

科学数据共享及其在互联网上实现的技术途径

Scientific Data Sharing and Its Solutions Based on Internet

廖顺宝

(中国科学院地理科学与资源研究所 北京 100101)

一、科学数据及其共享的重要性

数据是科学研究的基础,越是重大的科学研究,越是需要大量基础数据的支持。如果把科学研究的成果比做地面上的高楼,数据就是地面下的地基,楼越高,需要的地基越坚实。离开科学数据支持的科学研究就像在沙滩上建楼,永远不可能建起高楼大厦。

国内外许多重大的科学研究项目和计划都十分重视科学数据的采集、管理和共享工作。

美国长期生态研究计划(LTER,目前已发展成为长期生态研究网络—LTERN)是美国国家科学基金委员会(NFS)提供资助、从事长期生态研究的实体,其科学研究所依赖的主要数据源是由代表美国主要自然生态系统类型的17个生态站采集和提供的。

为使来自各种渠道的大量监测数据得以合理利用,联合国环境规划署(UNEP)于1985年制定了全球资源信息数据库计划(GRID),其目标是建立一个世界上任何国家都容易使用的世界范围的环境数据网络,使现有的环境数据成为对规划者和科学家有用的财富。

全球变化是当今学术界研究的热点,国际地圈—生物圈计划(IGBP)是国际上从事全球变化研究最具代表性的组织之一。为了支持其研究目标的实现,它把“为全球变化研究收集在时间上和空间上描述整个地球系统的高质量的数据,并以尽可能有效的方式为从事全球变化研究的组织和个人提供各种数据管理服务”作为其信息系统的目标。

在我国,中国生态系统网络(CERN)是由代表我国不同生态系统类型的29个生态站、5个分中心和1个综合中心组成的,以研究生态系统的结构、功能为主,并服务于环境、资源和大农业研究的综合研究网络体系。CERN的科学研究主要依赖于各生态站采集、收集和整理的长期定位观测数据、实验数据和社会经济数据。

为促使我国与国外进行数据方面的交流与合作,我国于1988年加入了世界数据中心(WDC)。目前,遥感已经成为地球科学研究获取数据的主要技术手段,我国已经建立了不同类型的遥感卫星地面接收站,如NOAA-AVHRR地面接收站、MODIS数据地面接收站。由此可见,数据不仅是科学研究的基础,同时也正在成为科学研究的驱动力。

科学数据对于科学研究十分重要,而科学数据的共享同样十分重要。科学数据的获取往往需要投入大量的人力、物力和财力,特别是像资源环境等具有地学特点的数据。众所周知,资源环境系统的变化是一个十分缓慢的过程,要科学地揭示和描述其变化的规律,必须要有长时间序列的数据支持,而且往往是时间序列越长,数据越有价值,因为许多变化过程是不可逆的。因此,这类数据的交换与共享尤其必要。同时,许多数据的产生者或拥有者在使用过几次数据后,一般就不再使用这些数据而搁置在一边,而他们的同行却渴望得到和使用这些数据,如果能够以某种方式共享这些数据,不仅能够最大程度地发挥这些数据的价值,而且还能避免重复投入,为国家节省开支。目前绝大多数的科学数据的投入均属国家投资,从这个意义上说,科学数据也必须实现共享。

二、科学数据共享的有关问题

与科学数据共享有关的问题很多,本文将其归纳为观念问题、政策问题、费用问题和技术问题。

1. 观念问题 要实现数据共享,必须解决观念问题,否则,共享无从谈起。首先,数据生产者或数据占有者应当认识到,他们在从事数据采集、收集和整理的工作中,或多或少地得到了国家投资的支持,因此不应该把数据当作自己的私有财产。其次,在数据交换与共享的过程中,数据提供者从事数据提供服务的同时,也可能在接受别人提供的数据服务。再者,数据产生者并不总是一直使用自己所产生的数据,往往是使用了一段时间以后就不再使

用了,那么,为什么不能把这些数据通过一定的方式与别人共享呢!而且通过数据共享,他们还有可能从中得到一些经费上的补偿。

2. 政策问题 这里所说的政策问题,包括两个方面的内容。首先,对于国家长期固定投资使之从事某些专项领域(如气象、水文、地震)的数据采集与数据生产的组织,国家应当制定明确的政策,使这些机构或组织生产的数据成为数据共享的主要数据源,至少是公益性数据共享的数据源。否则,国家投资就只能服务于这些机构或组织,不仅国家利益得不到完全体现,而且还会造成高价倒卖数据的流弊,因为其它的个人或组织要获得这些数据比较困难,甚至是不可能的。其次,要有一套鼓励数据共享的政策和措施,保护数据提供者的利益。数据共享并非数据“共产主义”,数据共享必须尊重和保护数据提供者的知识产权,如:被提供服务者为最终用户,对数据不具备散发权;数据提供者在被提供服务者的研究成果中有署名权,等等。此外,不论数据本身是免费还是有偿的,应当允许数据服务的提供者收取适当的工本费和服务费,以维持数据共享系统的正常运转。

