

地球科学数据共享平台建设思路与实践

苗茹¹ 宋佳² 王卷乐² 诸云强²

(1. 河南大学计算机与信息工程学院, 河南开封 475004; 2. 中国科学院地理科学与资源研究所资源与环境信息系统国家重点实验室, 北京 100101)

摘要: 提出地球科学数据共享的框架体系, 并从国家政策机制、共享服务规范、共享服务对象以及系统平台应用的角度对地球科学数据共享的建设和发展提出建议。通过建立健全地球科学数据共享过程中的规范条例、明确共享服务对象的权利义务、合理定位共享平台的功能体系, 促进地学领域数据共享事业的健康发展。

关键词: 地球科学; 共享标准规范; 共享要素; 数据汇交; 数据共享系统; 地球科学数据共享平台

中图分类号: P208, TP311

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2014.04.011

Study on Construction and Development of Geo-Data Sharing Platform

Miao Ru¹, Song Jia², Wang Juanle², Zhu Yunqiang²

(1. College of Computer and Information Engineering of Henan University, Kaifeng 475004; 2. State Key Lab of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Beijing 100101)

Abstract: This paper proposes the framework of the geo-data sharing system, which includes national policy mechanism, sharing service specification and standards, sharing service users, system platform application. In the process of data sharing, we should provide the sharing standards and specification to ensure the integrity of data resources. On the other hand, we should make it clear the rights and obligations of users and positioning the function system of sharing platform reasonable.

Keywords: earth sciences, sharing standards and specification, sharing features, data integration, data sharing system, geo-data sharing platform

现代地球科学(简称“地学”)主要研究大气圈、水圈、岩石圈、生物圈和日地空间的过程与变化以及这些过程之间的相互作用^[1-2]。地学研究需要跨学科、跨区域、多过程模拟和融合数据

的支持^[3], 因此, 地球科学数据(简称“地学数据”)是科研人员开展科学研究的基础。

为了促进地学数据资源的共享, 国内外开展了一系列基于信息网络的科学数据共享技术研究

作者简介: 苗茹(1979-), 女, 博士, 研究方向: 空间数据组织与可视化, 地学数据信息化; 宋佳*(1980-), 男, 博士, 助理研究员, 研究方向: 空间数据组织与可视化, 地学数据共享和信息化; 王卷乐(1976-), 男, 博士, 副研究员, 研究方向: 资源环境科学数据集成与共享; 诸云强(1977-), 男, 博士, 副研究员, 研究方向: 地学数据共享关键技术, 资源环境信息系统。

基金项目: 国家自然科学基金青年基金“面向地图可视的矢量数据流式渐进传输研究”(41201411); 国家科技基础性工作专项“中国北方及其毗邻地区综合科学考察”(2007FY110300-8); 资源与环境信息系统国家重点实验室自主部署项目“陆地表层系统数据—模型共享研究与应用示范”(O88RA900KA)。

收稿日期: 2014年7月12日。

和实践，建立了一批地学数据共享平台。如，为了促进全球地球观测数据的共享，于2005年成立了地球观测组织^[4]（Group on Earth Observations, GEO），国际山地中心2006年主持创建了兴都库什-喜马拉雅地区空间信息共享网络^[5]，美国在2003年建立了地理空间一站式共享网络^[6]（Geodata.gov）。我国在1997年成立了国家地理空间信息协调委员会，开始推动国家空间信息基础设施的建设^[7]，特别是2003年启动了“地球系统科学数据共享网”项目，构建了地球科学数据共享平台，并于2005年纳入国家科技基础条件平台，其中6大类43个科技资源共享平台得到了国家财政支撑^[8]，对地学科学数据的共享发挥了重要作用。

尽管科学数据共享已经在发展规划、共享标准制订、科学数据整合以及共享服务等方面取得了长足进展，但是仍缺少有效的数据质量管理体系。数据资源共享中规范化和标准化程度低，同时也缺少对数据资源有效的评估和评价体系等。

