到 40 个,专业子库数量达到 300 个,总数据量达到 10TB;标准规范建设方面,要研究、制定科学数据库相应的标准与规范,形成其特有的标准规范体系,用以规范数据资源的建设、数据共享与服务,保证数据资源的质量,提高科学数据库的管理和服务水平,为科学数据库的持续发展奠定良好的基础;系统平台建设方面,要运用 IT 技术的新进展,对大规模、分布式、异构的科学数据库中的海量数据资源进行整合,实现全方位、深层次的资源共享,并在高性能环境的支持下开发基于科学数据库的先进应用系统,使科学数据库的发展能够满足未来科学研究的需要,使科学数据库能够成为未来信息化科研环境的核心组成部分之一,为建设信息化的科学研究环境提供数据资源保障,为知识创新服务。因此,在今后的几年内,我们的任务光荣而艰巨,有很多新的技术需要我们去探讨,有很多新的问题需要我们去解决,还有大量的工作要我们去完成。

2002 年 8 月在新疆召开的第六届“科学数据库与信息技术”学术讨论会,主要是进一步研讨自 2000 年第五届学术讨论会以来,在科学数据库建库技术、网络技术、信息服务技术,应用服务、总体发展等方面的进展和取得的成果,以及对科学数据库今后发展相关的新技术、新思路的探讨。这次会议收到的论文 63 篇,经编辑委员会认真评审,选出了 52 篇论文进行会议交流并收入了本次会议的论文集。这些论文集中体现了近年来国内数据库科学与信息技术的研究和应用水平,可以提供从事数据库及信息系统研究的同行参考。

希望科学数据库在今后的发展过程中,为国家的经济建设做出更大的贡献。

中国科学院科学数据库专家委员会主任

胡启恒

2002 年 4 月 20 日

目 录

序言

总论

科学数据库“十五”期间的数据资源建设	李望平, 阎保平 (1)
科学数据库系统平台建设的设想	南 凯, 阎保平 (7)
推进科学数据共享, 增强知识创新能力	孙九林 (14)
论科学数据库的关联性	王 源, 陈维明 (23)
基于互联网的科学数据共享系统研究	廖顺宝 (30)
数字图书馆时代在线服务信息组织研究	丁小文 (35)
CODATA 中国理化数据库系统的建立	阎保平, 肖 云, 张 辉, 金化年 (41)

数据库系统及建库技术

科学数据库中基于内容的图像检索技术研究及其试验系统	唐俊华, 阎保平 (47)
科学数据库系统平台元数据系统设计中的两大问题与对策	马永征, 阎保平 (52)
Native XML 数据库技术初探	杨德娣, 阎保平 (58)
实现多媒体数据库管理系统的几个途径初探	董科军 (64)
Distributed Parallel Indexing for SRS	一柳芳浩, 马俊才 (70)
数据仓库技术及其在科学数据库中的应用	替振宇, 马 骊 (76)
化学标记语言及其在分子结构数据库检索中的应用	李 琰, 温 浩, 周家驹 (84)
大气科学数据的元数据标准研究	徐予红 (93)
确定化学结构微机检索系统	陈维明, 孙传涛, 朱翠娣, 徐衍波 (99)
COM 技术及其在活度系数模型参数估算中的应用	郑建龙, 温 浩, 张和珍 (104)
族性化学结构的色性匹配	徐衍波, 孙传涛, 朱翠娣, 陈维明 (110)
光学镜头数据库数据的分类与数据档案的建立	丛小杰, 张 新, 任 涛, 常 军, 张增宝, 翁志成, 姜会林 (115)
基于 COM 技术的湿地信息系统设计	张树清, 何福久, 姚 敏 (121)
矿山地质科学数据库系统初探	吴堃虹, 李宏辉 (127)
C-13 NMR 谱图数据库相似检索的网上实现	罗操操, 齐玉华, 许 禄 (132)
国家水土保持基础信息设施建设的现状与问题	杨勤科, 郭明航, 韩 琳 (136)
XML 在化学数据库中的应用	罗操操, 许 禄 (140)
应用元目录技术建设专家信息库	周 宁, 赵淑玉, 韩秀芝 (144)