3. 费用问题 费用问题是数据共享中比较敏感的问题,费用包括数据本身的费用和数据以外的费用。对于完全是国家投资所产生的数据(绝大部分为公益性数据),原则上不应当再收取数据费,因为生产数据的费用开支已经由国家投资所支付。而对于不是(或不完全是)由国家投资所产生的数据,被提供服务者可能需要对数据支付一定的费用,因为数据提供者在获取这些数据时进行了大量的人力、物力和财力投入,他们当然希望通过数据共享与数据服务来获得一定的经济上的弥补,这也在一定程度上为生产新的数据提供了经费支持。

不论数据本身是免费还是有偿的,数据服务的接收者都应当支付必要的工本费和服务费,工本费主要是指存储数据的介质费用和数据通过网络传输的费用,如:磁盘费、光盘费、网络传输费等;服务费主要指提供数据服务人员的劳务费。只有这样,才能鼓励和支持数据服务与数据共享工作的开展,维持数据共享系统的正常运转。

4. 技术问题 当数据共享的观念、政策和费用等问题解决以后,剩下的问题就是如何实现数据共享,即数据共享的技术问题。

随着信息技术和计算机网络的不断发展,数据的存储已由传统的模拟介质(纸、胶片)发展到以数字介质(磁盘、磁带、光盘)为主,而信息的传输也由人工传递、邮寄发展到以网络传输为主。据中国互联网络中心的统计,截止到2001年6月30日,我国上网计算机约1 002万台,上网用户约2 650万人,WWW站点数超过24万个。网络的发展为数据共享

提供了更加先进的技术手段。目前我国,互联网主要是用作公司、企业、组织的宣传、介绍以及公共免费信息的查询和检索的平台,尚很少利用互联网进行数据的共享与服务,以下将对这一问题进行研究和探讨。

三、基于互联网的数据共享的技术途径

基于互联网的科数据共享系统由数据提供者、计算机互联网络和数据用户3部分构成,数据提供者通过网络把数据传输给用户,同时,数据提供者和数据用户之间还通过网络进行与数据共享有关的其它信息的传送,如:用户注册信息、费用信息、用户反馈意见等。(见图1)

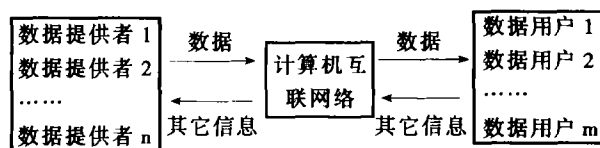


图1 基于互联网的数据共享系统的构成

在功能上,系统应具备元数据管理、用户管理、数据管理、费用管理以及对用户使用数据情况的监督与管理功能。

1. 元数据管理 元数据是对数据进行描述和说明的数据,即关于数据的数据。通过元数据查询检索系统,用户可以了解到某一数据提供者目前有哪些数据可以共享,可以了解某一特定数据集的产生背景、数据质量、数据格式、数据量大小等十分有用的信息。在了解这些信息之后,他们才能决定是否要使用其中的某些数据,进而向数据提供者提出使用数据的申请。除查询检索功能外,元数据管理系统还应具备元数据输入、编辑修改功能,因为数据提供者随时可能有新的数据集上网共享,需要不断地向元数据库中追加新的记录,同时,已有的元数据信息随着时间的推移也可能做某些修改。

2. 数据的分级管理 数据共享是在不同的层次和不同的级别上进行的。根据数据的对外保密程度不同,不同的数据具有不同的秘级,一般分为:绝密、机密、秘密、国内公开和公开等5级。对于大部分的数据用户来说,前3级数据(占总量的很少部分)是不能提供的;对于国外用户,前4级数据都是不能提供的;而对于一些特殊的用户,如党政机关、军事机关,所有的数据都应当提供。因此,在数据共享系统中必须记录数据集的保密(公开)级别。

3. 用户的分类管理 数据具有不同的保密和共享级别,而这种不同级别的划分主要是依据数据的使用对象(即数据用户),因此,在数据共享系统

(下转第56页)

然不是所有 200 万人以上的特大城市,都不让其合理发展,如上海浦东、大连、南京、重庆等城市还可以有控制地发展;而像北京、天津、广州、兰州、沈阳、西安、成都这些历史文化名城,以及其它环境恶化,淡水资源、地下水资源稀缺,市内交通十分堵塞的城市还是应当严格控制其人口、用地的规模。那些市区人口 200 万人以下的省会城市,用水、用地潜力较大的重要港口、交通枢纽城市等,应允许其在完善基础设施和提高环境质量的基础上适当增加城市人口,让其充分发挥中心城市的作用。市区非农业人口 50 万左右的中等城市,人口规模扩张速度可以快一些,以便进一步发挥区域经济中心的辐射作用。此外,还应注意发挥中小城市联系城乡经济文化的纽带作用,同时应发挥中小城市和建制镇吸纳农村剩余劳动力的主导作用。