因此，基于数据用户、数据资源是地学数据共享核心主体，本文将从数据使用者、数据贡献者和数据服务者3个层次探讨地学数据共享过程中的机制、需求。最后，通过国家地球系统科学数据共享平台，构建科学的地球系统科学数据共享框架，分析共享软件平台的功能体系和支撑技术，促进科学数据共享事业更健康、科学、全面的发展。

1 地学数据共享框架

地学数据共享框架包括共享政策、数据资源规范、用户服务对象以及系统实现模型等，如图1所示。

(1) 共享政策

开放的共享政策和严格的版权保护是科学数据共享所要遵循的原则。科学数据要采用对数据贡献者激励的机制，例如：通过“积分奖励”的形式，用户贡献的数据越多则共享到别人的数据也越多，只有这样才能让更多的数据贡献者或单

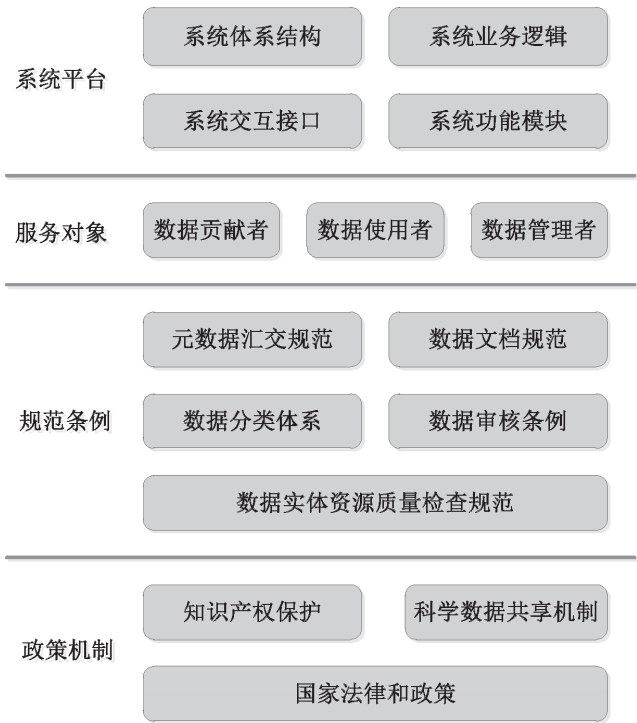


图1 地学数据共享框架

位愿意将他们的数据给广大用户分享。此外,利用数据标识、引用机制保护数据生产者或加工者的知识产权,提供数据使用声明与引用规范、数据申请审核等机制,保障数据贡献者的利益。

(2) 规范条例

地学数据的共享规范条例包括元数据描述规范、数据文档书写规范、数据分类体系设置、数据汇交时对数据实体进行质量检查的规范以及数据发布审核条例的规范。元数据描述规范是指数据汇交过程中元数据填写的详细说明文档。数据文档是对数据的详细描述,应该指导对数据详细描述过程中涉及到的要素、说明形式、描述方法等。数据分类体系设置应面向数据共享的应用环境,选择适合用户搜索和满足数据归档的科学分类体系。数据实体质量检查是数据管理者对需要共享的数据资源以及和数据相关的材料进行检查、测试、试验的过程,并将检查结果根据质量检查规范要求进行比较以确定数据质量是否合格。检查过程中首先要保障数据的完整,材料齐全,然后根据数据内容进行必要的测试或试验,保证提交的数据真实可用。数据审核是在数据公开发布之前由数据管理人员或评审专家对元数据、数据文档、数据实体等相关要素进行评审的过程。审核条例可包括审核机构、审核标准、审核程序、审核结果的处理方式等。

(3) 服务对象

数据共享面向用户需求可以分为3类不同的用户:数据贡献者、数据使用者和数据管理者。

一方面,数据贡献者提供数据源授权并保障数据质量、提供数据更新;另一方面,还需要确保数据贡献者对数据使用的被引用权、知情权和决定权。被引用权指的是数据贡献者可以要求数据使用者对数据来源进行声明,引用数据直接相关的论文。知情权指的是数据贡献者应该知道所提供的数据被哪些用户使用了,用做什么途径。决定权指的是对于申请的数据,可以根据数据应用目的,由数据提供者决定数据提供哪种使用方式,如直接在线下载或者通过申请审核之后获取。另外,需要给予数据贡献者一定的奖励,以