Internet/Intranet 相关技术

利用 ASP 实现分布式异构数据库共享服务的几点技术探讨	张耀南, 谭振华, 韦五周 (148)
科学数据库系统平台的安全认证机制的探讨	王东安, 阎保平 (154)
网上分布式数据库智能集成检索引擎	马俊才, 一柳芳浩 (161)
一种可以在数据库系统中应用的网络并行计算平台	王鹏飞 (169)
smart—一种高效的 web 应用服务器	吴开超, 何洪波 (176)

连接池技术研究及其在科学数据库上的应用	沈志宏,周宁,王龙潇(184)
软件测试在 SDB 系统平台开发中的应用	李树仁,吴开超(190)
数据库服务器流量控制和流量统计	王鹏飞(196)
流媒体技术探讨以及在“中国科普博览”网站中的应用	褚民学(200)
网站设计与网页制作技术	陈立立(205)
文献数据库及全文检索技术	
应用软件服务器在图书馆数字化资源利用中的应用	秦幸昌(211)
语义互联网的基础——后控词表的研究	赵英莉,王源,陈维明(215)
各种文献信息的 Web 整合	贾红阳,梁春燕,郭力,李晓霞,杨章远,许志宏(221)
数据库应用	
光学镜头数据库在多光谱遥感成像中的应用	翁志成,丛小杰,常军,姜会林(227)
从天文数据库到虚拟天文台	郭红锋(233)
湖泊——流域数据库在湖泊变迁研究中的应用	高俊峰,赵锐,施晶晶,史岚(239)
科学数据库在精准农业中的应用	陈沈斌(243)
数据库技术在中国农业宏观经济决策中的应用	刘志强,赵军,韩晓增(249)
东北农业生态数据库管理系统的建立与应用	赵军,刘志强,孟凯,于莉,张庆平(256)
中国自然资源数据库在气候变化影响研究中的应用——黄土高原农业生产的气候脆弱性研究	刘文泉,李泽辉,王颖棠(262)
农业增长贡献率的计算——中国自然资源数据库的应用	陈安宁,李泽辉(269)
网络型河流水污染信息系统及初步应用	王云鹏,王岩,夏昊(275)
云南省黄金资源地质信息库及其应用	黄怀勇,张洪恩,张湘炳,陈广浩(280)
中国动物物种编目数据库及其应用	伍玉明,贺大为(284)
CO ₂ 倍增对中国自然带及自然地区分布的影响研究	卢显富,孙九林,李泽辉(288)
长江流域脊椎动物数据库	伍玉明(295)
数据挖掘与电力系统的信息化	柳进,周苏荃,唐降龙(298)

科学数据库“十五”期间的数据资源建设

李望平 阎保平

(中国科学院计算机网络信息中心, 北京 100080)

