中国科技资源导刊 ISSN 1674-1544 CHINA SCIENCE & TECHNOLOGY RESOURCES REVIEW
2008年5月 第40卷第3期 6-11 ISSN1674-1544 Vol.40 No.3 6-11,May,2008

分散数据资源整合策略和模式研究

孙九林
(中国科学院地理科学与资源研究所,北京 100101)

摘 要:科学数据资源通常可以分为两类,一类是行业部门产生的科学数据;一类是科研项目研究过程和结果中产生的分散科学数据。本文通过对地球系统科学数据的特点及"地球系统科学数据共享网"的分析研究,总结出分散科学数据资源的十种整合策略以及不同的整合模式。

关键词:分散数据;地球系统科学;科学数据;数据共享;资源整合
中图分类号: G30 **文献标识码**: A DOI: 10.3772/j.issn.1674-1544.2008.03.002

Tactics of Dispersed Data Resources Integration and Model Research

Sun Jiulin
(Institute of Geographic Science and Resources Research, CAS, Beijing 100101)

Abstract: Science data resource is generally divided into two categories, one is the science data produced by professional units, the other is the dispersed data resulted in the process of scientific project research. Through the characteristics of earth system science data and the analysis of "earth system science data sharing network", this paper summed up ten integration tactics and different integration models of dispersed science data resources.

Keywords: dispersed data, earth system science, science data, data sharing, resources integration

科学数据资源通常可以分为两大类,一类是行业部门通过统一的规范标准产生的科学数据资源;一类是科研项目过程和成果中产生的科学数据资源。在国家实施的科学数据共享平台建设中,两类科学数据资源都需要在资源整合过程中实现共享。由于这两类数据资源所产生的背景不同,它们在整合集成中所采取的方式也各有不同。

行业部门产生的科学数据资源,一般在产生过程中遵循较完善的规范或标准,同时在行业内部有较强的行政机制进行管理和规范,所以这类数据资源的共享可以通过行政系统的管理职能

第一作者简介:孙九林(1937-),男,江苏盐城人,中国工程院院士,研究方向是地球信息科学研究。

收稿日期:2008年5月12日。

实现；在科研项目过程和结果中产生的科学数据资源，缺乏统一的标准和规范，而且很少进行过系统的整理与提供应用，大部分还分散在科研院所和科研组或科学家手中，而这些分散的科学数据资源要进行整合集成与共享服务，不可能依靠行政系统管理职能的作用来实现。对于共享分散科学数据资源的有效整合，需要考虑的问题较多：例如，分散数据在网络中实现共享后，科学家们能否从共享网中获取在科学研究中他们需要的数据，即能否实现共赢？个人手中的科学数据实现共享后，能否保证数据的安全？单位或个人提供的数据资源在数据实现共享过程中，如何体现原数据拥有者的权利？在整合、集成多学科的分散数据资源，实现共享的过程中，采用什么样的规范和标准？等等。

“地球系统科学数据共享网”主要整合集成分散科学数据资源，实现共享。它在整合和集成过程中，必须根据数据共享中的实际情况和自身的特点，努力调动参与者的积极性，达到共建共享的目的。

1 地球系统科学数据的特点

1.1 交叉性和综合性

地球系统科学数据资源涉及地球科学及其相关的多个交叉学科和领域，同时，还涉及地球系统科学研究的各个圈层，尤其关注各圈层之间相互作用研究的数据。

1.2 分散性

地球系统科学数据的空间分布性、学科分散性以及地学研究数据来源的区域差异性、地学研究数据存储的物理分散性等，共同决定了地学研究数据具有分散性的特点。

1.3 异构性

地球系统科学数据的异构性主要表现在：数据源所依赖的应用系统、数据库管理系统乃至操作系统的异构；数据源在存储模式上的不同；语义异构表现在相同的数据形式表示不同的语义，或同一个语义由不同形式的数据表示。