此激励更多的科研工作者和单位共享数据。只有切实保护好数据贡献者的权益,给予数据贡献者应有的地位,才能真正激励和推动科学家和科研单位主动共享自己的科学数据。

数据使用者主要依据自己的需求查找自己需要的数据,然后进行下载或申请。为了保障数据贡献者的权益,要求数据使用者在使用数据时需要提供全面的产权引用信息,包括同意数据使用声明、引用相关的参考文献、说明项目资助信息以及直接引用数据的DOI(数字对象唯一标识符, Digital Object Identifier)。另外,数据使用者也可以对使用过的数据评价,通过建立数据评价体系,筛选出一批高质量、高使用率的数据,作为优质数据资源向用户推广。为了激励用户使用数据并进行评价,可以建立“积分机制”,根据对数据评价内容的详细程度和专业性给予一定的积分奖励,积分越多其获取的数据使用权限就越大。

数据管理者是联系数据贡献者和数据使用者之间的“纽带”。一方面需要将数据贡献者提供的数据进行整编集成,在此过程中要整合数据信息,控制数据的质量,并将审核通过的数据公开发布以供数据使用者下载或申请使用。另一方面需要协调保障数据贡献者和数据用户的权力与义务,保护数据提供者的知识产权,使他们提供真实可靠的数据资源给用户使用。另外,数据管理者还需要提供数据共享软件平台及数据申请服务、收集跟踪数据使用情况、对数据资源进行日常的维护管理等。

数据贡献者、数据使用者和数据管理者应用数据的过程如图2所示。

(4) 系统平台

基于数据用户、提供者、管理者是参与数据共享核心主体,从数据使用者、数据贡献者和数据服务者3个层次分析和定位共享平台的功能体系和支撑技术更具有应用价值和共享效果。在数据使用者参与数据共享的“寻找数据”“获取数据”“反馈质量”3个核心环节中,“多元导航式筛选与搜索”“一站式数据订单”“大众数据评