摘要 本文介绍了科学数据库“十五”期间的数据资源建设。包括目标与任务, 建设内容, 组织管理, 项目实施等内容, 以及今后的发展方向。

关键词 科学数据库 数据资源 建设

一、引言

20 世纪 50 年代, 由于计算机技术的发展和成熟, 收集加工储存和应用大量数据成为可能, 并决定了数据的价值。中国科学院作为中国自然科学的研究中心, 在长期的科学研究实践中产生和积累了大量具有重要科学价值和实用意义的科学数据和资料。60 年代发展起来的数据库技术, 为有效管理和开发应用科学数据创造了有利条件。70 年代, 中国科学院的一些研究所开始在自己学科领域中建立专业数据库, 1983 年中国科学院开始组织建设“科学数据库及其信息系统”。近 20 年来, 随着计算机技术的进步和用户需求的不断变化, 系统也不断改进和完善。科学数据库的研制和建设过程是知识积累、加工、利用和传播的过程, 是跟踪数据库技术、网络技术发展和应用的过程, 是数据资料逐步规范标准和积累的过程, 也是广大系统建设者与用户不断沟通, 向实用化发展的过程。到“九五”末期, 科学数据库共有专业数据库 180 个, 总数据量达 725GB。科学数据库专业库分布在中国科技网上站点已达 19 个, 网上 Web 可检索专业数据库达 153 个, 可检索总数据量达 325GB。目前, 科学数据库由中心站点和分布在网上海本地和外地的相互独立的若干个专业库子站点组成网上服务体系, 已经成为国内信息量最大, 学科专业最广, 服务层次最高, 综合性最强的科学信息服务系统。多年来, 科学数据库的研制者利用这批宝贵的数据资源进行了大量的专业咨询服务、专业应用服务和网上综合科研信息服务, 已在国家经济建设、国防建设、规划决策、科学研究、科技攻关、学科发展、国际合作等诸多方面得到应用, 取得了显著的社会效益和一定的经济效益, 在国内外产生了一定影响。

科学数据库经过 20 年的建设和发展, 已经成为科研工作的基础设施之一, 并将成为未来科学研究必不可少的资源。目前, “科学数据库及其应用系统”作为中国科学院信息化建设的重要基础设施, 已经纳入了中国科学院“十五”期间信息化建设的重大项目, 给科学数据库的建设和发展带来了一个有利的时机, 科学数据库将在“十五”建设期间得到长足的发展, 为社会做出更大的贡献。

二、目标与任务

科学数据库在“十五”期间的总体发展目标是在继续扩大数据资源的基础上, 促进数据向知识的转化, 完善信息化的科学研究环境, 初步建成面向科学研究和社会的科技信息服务体系。其中, 数据资源建设, 标准规范建设和系统平台建设分别确定了明确的建设目标,

用以保证总目标的实现。在加强科学数据资源的采集和积累,扩大现有专业数据库范围的同时,研究并制定科学数据库相应的标准与规范,用以规范数据资源的建设、数据共享与服务。运用 IT 技术的新进展,对大规模、分布式、异构的科学数据库中海量的数据资源进行整合,实现全方位、深层次的资源共享,并在高性能环境的支持下开发基于科学数据库的先进应用系统,使科学数据库的发展能够满足未来科学研究的需要和建设信息化科研环境的需要。

科学数据库资源建设的总体目标是继续加强科学数据资源的采集和积累,扩大现有专业数据库范围,注重数据的完备性,系统性,内容和形式的多样性,在全院范围内逐步建立学科门类完整的专业数据库群体,为建设信息化的科学研究环境提供数据资源保障,为知识创新服务。

在“十五”末期,科学数据库资源建设的预期成果是在全院范围内初步建成包括化学、天文学、材料科学、能源科学、海洋科学、生物学、地球科学、信息科学等学科门类基本完整的科学数据库群体,使科学数据库系统的建库单位的数量接近 40 个,学科门类逐步趋于完整;专业子库数量达到 300 个,专业库种类包括数值库、事实库和多媒体库;科学数据库总数据量达到 10TB。同时,建成 1~2 个数字标本馆应用实验系统。

科学数据库资源建设在“十五”期间的主要任务有以下两个方面:

(1) 专业数据库建设

科学数据库“十五”期间资源建设从应用服务的角度出发组织专业数据库的建设。侧重于从学科领域整体规划数据库的发展,组织学科领域标准规范的制订和实施,注重数据的完备性,系统性,内容和形式的多样性,在全院范围内逐步建立学科门类完整的专业数据库群体,为建设信息化的科学研究环境提供数据资源保障。