1.4 不可重复性

地球系统科学数据反映地球系统各组成部分的变化过程和局格状况，由于现实世界不断变化，所以，无论采用什么样的技术手段所获取的反映地球系统各组成部分的数据，很难完全相同，它们在时间上或空间上都有差别。这种不可重复性的特征更增加了地球系统科学数据长期保存和共享的重要性。

2 地球系统科学数据共享网

2.1 概况

“地球系统科学数据共享网”作为国家“科学数据共享工程”首批的9个试点之一，于2002年启动。2004年，该网通过科技部和财政部的评审，纳入国家科技基础条件平台项目，属于科学数据共享工程规划中的“基础科学与前沿研究”领域。“地球系统科学数据共享网”是我国目前科学数据共享工程中唯一以整合，集成科研院所，高等院校和科学家个人通过科研活动所产生的分散科学数据资源，引进国际数据资源，接收国家重大科研项目产生的数据资源为重点的建设项目，在数据资源整合集成的基础上生产、加工数据产品，并健全标准规范和运行机制，构建分布式共享平台，为地球系统科学、全球变化等基础和前沿科学研究和国家重大战略决策提供数据支撑服务。

2.2 发展阶段

“地球系统科学数据共享网”的建设分为3个阶段：2003年到2005年是第一阶段，为自愿入网阶段；2006年到2008年是第二阶段，为规定入网阶段；2009年以后是第三阶段，为申请入网阶段。在第一阶段，由于社会公众对科学数据共享缺乏共识，此阶段采用自愿的原则，主要任务是提高大家的共享意识，制定标准规范，积极利用建设单位的数据资源优势、无偿提供服务，扩大影响，对自愿加入共享的单位和个人给予经费补

偿。经过两三年的工作，为后期的工作建立了基础。现在处于第二阶段，此时科技界已经对科学数据共享有了普遍的理解，申请加入数据共享网的单位和个人逐渐增多。在这一阶段，需要制定规范，进行有条件的遴选。从2005年到2007年期间，每年都有几十家单位申请加入共享网，包括科研院所和个人，每年筛选率约为50%，即每年约有40家单位申请，一般接纳18～20家加入。这一阶段主要通过制定国家政策和标准规范，建立功能齐全的网络服务平台，让用户十分方便地从网上获取所需要的数据，进一步提高数据服务质量，开展用户行为分析，以需求进行引导，并进行技术驱动，同时对筛选入网的数据源单位给予经费支撑。第三阶段从2009年开始，要求加入共享网的单位和个人根据国家政策、标准规范和技术以及共享网的服务目标和数据资源发展规划提出申请。申请的标准和条件相对于第二阶段将更加严格和规范。在三个阶段的运作中，同时实行"边建设、边服务、边完善"的原则。

2.3 成效

截至2007年12月底，地球系统科学数据共享网已在北京初步建成1个总中心，以及分布在国内相关单位的13个分中心和若干数据源点的分布式数据共享网络；并与国际相关地学数据网络相联，包括与美国十几家网站、蒙古、俄罗斯、国际山地中心等国家和地区的网站相联，是国内外地球科学领域中较大的网络之一；已整合集成了超过10TB的数据资源，形成了一批特色数据产品；初步构建地理科学、人文过程、资源科学、极地研究、固体地球、空间科学、对地观测、海岸近海等8个主体数据库；收集整理了4000多个国际数据资源站点，并对它们进行分类、导航和翻译，建立了4个国际数据资源镜像站点。

截至2007年底，共享网注册人数共计34597人，总访问量为2447846人次，提供了17TB的数据服务量。为500多个国家重大科研项目和重点工程项目提供数据支撑服务，为3469人次提供数据定制服务，产生了较大的社会和经济效益。

3 分散数据资源整合策略

在国家还没有统一政策，而部门的行政管理机制的职能在科研院所和科学家个人中还不便使用的情况下，分散的数据资源要实现整合和共享，需要采用多种灵活机动的策略才有可能，根据我们多年的探索和实践，以下的策略是可行的。