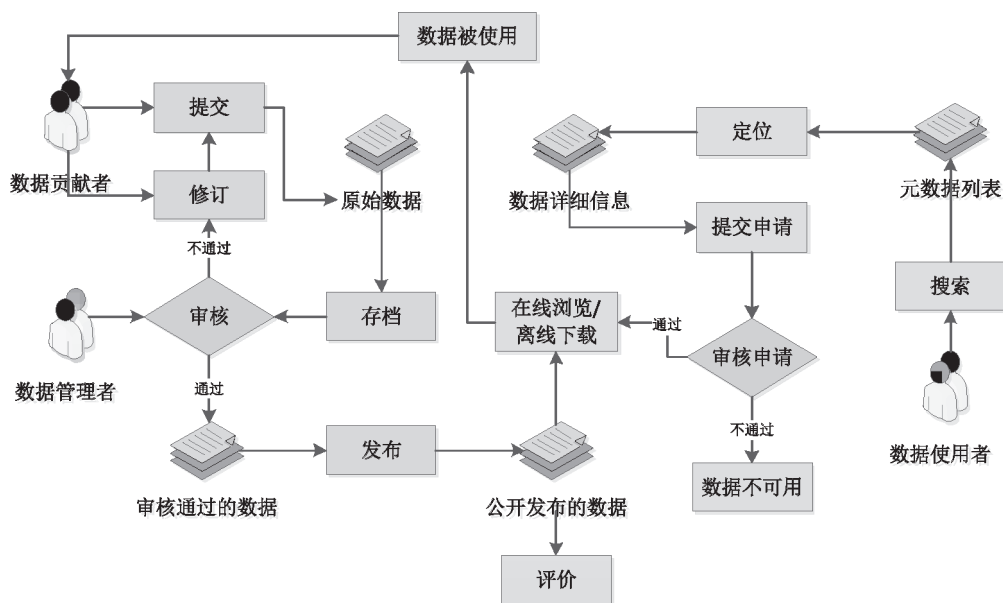


图2 3种服务对象应用数据的过程

价”等是支撑这些环节的重要功能体系；面向数据贡献者所关注的“数据产权”“数据使用许可权”“数据使用知情权”3个核心问题中，提供可靠的“数据使用声明与引用规范”“数据申请许可”“数据使用反馈”功能是共享平台保障数据贡献者利益的关键；数据服务者作为数据使用者和数据贡献者的纽带，在参与“数据整合集成”“数据质量控制”“数据申请响应”3个重要环节中，需要共享平台在“多标准元数据发布”“数据专家评审”“数据申请审核”等方面提供功能保障。

2 数据共享要素

数据资源是地学数据共享环节中的重要组成部分，通过详细全面的数据资源汇交模式，可以提高元数据检索的精度，帮助用户快速检索和正确使用共享数据。为了对数据资源进行更加详细的说明，数据汇交可采用“六位一体”的数据描述方式，包含元数据、数据缩略图、数据说明文档、数据样例、数据实体、数据关联分类与筛选组等，如图3所示。

2.1 核心元数据描述规范

元数据是数据描述的核心信息，是用户对数据了解的关键。通过建立元数据填写规范，全面

提升元数据的质量，提高元数据检索的精度。元数据包括数据核心描述信息、数据贡献者信息以及数据服务单位及联系人信息。

(1) 数据核心描述信息

数据核心描述信息可包括数据标题、主题要素词、空间范围及尺度、时间范围及分辨率、数据产生来源、数据类型、学科类别、数据访问方式和DOI号。标题中需包含内容、时间、地点、尺度等信息。主题词是能够体现元数据内容的简短词语，应该有两个以上，并且尽量集中，便于用户搜索。时间范围可按“年代”“年”“月”“日”进行划分，空间范围可用地理位置进行描述或以经纬度坐标作为搜索条件，且经纬度坐标可在交互地图中通过选定一个空间来自动确定坐标值。数据详细描述信息可以看成是数据文档的摘要，主要对数据资源进行一个简要的介绍，主要包含数据内容、投影方式、数据量、数据类型等。数据访问方式可以根据数据等级或数据贡献者的要求分为“在线下载”或“离线申请”等，为了简化数据获取流程，尽量提供便捷的在线数据获取方式。对于离线数据要提供在线填写离线申请以及详细的联系方式，保证用户能够顺利获取数据。