• 科研数据的采集与积累

将实验室、科研课题项目等科研活动中积累的科学数据和成果收集、整理、加工、建库上网,提供数据共享与服务。

• 国外数据的合作交换

通过与国外同行科学数据的合作交换,建立互利互惠的合作关系,扩大专业领域的数据资源。

• 建立国外数据库网上镜像节点

根据科学研究的需求,联系专业领域内国际上最权威的网上免费数据服务系统,在我国建立生物、天文等相应的网上镜像节点,实现在国内网上提供全球性科技信息资源共享。

(2) 数字标本馆建设

中国科学院在生物和地学领域的研究所中分布着 24 个实物标本馆,其中相当一部分标本馆在国际上有很高的地位。数字标本馆是将原始实物标本系统地转化整理成易于传递的数字化电子信息,将标本在网上进行实物展示,具有很大的方便性和灵活性。数字标本馆的建设是一个新的探索,由于标本馆标本数量多,经费投入大,国内外尚无成功的范例。如何三维、直观和综合有效地展示标本的技术问题还有待探讨和解决。

“十五”期间,将在生物和地学领域选择 1~2 个标本馆,进行数字标本馆建设的试点工作。

三、建设内容

中国科学院现有一半以上的研究所在科学研究活动中都不同程度的建立了科学专业数据库,其中相当一部分研究所具有很好的建库基础和经验。这些专业数据库有自建的,也有引进的,内容涵盖了化学、物理、天文、材料、能源、海洋、生物、地学以及信息技术等多种学科,数据库类型包括数值、事实、多媒体等多种类型。这些科学数据库资源都是由研究所科研工作的需求产生,应用范围各不相同,总数据量约 20~30TB。如果将其中应用范围广泛,并能提供公用的科学专业数据库纳入科学数据库系统加以组织、利用和共享,从学科领域整体规划数据库的发展,提供专业领域的综合科技信息服务,将会取得更加明显社会效益和经济效益,使这些宝贵的信息资源更好地发挥作用。

科学数据库在 2001 年底启动了“十五”期间的第一批专业数据库的建设:

(1) 化学

工程化学数据库,化学专业数据库,应用化学数据库,化学物质毒性库;

(2) 天文学

天文数据库;

(3) 材料科学

材料数据库,光学镜头数据库,纳米材料数据库;

(4) 能源科学

中国能源经济数据库;

(5) 海洋科学

南海海洋数据库;

(6) 生物学

中国植物数据库,植物图谱数据库,中国动物数据库,中国西南地区动物资源库,中国微生物资源数据库,病毒数据库,中国核酸序列数据库,中国水稻基因数据库,人类基因组多态性数据库;

(7) 地球科学

自然资源数据库,全国资源环境遥感数据库,中国遥感卫星地面站检索库,中国寒区旱区特色数据库,中国湖泊数据库,中国土壤数据库,中国沼泽数据库,大气科学与环境数据库,大地构造图件数据库,中国山地环境与灾害数据库,中国东北农业生态环境数据库,新疆资源生态环境数据库,中国地质数据库,黄土高原基础数据库;

(8) 其它

中国科学院科学家数据库。

2002 年将启动“十五”期间的第二批专业数据库的建设

四、组织管理

科学数据库建设是中国科学院全院范围内的跨所联合,力图在全院范围内将研究所多年积累起来的科技信息资源组织起来,与计算机、数据库和网络等先进技术相结合,对社会提供科技资源共享与服务。科学数据库的组织形式是以综合型重大工程项目的方式进行的。从十几年来科学数据库的建设和发展的实践经验来看,这种组织方式是行之有效的。

科学数据库的组织管理结构采用以下方式:

(1) 科学数据库专家委员会

专家委员会是科学数据库的学术领导机构,在院的领导下,负责中国科学院科学数据库有关重大问题的决策和管理。负责制定科学数据库的发展规划和实施计划,确定经费的具体资助范围和原则,审定资助的项目及经费分配方案,检查监督项目的执行情况。课题成果的评审、鉴定、验收,由专家委员会组织进行。