3.1 顶层设计

作为地球系统科学，数据源多而分散，科学家的需求迫切而强烈，所以，我们应该面向《国家中长期科技发展规划纲要》等战略需求，制定本项目发展的顶层规划与设计。结合本项目开展的多方面用户需求调查结果，构建本项目的总体数据资源建设体系，形成分散数据资源的整合集成、共享服务的网络体系。在顶层设计中，划分出自愿入网、规定入网和申请入网3个阶段，并对每个阶段的数据资源建设提出具体的内容和实现的措施，从而保证项目的有序实施。

3.2 明确责、权、利

由于项目属于科研项目，与科学家、行业部门、高等院校、科研院所等都有联系，要构建数据资源体系，必须明确数据拥有者、使用者、管理者的责、权、利，制定相关条例。对于数据拥有者，需要明确其数据资源实现共享以后，拥有什么权利和责任，能够获得什么利益。同时对数据的使用者和管理者也应该有相关条例进行规范。

3.3 先易后难

先易后难是一个重要的理念和做法。"地球系统科学数据共享网"由中国科学院地理科学与资源研究所组织建设。首先整合中科院地学领域的相关研究所、挂靠在中科院的"世界数据中心"(WDC)的5个学科中心以及有数据共享意识的高等院校和科学家个人手中长期积累的科学数据资源，实现共享。这部分数据资源的整合相对比较容易。通过实现相对容易的数据资源的整合，形成一定的规模。然后在此基础上，逐步推进

其他相关数据资源的整合。

3.4 先服务,后集成

由于分散数据资源分散性的特点，没有也不能使用任何行政命令进行资源的整合，所以实行先服务的方式。针对各类科研项目，利用已整合的数据资源，无条件地为相关科研人员提供上门服务。通过寻找服务对象,获取用户需求,主动提供无偿服务，让广大资源使用者逐渐了解共享网，使用共享网，并逐渐为共享网提供自己拥有的数据资源，最终实现网上共享。在数据共享服务在先、项目数据汇交集成在后、互赢互利、共建共享的原则指导下,不断扩大数据来源。

3.5 多渠道吸引国际数据资源

分布在科研部门、科学院、高等学校以及其他研究部门的科学家，正是产生数据资源的群体,他们能与很多科学研究的组织以及国际上有名的专家进行联系。利用科研单位和科学家与国外多渠道联系的优势,吸引国际数据,如:进行国际数据资源导航,直接引入数据源或者建立合作研究网络的方式。像俄罗斯、蒙古拥有大量的数据资源，尤其是俄罗斯的科学家拥有很多、很有价值的而国内科学家又特别需要的数据资源,我们通过建立“东北亚生态环境合作研究网络”,把我国可以提供给国外科学家使用的数据放到网上实现共享，按照协议他们也这样做，实现三国科学数据的共享，促进学术交流，通过这种方式建立的共享,能够吸引大量的国际数据资源。我们与“国际山地中心”建立的合作网吸引了喜马拉雅山周边国家的数据源;我们通过科学家之间的关系引进了大量免费共享的遥感信息源,供国内科学家使用，为他们节省了大量的费用；利用世界数据组织和国际科学计划使我们与国际上10多个相关的数据中心产生数据交流或镜像。吸引国际数据资源,在交流的过程中建立良好的关系。这些数据资源可以供国内的科研人员使用。

3.6 建立精品数据集、数据库

很多科学家需要原始数据，但还有相当一部分科学家在科学研究中需要综合的数据集。为了满足这部分科学家的需求，为他们提供服务，需要建立精品数据库或数据集。当前，基于地球科学已有分支学科的数据，生产了一系列数据产品，受到很多科研部门和科研人员的欢迎，发挥了重要的作用。但是科学家以各种方式只能获取一些专业数据,而难以获取综合、全面、长期和空间分布的数据集。“地球系统科学数据共享网”通过建立精品数据库或数据集的方式，为科学家服务，支持他们的科研项目，并吸引科研项目数据入网,通过网络提供服务。实现数据共享中的“共建共享”理念。