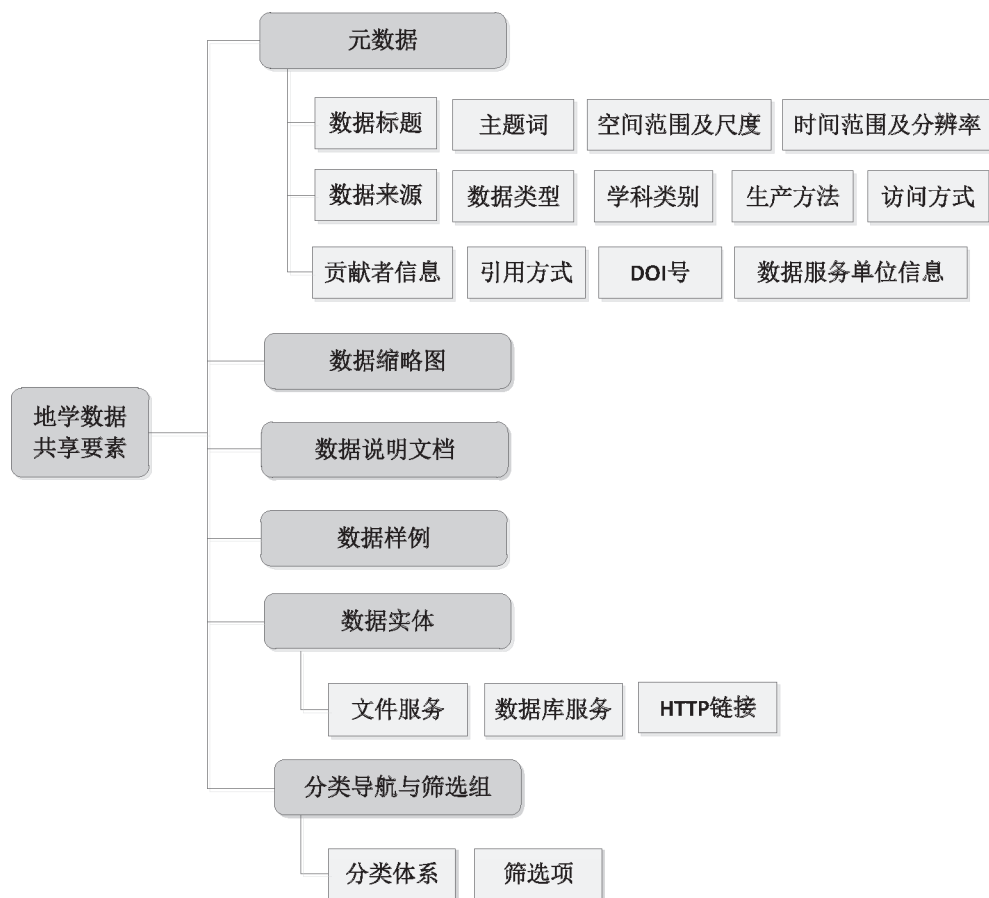


图3 “六位一体”的数据汇交模式

(2) 数据贡献者信息

数据贡献者信息主要为了突出数据提供者，通过提供数据贡献者的姓名、邮箱、单位以及数据引用方式保障数据的知识产权。数据引用方式建议包括以下3种。一是引用平台格式：“数据贡献者名称. 数据集名称. 共享平台”；二是遵照数据贡献者意愿引用相关论文；三是数据贡献者对使用本数据的其他声明数据。

(3) 数据服务单位信息

数据服务单位及联系人作为数据管理方，主要负责对数据的存档和维护。数据管理方一般需要提供数据联系人、联系方式（邮箱、电话）、负责单位名称、地址等信息，方便用户联系申请数据。

2.2 数据资源汇交规范

数据资源包括能够反映数据特征的缩略图、数据的详细说明文档、数据样例文件以及数据资

源实体。

(1) 数据缩略图

数据缩略图是反映数据内容的代表性图片，空间数据可以通过数据可视化展示，属性数据可以用代表性要素景观图片代替，如台站可用台站景观照片，森林数据可选择森林照片。为了能够让数据使用者更清楚地了解数据集，还可以提供数据在线预览。对于属性数据，用户可以预览前10~20条信息；对于矢量数据，用户可以筛选大斑块几何要素预览；对于栅格数据，用户可以预览其缩略图。

(2) 数据文档

数据文档是对数据的详细说明，数据文档的质量直接关系到用户能否正确使用数据资源。数据文档中应该对数据集的内容字段及其数值单位、数值代码类型以及数据来源、生产方法、质量控制措施、数据使用环境、使用限制、数据的

产权信息和引用方式等进行详细的说明,辅助用户正确使用数据。

(3) 数据样例

数据样例是从数据实体中截取极少量(或非关键)内容作为数据的代表,可被用户自由下载。数据样例可以是数据集的某一个空间子区域或是某一个时间段的子集,它是原始数据的真实组成,并且能够包含完整的数据描述信息,具有一定的代表性。通过数据样例文件,能够弥补对数据说明中的不足之处,这样用户就更加明确该数据是不是自己真正需要的,避免出现经过申请之后才发现并不是自己需要的数据的问题。

(4) 数据实体与获取方式

数据实体资源应该严格按照质量检查规范进行提交,保证用户得到的数据真实可靠。数据获取方式可以设置为“在线下载”或“离线申请”。在线下载的数据通过提交在线申请之后,便可获取数据下载的连接地址,一般在有效时间期限内下载即可。离线申请的数据需要在线填写数据申请表,并提交管理员审核,审核通过后获取数据下载方式。数据的下载方式可以提供文件服务、数据库服务以及原始数据的HTTP网址。如果提供给用户的下载地址为临时地址,则需要设定一定的失效时限,以保障数