(2) 科学数据库办公室

科学数据库办公室是专家委员会的办事机构,设在中国科学院计算机网络信息中心。课题的内容、目标和经费经专家委员会审定立项后,由科学数据库办公室按照院重大工程项目的管理办法进行管理。课题合同的检查验收,由科学数据库办公室组织进行。

多年来,科学数据库在科学数据库专家委员会的领导下,由科学数据库办公室进行课题管理,各建库单位负责对本单位的建库课题进行组织实施。科学数据库在组织实施过程中,实现了跨学科的联合和数据、人才、技术的集中,体现了中国科学院的整体优势,在建立中国科学院和国家科研工作的基础体系中起到了龙头的牵引带动作用,提升了专业数据库的技术能力,综合能力和整体服务能力。

五、项目实施

(1) 资源调研

2001 年科学数据库办公室在全院范围内通过发函和实地考察进行了科学数据库资源调研工作。考察调研以实现“十五”期间科学数据库的主要目标为原则,学科领域布局为重点,调查对象以研究所为单位,重点考察各研究所科学数据资源积累情况,现有建库工作基础,专业数据库内容、性质与意义,服务对象,应用领域,建库负责人与建库队伍情况,现有网络环境、数据库硬件及软件环境情况,系统功能和今后的可持续发展等。

科学数据库建库单位的考察和挑选重点考虑了其学科专业布局和学科地区布局,数据资源积累与今后的延续性,建库工作基础,服务对象和应用范围,建库负责人与建库队伍情况等。经过认真的考察和论证,在原有 21 个建库单位的基础上新增选了金属所,南海海洋所,武汉植物所,昆明动物所,武汉病毒所,上海生命科学院,国家基因中心,生物物理所,遥感卫星地面站,南京土壤所,长沙大地构造所,成都山地灾害与环境所,黑龙江农业现代化所,新疆生态与地理所,广州地球化学所,水土保持所等有科学数据资源和技术力量基础的十几个研究所加入科学数据库系统予以支持。从而,科学数据库的学科覆盖面更加广泛,数据资源更加丰富。

(2) 投资原则

科学数据库“十五”期间资源建设的进一步发展从应用服务的角度出发,在学科布局和数据资源的完整性方面进行了调整。同时,强调了以实现数据共享作为资源建设的主要投资目的,要求共享数据量达到 80% 以上。

资源建设的组织方式以建库单位为课题单元,投资方式分为专业领域布局的核心库、重点库和一般库三个层面。

核心库要有很好的建库基础和经验,具备一支专业建库队伍,具备先进的网络和计算机环境,能够承担引入本学科领域内国外科技数据库站点的镜像,能够在“十五”期间完成 TB 级数据量的建库任务,能够从整体上规划本学科领域数据库的发展,组织实现本学

科领域数据共享, 牵头承担本学科领域数据标准规范的制订和实施, 承担学科领域数据分中心的系统平台建设任务, 并从技术上成为该专业领域建库单位的技术支撑, 提供专业领域的综合科技信息服务。在化学、生物、地学等学科领域中各选择一个建库单位承担建立该学科领域的核心库的建设, 并作为科学数据库资源建设的重点投资对象。

重点库是根据学科领域重点发展布局确定。要求符合专业领域学科重点布局 and 区域代表性, 有较好的建库工作基础, 有一定量级的数据积累和延续性, 有一支稳定的建库队伍, 专业数据库应用面广。拟在科学数据库专业建库单位中, 选择四分之一的建库单位作为学科领域发展的重点库, 给予一定的投资强度支持。

一般库要求符合学科专业布局和学科地区布局, 有一定的数据资源积累和延续性, 有一定的数据积累和建库工作基础, 建库队伍情况良好, 可提供数据共享并有一定的应用面。一般库约占全部建库单位的 2/3, 给予一般性的投资强度支持。