3.7 边服务,边建设

这个理念是做数据共享所有单位的共识。“地球系统科学数据共享网”作为一个专业服务型网络,要提倡在建设中服务,在服务中促进建设,只有让广大科技人员在共享中得到收益,才能提高他们对科学数据共享的认识，进而吸引他们加入到共建共享的行列中来。多年来,我们始终把做好数据应用服务放在首位，无论是数据资源整合,还是数据产品的加工生产,乃至网络平台的构建等，均把用户的需求放在第一位，即以需求为主导,建立共享服务系统。只有提供优质的服务，才能吸引和整合集成更多的分散数据资源。

3.8 建立结构合理的专职建设队伍

队伍是一切事业成功的根本。没有一支坚持做科学数据共享的建设队伍，科学数据难以实现持久共享服务，队伍的组织在科研部门有较大的难度，因为数据共享服务为科学研究提供服务，他们无法发表高水平的文章。对这部分人员的考核和评价应有专门的指标体系，以稳定队伍，保证科技基础条件平项目建设和运行服务的长期性。我们在队伍组织中,首先要求中心、分中心和数据源点有专职人员，相关单位应对专职人员有一定的稳定措施。建设队伍的成员应由三部分人员组成，一部分是固定从事数据的人员，一部分是从事IT技术的人员,这两部分组成专职的建设和服务人员，第三部分是兼职的科研一线人员，

他们是科学数据的使用者，又是科学数据的生产者。我们聘请了一批这样的专家直接参与共享系统建设，这种形式即专职人员与兼职人员相结合。

3.9 聘用离退休人员挖掘数据资源

广大的离退休科研人员手中掌握了一大批宝贵的、分散的科学数据资源，他们对自己几十年科研活动中所保存下来的科技数据和资料都十分珍惜，特别愿意能够亲手再整理出来供后人利用。我们经过调研，采取自愿的原则，给予一定经费补偿，请他们整理手中的数据资料，参与到我们的共享体系建设中。

3.10 构建共享联盟，形成联盟体系

逐步构建地球系统科学数据共享联盟，调动国内外各方面的力量形成科学数据共享联盟体系，促进数据流通、共享增值。这也是地球系统科学数据共享网正在进行的一项工作。将来要更加完善数据共享工作，尤其是让更多的科学家拥有的数据资源放在共享网上共享。共享联盟的逐步形成，核心问题是共享网能否满足参加联盟的成员对科学数据的需要，能否处理好数据拥有者、数据使用者、数据管理与服务者之间的责、权、利的关系，如果科学数据共享联盟体系的运行机制得到社会认可，则能够很好地代替行政部门的行政命令，使科学数据资源整合集成产生重要影响。

4 分散数据资源整合模式

4.1 支付一定的整合集成经费

由国家在科技基础条件平台建设计划中拨付建设经费，是保障科学数据资源整合集成共享服务的先决条件，目前整合的是过去已产生而不符合一定规范标准的数据资源，必须经过系统整理和加工产生符合要求的数据集。这个过程需要一定的经费。对于未来所产生的大量更新数据，同样需要投入资金以保持科学数据共享事业永久持续。多年来，我们所提供的科学数据资源，除了从国外引进的数据资源外，均是以投入一定的经费而实现的，这种情况在数据共享领域各单位基本是一致的。

4.2 先提供数据服务，后整合全部结题项目数据系统

这种模式是先向在研项目提供数据服务，然后将该项目所建立的数据管理服务系统，整合集成到我们所建立的“地球系统科学数据共享网”，成为该网的一部分。例如，“中国草业开发与生态专题数据库”开始是一个项目，由共享网提供部分数据支撑，项目结题后，将其开发的“草业开发与生态专题数据库”全部纳入到“地球系统科学数据共享网”中。现在，草业开发与生态专题数据库光盘已经免费寄送到全国3000多个县级农牧局和科技局。