据的安全性。

2.3 数据分类与筛选组

数据分类是数据管理中必不可少的环节,面向地学数据共享,需要设计一套多元的数据分类与组织体系,满足用户需求和数据集成要求。

地球系统科学数据分类体系可采取两种模式:一是数据分类目录模式,二是为了便于数据检索的关键词表分类模式。数据分类目录模式主要用于后台数据管理和前台门户展示,为用户呈现一个完整的、可管理的数据目录体系。数据关键词表分类主要是为了平台内数据快速检索、导航而建立的规范的关键词分类词表,用于数据的筛选。根据地球系统科学数据的特征和用户检索的高频关键词,可将其分为13个一级类、71个二级类以及686个关键词^[9]。

数据导航按照数据一级和二级分类体系,各分类之下可以再建立数据筛选组。筛选组包含时间、空间、数据类型、空间分辨率、产生来源、学科类别等主题词。用户查找数据时可以按照更加精细化的导航查询并结合关键词筛选组,并将搜索结果按照语义关联程度和数据应用特征(例如数据相关性、数据使用热度、数据发布时间等)优化排序呈现给用户,同时给出与数据实体的关联,如图4所示。

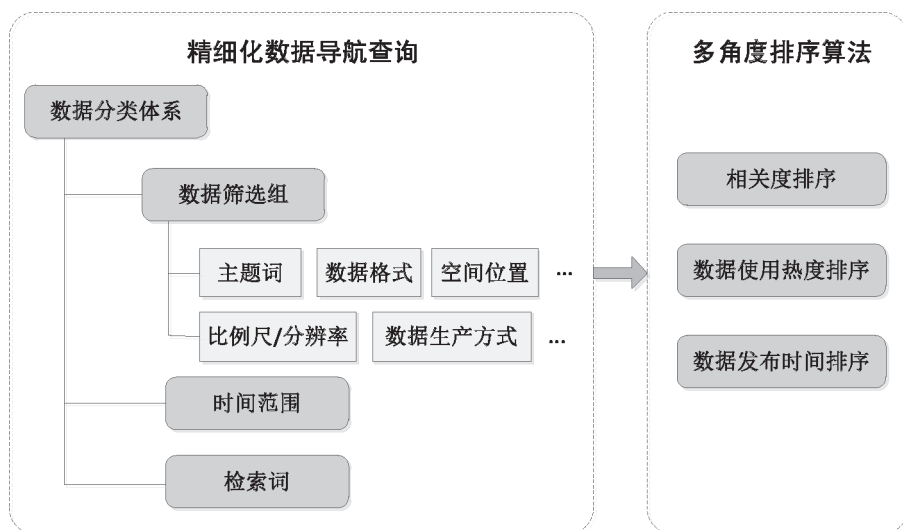


图4 数据导航分类和筛选

3 面向服务的国家地球系统科学数据共享平台

国家地球系统科学数据共享平台是国家科技基础条件平台之一。2003年,科技部科学数据共享工程启动了“地球系统科学数据共享网”项目,开展了地学数据共享平台建设实践,构建了国家地球系统科学数据共享平台,并于2005年纳入了国家科技基础条件平台。2011年,“地球系统科学数据共享网”通过财政部、科技部的联合认定,成为首批23个国家科技条件平台之一^[10]。平台在新的以“数据服务为中心”的大背景下对原有的平台系统进行了升级改造,引入了数据订单、数据引用声明等方式体现平台高效服务的意识,并通过对原有地球系统科学元数据、数据文档、数据分类等共享要素规范的修订更好地促进现有数据资源的整合与规范,推进数据组织与分类的多元化,突出体现简单明了的数据查询、浏览、访问机制与流程。平台还积极吸收并应用信息产业化领域的先进、主流技术体系,能够保障稳定运行的能力(资源响应时间和并发请求)并提高用户获取数据的效率,改善用户体验。