(3) 管理方式

• 专家论证

科学数据库课题的立项, 要经过科学数据库专家委员会的论证和答辩, 确保课题实施的可行性。

• 目标管理

经专家论证立项的课题, 由科学数据库办公室负责组织签订《中国科学院知识创新工程信息化建设专项任务书》, 确定项目内容和目标, 实行目标管理。

• 过程控制

为了有效地促进数据库资源建设的发展, 实现“十五”期间设定的目标, 科学数据库办公室将按照签订的合同指标, 严格进行建设过程的控制。每年对课题组织年度工作检查, 进行过程控制。一旦发现课题执行过程中的问题, 及时提交专家委员会决策解决。

• 滚动支持

科学数据库资源建设课题的管理, 实行滚动支持。鼓励先进, 淘汰落后, 以便促进科学数据库事业的发展。

六、结 语

“十五”期间, 科学数据库资源建设将有一个高速的发展, 数据资源愈来愈丰富, 综合服务能力愈来愈强, 运行环境愈来愈先进。通过标准规范和系统平台的建设, 科学数据库与高速网络、超级计算机结合, 将真正成为科学研究的支撑, 并构筑出信息化的科学研究环境, 为中国科学院知识创新工程, 为国家创新体制, 为国民经济建设, 为我国的高科技的发展做出更大的贡献。

Construction of Data Resources for SDB During 10th "Five-year Plan" Period

Li Wangping Yan Baoping
(Computer Network Information Center,
Chinese Academy of Sciences, Beijing 100080)

Abstract This paper gives a description of the further construction of data resources for the Scientific Database during the 10th "Five-year Plan" period, including the goals, tasks, contents, organizational management, project implementation and development orientation in the coming years.

Key words Scientific Database, data resources, construction

科学数据库系统平台建设的设想

南 凯 阎保平

(中国科学院计算机网络信息中心, 北京 100080)

摘 要 建设科学数据库系统平台是科学数据库发展的需要, 已被列为“十五”科学院信息化建设项目“科学数据库及其应用系统”的重要建设内容。本文就科学数据库系统平台的意义、目标、技术路线等问题提出一些设想, 概要介绍了以数据网格技术为核心, 建设科学数据库系统平台的思路 and 计划。

关键词 科学数据库 系统平台 数据网格

一、引 言

中国科学院科学数据库经过近 20 年的发展, 已经积累了相当可观的数据量, 大部分数据实现了基于 Web 的数据共享, 并在此基础上开发了一些应用系统。到“九五”末期, 海量数据的存储与处理问题, 学科领域中的先进应用系统对高性能计算资源、高速网络环境的需求, 都日益突出, 成为科学数据库中一个有共性的重要问题。基于对上述问题的认识和考虑, 我们产生了建设科学数据库系统平台的设想。

世纪之交, 随着整个国家综合国力的蒸蒸日上和科技事业的蓬勃发展, 科学数据库也赢得了宝贵的机遇, 开始新的腾飞。作为中国科学院“十五”信息化建设的重点项目, 作为科学院信息化建设中的重要基础设施, 科学数据库在“十五”期间的总体发展目标定位为在继续扩大数据资源的基础上, 促进数据向知识的转化, 完善信息化的科学研究环境, 初步建成面向科学研究和社会的科技信息服务体系。为了实现这个宏伟目标, 科学数据库系统平台的建设将发挥至关重要的作用。本文将介绍我们在科学数据库系统平台的意义、目标、技术路线及实施计划等方面的主要设想。

二、科学数据库现状分析及系统平台建设的意义

中国科学院科学数据库始建于 1983 年, 至今已成为国内信息量最大、学科专业最广、服务层次最高、综合性最强的科技信息服务系统。到“九五”末期, 科学数据库共有专业数据库 180 个, 总数据量达 7250 亿字节 (725GB), 在中国科技网上建立 Web 站点 19 个, 上网专业数据库 153 个, 数据量约 3250 亿字节 (325GB)。