在严格控制大城市规模、合理协调发展大中小城市的同时,应当积极推进小城镇的发展。但在我国农村人口基数大的情况下,小城镇面广量大,加上经济实力有限,所以小城镇的发展应合理引导、分期分批、有重点地稳步发展。在小城镇建设与发展的起步阶段就应注意发展生态农业与生态工业,

走集约化、专业化的道路,以免重蹈传统农村工业“家家点火、村村冒烟”的覆辙。小城镇应当成为我国农村人口城市化的一个重要生长点和周边农村经济的中心。

参考文献

- [1] 崔功豪. 近 10 年来中国城市化研究的进展. 地域研究与开发, 1989(8)
- [2] 赵燕青. 战略与选择: 中国城市化道路回顾. 城市规划, 1990(14)
- [3] 郑弘毅主编. 农村城市化研究. 南京大学出版社, 1998
- [4] 中国城市发展报告. 中国科学院南京地理与湖泊研究所, 1999
- [5] 赵士修. 关于我国城市化进程若干问题的思考. 城市规划, 1996(20)
- [6] 俞正声. 要从战略的高度认识城市化问题. 城市规划, 1998(3)
- [7] 姚士谋, 陈振光等. 中国城市群新论. 中国科技大学出版社, 2001
- [8] 姚士谋, 朱振国. 城市规模切忌盲目扩大. 中国土地科学, 2001(4)

(责任编辑 肖庆山)

(上接第 26 页)

中,除对用户的常规信息进行管理外,还必须对数据用户进行不同类别的划分,这种划分应当与数据保密级别的划分是一致的。科学、合理地对用户进行分类,既可以最大程度、最大范围地共享科学数据,使其最大限度地发挥其价值,同时又不至于危害国家安全和国民经济。

4. 记费管理 数据共享并不等于数据的无偿使用,数据使用者必须向数据提供者支付用于产生数据和进行数据服务(工本费、劳务费、数据传输费)的必要费用,这样才能鼓励数据提供者向外提供数据、维持数据共享系统的正常运转。因此基于互联网的数据共享管理系统应当具有记费管理功能,包括:数据本身的费用(如每兆字节应支付多少钱)、劳务费、工本费、邮寄费、网络传输费,等等。一方面,数据提供者要使用这些信息来计算数据用户应支付的费用,因此,这些信息应当公布在网上供用户查询,以增加数据服务的透明度;另一方面,数据用户在被提供数据服务后实际所发生的费用也应记入费用数据库。

5. 系统对用户使用情况数据的监督与管理 作为一个完整的网络数据共享管理系统,既要有数据共享管理的基本功能,同时又要对用户使用数据情况的监督与管理功能,如哪些用户使用过本数据中心的数据、使用的数据集的名称、使用的时间、本数据中心哪些数据集被使用的频率较高以及用

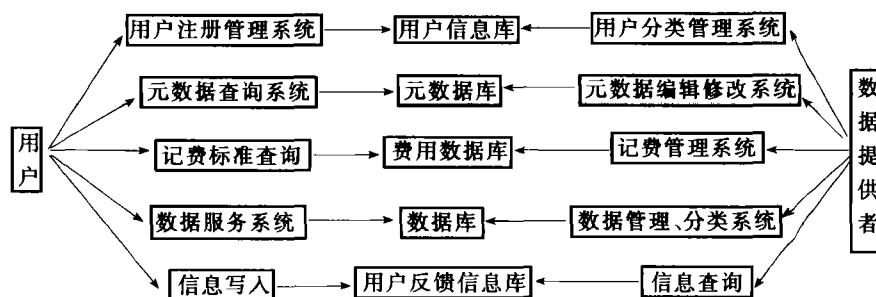


图 2 基于互联网的科学数据共享系统的功能结构

户在使用数据过程中产生的反馈信息,等等,以便进一步改进和提高数据服务的质量。

6. 网络数据共享管理系统的功能结构 根据上述分析,我们可以构建出一个完整的基于互联网的科学的科学数据共享系统的功能结构图(见图 2)。

在互联网日益成为信息发布和信息传输主要手段的今天,以互联网为载体构建科学数据共享新的信息平台必将进一步推进数据共享工作向前发展。

参考文献

- [1] 世界数据中心中国中心协调办公室. 世界数据中心中国中心与地球系统科学数据. 北京:科学出版社, 1995
- [2] 中国 21 世纪议程管理中心. 中国地理信息元数据标准研究. 北京:科学出版社, 1999
- [3] <http://www.cnnic.net.cn/develst/rep200107-1.shtml>

(责任编辑 孙立明)