3.1 平台架构设计

地球系统科学数据共享平台(以下简称共享

平台)采用分布、一站式服务,按照“总平台—认证中心—分平台”的体系架构进行分布设计,由多个物理上分布、逻辑上统一的网络服务系统提供数据共享服务。为了集中管理数据资源,由总平台统一管理数据实体,同时推送给各分平台,形成“自上而下”的数据汇交模式。各个分平台则自动备份数据资源,保障与总平台的同步,如图5所示。

3.2 平台功能特点

纳入国家科技基础条件平台后,围绕“数据运行服务为中心”,共享平台加强了数据贡献者、数据使用者的权力义务的技术功能支持。针对数据用户,改善了数据检索功能,实现了“导航+筛选”一体化的数据查找途径,并通过主动推送服务,吸引了更多的科研用户群。数据申请采用订单模式,简化了申请数据的流程。采用“一站式”单点登录系统,方便用户在各个分平台之间的转换。平台总门户如图6所示。

(1) 数据汇交功能

数据汇交采用“六位一体”的模式,包含元数据、数据缩略图、数据文档、数据样例、数据实体、关联分类等描述信息。规范增强元数据特征描述,突出时空特征、数据贡献者、DOI、数据详细描述、数据引用方式,并采用“自动提

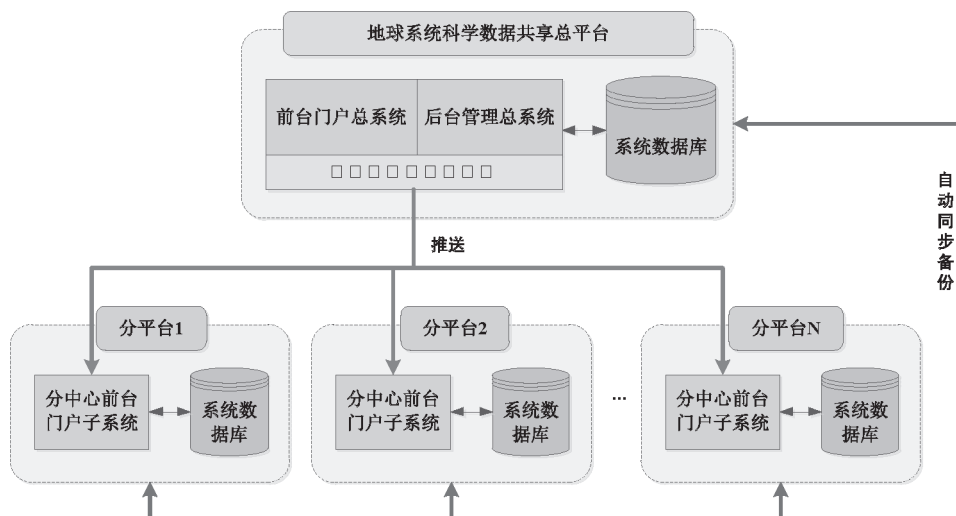


图5 地球系统科学数据共享分布式平台



图6 地球系统科学数据共享平台主页

示/自动完成”技术提高元数据的著录效率。数据提交时能够自动检验数据的完整性，并且通过设置不同的审核状态规范数据汇交过程。能够提供数据样例下载和数据在线预览以及相关文献链接功能，方便了用户文献检索。

(2) 数据检索功能

系统提供多种数据查找入口，可以通过主题分类、空间、时间、学科类别、关键词等精细化数据分类和导航筛选综合查找。提供数据推荐入

口，显著展示主要数据和特色数据，以采样相关度、数据使用热度、最近更新时间等多种排序方式展现数据，方便用户发现需要的数据。

(3) 数据订单功能

用户申请数据时引入了数据订单概念。数据获取均以数据订单方式处理，用户可以在一个订单中申请多个数据。数据申请处理过程通过电子邮件主动通知用户每步的数据处理情况，用户可以跟踪订单的状态获知数据申请处理的全过程。