毋庸置疑, 经过十多年的发展, 科学数据库已经取得了相当的成就。然而, 面对信息技术日新月异的发展进步, 以及应用领域新需求的不断涌现, 科学数据库也面临着新的挑战。从技术角度来看, 主要有以下五个比较紧迫的问题。

首先, 随着数据量的不断增长, 对海量数据的存储、处理和应用都产生了新的问题。对科学数据库而言, 专业子库的数量在增长, 每个子库的数据量增长得更快, 特别是某些十分活跃的学科领域, 如生物学, 数据量的积累速度更是惊人。十五期间科学数据库数据资源建设预计将达到 300 个专业子库, 10TB 的数据总量。虽然说这与国际上前沿领域中

PB 级的数据规模还有相当差距,但在大数据量的存储、处理、有效管理及应用等方面已经感受到不少问题和压力。因此,超大规模的数据量所带来的问题是我们必须解决的第一个技术难点。

其次,到“九五”末期,科学数据库共有 21 个建库单位,这些研究所分布在全国各地。而且,跨学科的 180 个专业库无论从数据类型,还是数据库管理系统上,都是异构的。这种分布而又异构的特点在“十五”期间必然会继续存在,甚至可能加强。所以,如何对科学数据库分布而又异构的数据资源进行整合,实现统一、透明、高效的访问,是一个非常重要的问题。这对于科学数据库数据资源的利用,科学数据库先进应用系统的开发,乃至科学数据库能否成为未来信息化科研环境中的重要的信息资源,都具有举足轻重的意义。

第三,是数据的规范化以及元数据的问题。信息交换是信息社会的基本要素,为了确保科学数据库中的数据资源能够方便地进行交换,有效地开发利用,数据的规范是必需的。虽然我们已经在在这方面做了一些工作,但还远远不够。另外,元数据技术是这方面的一个核心技术,科学数据库的元数据工作是“十五”期间的重要工作。实现科学数据库的数据规范化与元数据化,是对科学数据库进行有效的数据整合,达到统一、透明、高效访问这样一个目标的前提和基础。

第四,数据资源与计算资源、通信资源等其他资源的密切关系。今天,对大规模数据做处理,必然需要高性能计算资源,在数据分布的情况下还要有高性能通信环境,在数据可视化以及多媒体数据等方面还会要求其他资源,等等。总而言之,对于基于科学数据库的先进应用系统来说,需要有一个集成了上述各种资源的高性能环境的支持,这也是一个亟待解决的重要问题。

最后,是基于数据的协同工作问题。资源共享是科学数据库的科技信息服务体系中最重要基础内容,而全方位、深层次的资源共享必然会演变为协同工作的问题。如何创造这样的资源共享与协同工作环境,涉及到网络通信、信息安全等多方面的技术,随着科学数据库的发展应在“十五”期间着手解决这个问题。

考虑到上面提到的这些问题以及科学数据库未来的发展,我们提出了建设科学数据库系统平台的设想。所谓“科学数据库系统平台”应该是一个综合的软硬件环境,它集成了海量存储、高性能计算、高速网络等各种资源,由一套系统软件和中间件软件体系提供对分布式、异构数据资源的共享、处理和管理服务,为科学数据库信息资源的积累、开发和应用提供支持。可以说,科学数据库系统平台是从信息技术的角度,为科学数据库的发展提供全面支持的一个支撑环境。随着信息技术的不断进步,这个系统平台的内容和功能也会相应的发展。

科学数据库系统平台的意义主要有以下三个方面。第一,科学数据库是一个跨越多种学科、由几十个研究所共同建设的超大型综合数据库,建库的人员大多是各学科领域中的专家,在信息技术方面各单位的情况各不相同,很多时候因为计算机技术、网络技术等方面的问题,阻碍了专业数据库更快更好的发展。因此,通过建设科学数据库系统平台,我们将对科学数据库中主要的共性问题提供统一的支持,让科学数据库中的各专业子库都能充分利用信息技术的最新成果。