4.3 边提供数据服务，边集成项目研究过程数据

科学研究整个过程都可能产生若干研究过程数据。数据研究者把其作为进一步研究的过程数据，而对其他研究人员所从事的其他研究项目同样有用。如果原研究的群体或科学家个人愿意在最终成果数据产出前就提供给社会共享，我们就从项目一开始为他们提供原始的和加工的数据产品，尽可能地满足他们的需求，使他们早日产生过程数据，供更多的研究人员使用，节约了使用这类过程数据的科学家的劳动和时间。这种数据使用过程，是我们共享网数据流动增值的过程之一。

4.4 委托建库

进入21世纪信息时代，数据和信息受到人们普遍重视，但是数据库需要人、财、物的支持，这样对很多科研项目来说，需要有较多开销，特别是要组织建库的队伍。数据库建成后还需要有人运行维护和提供服务。针对这种情况，利用建设队伍中数据资源和技术力量的优势，为研究项目承担建库任务。在这种模式中，我们通常根据委托部门或项目的要求，设计数据库整套软件，并按委托项目的要求，将数据共享网中与委托建

库内容有关的数据资源整理好，与委托方的数据资源一起进入所建的数据库，成为共享网的一部分，但在数据服务中，必须遵守委托方的有关规定。

4.5 直接承担科研项目(重大科学工程)数据管理系统建设

“地球系统科学数据共享网”在全国有20多个研究所、高校设有分中心或数据源点。国家与地球系统科学、资源环境等领域的重大科研项目也大多集中在科研院所和高等院校，所以我们除了总中心以外，还充分利用分中心或数据源点的优势，尽量从一开始就介入本单位所承担的科研项目，利用共享网的数据规范标准及共享平台的技术标准直接为项目建设项目数据管理系统，使其成为“地球系统科学数据共享网”的分中心或数据源点的组成部分。

4.6 镜像国外数据资源

国外数据资源是国内科学家十分喜用的数据资源之一，尤其是地球科学领域更对全球的数据资源感兴趣，国外数据资源的取得在网络十分通畅的今天还是比较容易的，但要通过国际网去下载大量的数据资源所发生的费用却十分可观，所以镜像国外的数据资源仍然是十分可取的模式。

4.7 引进国际免费数据源

地学领域的研究，普遍要用到遥感信息源，例如TM信息源，根据中美协议，美方不能直接从美国向中国用户提供中国范围的TM信息源。我们与美方高等学校签订这方面资源使用和免费发放的协议，通过“地球系统科学数据共享网”免费向国内用户提供使用。几年来，按国内现行价格计算，仅这一项服务可以为国内的科学家节省8000万元的经费，产生了很大的经济效用。

4.8 共享软件成果

“地球系统科学数据共享网”建设队伍中，有一支较强的数据管理和应用服务的软件开发人员，他们对地球系统的科学数据有一定程度的了解，有针对这一领域开发管理软件的能力。相关单位利用我们开发的软件系统，可以顺利地进入国家科学数据共享的行列。

4.9 建立国外数据网站导航系统

我们充分利用互联网查找国际上与地学领域有关的网站，建立网站导航数据库，按网站所含数据的类型进行分类，对每一类每一个网站都用中文进行介绍，使国内科学家能非常方便地查到他所想查的网站，再进一步查找所要的数据。目前“地球系统科学数据共享网”中已有近4000个国外网站。

4.10 建立英文“地球系统科学数据共享网”网站

为了加强国际交流，吸引国际数据资源，在保证国家安全的前提下，我们构建了英文版的“地球系统科学数据共享网”网站，并直接与国际相关网站相联，交换元数据。

在讨论分散科学数据资源整合集成和共享过程中，有一点需要指出，我们所说的将分散数据资源整合集成，不是要求数据拥有者一定汇交到某一个固定的地方，而是将那些有用的、分散的数据资源，按一定的规范标准整理好，提供到一个方便用户、实现共享的环境中，数据拥有者有权选择数据的存放环境，只要达到共享的目的即可。对于那些有必要进行相对集中的数据资源，相关部门会制定具体的规定，例如，国家973计划资源环境领域项目数据，现在科技部已经规定要汇交到科技部的973计划资源环境领域项目数据汇交中心。