(4) 第三方用户登录认证

在单点登录的基础上,系统引入第三方登录认证,用户既可以通过注册新用户创建一个账户,也可以使用其他凭证登录系统。除了可以使用平台注册账户登录外,还可以使用中国科技网通行证(如中科院邮件系统账号)以及QQ账号(腾讯微博账号)登录平台。通过第三方登录认证能够挖掘潜在的用户群体,吸引更多的科研人员和社会人员使用。

(5) 统计分析功能

统计分析功能主要提供用户分析统计、数据资源量统计、数据访问量统计和回访率统计等方面的功能,能够进行所有用户访问统计的对比查看与分析。通过对数据和用户的分析可以看出哪些用户贡献的数据较多,哪些数据的访问量较高,哪些时段的访问量较高,进而对系统进行加强和优化。

4 结论与展望

地学数据共享平台是推动地学数据共享事业不可缺少的技术组成部分,数据用户和数据资源是整个数据共享过程中的核心环节。本文针对地学数据共享过程中的关键问题,提出了地学数据共享的框架,从国家政策机制、共享服务规范、共享服务对象以及系统平台的实践与应用角度提出了地学数据共享建设的思路。

开放共享机制是促进科学数据共享的关键要素,要激励研究单位和科研人员贡献自己的数据,同时要保障他们的知识产权,使数据得到应有的价值。在数据共享过程中,要针对地学数据的共享要素,提供相应的规范条例,建立“六位一体”的数据汇交模式,保证数据资源的完整可靠。基于使用者、提供者、管理者是参与数据共享核心主体的角度考虑,可以将用户分为数据使用者、数据贡献者和数据管理者3个层次,明确用户的权利和义务,促进数据共享良性发展。合理定位地学数据共享平台的功能体系和支撑技

术,使得数据资源更具有应用价值和共享效果。

随着大数据时代的到来和云计算技术的发展,需要形成一个“人人都是数据的提供者,人人都是数据的使用者”的数据共享服务环境^[1]。从数据共享技术的长远发展与未来展望角度来考虑,面向数据使用者和数据贡献者,应该瞄准“开放、自由、自愿、平等”的数据共享模式,促进科学数据共享事业更健康、科学、全面的发展。

参考文献

- [1] 汪品先. 我国的地球系统科学研究向何处去[J]. 地球科学进展, 2003, 18(6): 837-851.
- [2] 孙枢. 对我国全球变化与地球系统科学研究的若干思考[J]. 地球科学进展, 2005, 20(1): 6-10.
- [3] 孙九林, 林海. 地球系统研究与科学数据[M]. 北京: 科学出版社, 2009.
- [4] Bai Yuqi, Di Liping, Wei Yaxing. A Taxonomy of Geospatial Services for Global Service Discovery and Interoperability[J]. Computer & Geosciences, 2009(35):783-790.
- [5] Shresta B. Mountain Knowledge Hub Initiative in the Hindu Kush-Himalayan Region[J]. Grazer Schriften der Geographie und Raumforschung, 2007, 43: 227-234.
- [6] Goodchild M F, Fu P, Rich P. Sharing Geographic Information: An Assessment of the Geospatial One-Stop[J]. Annals of the Association of American Geographers, 2007, 97(2): 250-266.
- [7] 胥燕婴. 中国国家空间数据基础设施建设的进展[J]. 中国测绘, 2000 (4): 6-8.
- [8] 刘润达, 褚文博, 诸云强. 国家科技基础条件平台运行服务阶段关键问题探析[J]. 现代情报, 2013, 32(11): 51-53.
- [9] 王卷乐, 林海, 冉盈盈, 等. 面向数据共享的地球系统科学数据分类探讨[J]. 地球科学进展, 2014, 29(2): 265-274.
- [10] 诸云强, 宋佳, 冯敏, 等. 地球系统科学数据共享软件研究与发展[J]. 中国科技资源导刊, 2012(6): 11-16.
- [11] 苗茹, 诸云强, 宋佳, 等. 基于云计算的地球系统科学数据共享研究与实践[J]. 地球信息科学学报, 2014, 16(2): 264-272.