第二,科学数据库虽然是由众多不同学科的专业子库构成,但它应该是一个有机的整体。如何对科学数据库中五花八门的数据资源进行有效管理和统一访问,是科学数据库发

展到今天所急需解决的问题。科学数据库系统平台的建设正是在朝着这一目标努力,相信必将大大促进科学数据库的资源共享与开发利用。特别是对于需要同时访问不同学科专业子库的应用,没有系统平台的支持很难实现或者效率很低。

第三,科学数据库的建立和发展都是建立在数据库技术及其他计算机技术的基础上,相关技术的进步无疑将有力地促进和牵引科学数据库的发展。科学数据库系统平台的建设采用信息技术的最新成果,并将持续跟踪新的进展,这一方面是为了满足科学数据库先进应用的需求,另一方面,信息领域的新技术也会刺激和启发科学数据库新应用的出现。

总而言之,科学数据库系统平台是科学数据库发展到一定阶段的必然需要,也是信息技术新成果的产物。我们相信,科学数据库系统平台的建设必将为科学数据库的发展起到重大的推进作用,为科学数据库“十五”战略目标的实现做出重大贡献。

三、科学数据库系统平台建设相关的关键技术

科学数据库系统平台的建设涉及到信息技术的许多方面,这里只简要介绍几个重要的关键技术:元数据技术、XML技术与数据网格技术。

元数据是关于数据的数据,它描述了数据的重要属性,对于有效的管理和使用数据具有关键性的意义。元数据的工作主要有两个方面,一是元数据标准的制订,二是元数据的生成、使用与管理。目前,科学数据库中的元数据工作还比较欠缺,是我们将在“十五”期间重点突破的一个问题。元数据技术是实现科学数据库资源整合所依赖的一项核心技术,对于科学数据库系统平台的实现也具有关键性的作用。

XML作为一种可扩展性标记语言,以其很强的开发性日益受到青睐。目前XML正在成为各种数据特别是文档的首选格式。其自描述性使其可充当公共传输工具,非常适用于不同应用间的数据交换,而且这种交换是以不预先规定一组数据结构定义为前提。由于XML具有标记不同字段的能力,使得搜索变得更简单和动态化。XML把内容从显示格式中解放出来,使文件可以多次重复使用,满足不同的应用需求。此外,XML还可以处理各种数据,包括文本、图像和声音,并且可以由用户进行扩展以处理任何特殊类型的数据。我们将在“十五”期间的科学数据库建设中大量使用XML技术。

网格技术是近年来迅速发展并趋于成熟的一种新技术,它主要解决分布式、跨组织的资源共享与协同工作的问题。网格技术提出的初期,主要针对高性能计算资源的共享;而数据网格技术则是在网格技术的基础上,专门解决大规模、分布式、异构的数据资源共享问题。可以看出,数据网格技术对于科学数据库系统平台具有特殊重要的意义,它为科学数据库当前所面临的一些主要问题提供了建设性的思路。或者说,正是科学数据库这一类应用项目的发展及其所带来的问题,刺激并推动了数据网格技术的产生和发展。因此,我们将以数据网格技术作为“十五”期间科学数据库系统平台建设的核心支撑技术。

四、科学数据库系统平台的总体目标与主要任务

科学数据库系统平台建设的总体目标是运用信息技术的新进展,对大规模、分布式、异构的科学数据库中海量的数据资源进行整合,实现全方位、深层次的资源共享,并在高性能环境的支持下开发基于科学数据库的先进应用系统,使科学数据库的发展能够满足未来科学研究的需要,使科学数据库能够成为未来信息化科研环境的核心组成部分之一。在“十五”期间的主要任务是建设面向科学数据库未来五年发展所需要的软硬件支撑环境;