

科技基础性工作专项数据汇交共享平台建设

杨杰^{1,2} 宋佳^{1,3} 诸云强^{1,3,4} 王筱萱¹ 孙凯^{1,2} 李威蓉⁵

(1.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;
2.中国科学院大学, 北京 100049; 3.江苏省地理信息资源开发与利用协同创新中心, 江苏南京 210023;
4.白洋淀流域生态保护与京津冀可持续发展协同创新中心, 河北保定 071002; 5.山东理工大学
建筑工程学院, 山东淄博 255000)

摘要: 科技基础性专项工作产生了大量涉及多领域、涵盖多类型的数据资源, 只有对这些数据资源进行汇交共享, 才能实现数据资源的增值, 更好地支撑科技创新、国家战略决策和社会经济发展。在充分分析科技基础性工作专项数据资源跨领域性、类型复杂、分散性、异构性等特点的基础上, 设计实现了科技基础性工作专项数据汇交共享平台。该平台基于自下而上、逐步集成的数据组织框架, 可有效实现面向多领域、多类型数据资源汇交共享, 促进科技基础性工作专项数据资源汇交共享的发展。

关键词: 汇交共享平台; 科技资源; 数据资源; 数据汇交; 数据共享

中图分类号: G203

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2017.05.007

Construction of Special Data Archiving and Sharing Platform for the Science and Technology Basic Work

YANG Jie^{1,2}, SONG Jia^{1,3}, ZHU Yunqiang^{1,3,4}, WANG Xiaoxuan¹, SUN Kai^{1,2}, LI Weirong⁵

(1.State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, CAS, Beijing 100101; 2.University of Chinese Academy of Sciences, Beijing 100049; 3.Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023; 4.Center for Collaborative Innovation in Baiyangdian Basin Ecological Protection and Sustainable Development of Beijing-Tianjin-Hebei, Baoding 071002; 5.School of Architecture Engineering, Shandong University of Technology, Zibo 255000)

Abstract: A large amount of data resources is produced from research projects funded by National Special Program on Basic Works for Science and Technology of China(NSPBWSTC). The content of these data resources covers multiple fields and data types. Sharing these data is good to make the data value-added and thereby foster the innovation in science and technology, national strategy, social and economic development. The study analyzes the characteristics of the cross-domain, complexity, dispersion and heterogeneity of these

作者简介: 杨杰(1990—), 男, 中国科学院地理科学与资源研究所硕士研究生, 研究方向: 地学模型数据匹配方法; 宋佳(1980—), 男, 中国科学院地理科学与资源研究所助理研究员, 博士, 研究方向: 地球信息科学(通讯作者); 诸云强(1977—), 男, 中国科学院地理科学与资源研究所研究员, 博士, 研究方向: 地学数据本体与共享、资源环境信息系统; 王筱萱(1983—), 女, 中国科学院地理科学与资源研究所工程师, 硕士, 研究方向: 科学数据共享与集成; 孙凯(1990—), 男, 中国科学院地理科学与资源研究所博士研究生, 研究方向: 地学本体及数据关联; 李威蓉(1991—), 男, 山东理工大学建筑工程学院硕士研究生, 研究方向: 地学数据来源及数据关联。

基金项目: 科技基础性工作专项重点项目“科技基础性工作数据资料集成与规范化整编”(2013FY110900)。

收稿时间: 2017年7月31日。

data resources, and designs and implements data archiving and sharing platform. The platform is based on the framework of bottom-up and step-by-step integration method. The data sharing towards multiple fields and data types is implemented in this platform, and also promotes the development of data archiving and sharing of NSPBWSTC.

Keywords: archiving and sharing platform, scientific and technical resources, data resources, data archiving, data sharing

1 引言

科技基础性工作一般指围绕国民经济和社会发展和科学研究的需求而开展的获取自然本底情况和基础科学数据、系统编研或共享科技资料和科学数据、采集保存自然科技资源、制定科学标准规范、研制标准物质等科学活动的统称^[1]。作为现代科技“大厦”的地基与基础，科技基础性工作体现着国家整体科技水平的重要方面，对于科技进步、经济与社会发展和国家安全具有不可替代的作用^[2]，具有长期性、稳定性、系统性、原始性和公益性的明显特点。

自1999年我国启动科技基础性工作专项以来，已经在气象、地球科学、生物学、农业、林业、医学、环境、材料等领域设置了大量项目，并通过这些项目采集产生了一批重要的科学数据、文字资料、图集典籍、科学规范、标准物质、样本样品等^[3-7]。这些成果属于科技资源的范畴，高度重视其开放共享有利于科技资源的优化配置、高效开发和利用^[8]，实现数据资源的增值，从而更好地支撑科技创新、国家战略决策和社会经济发展。为推动科技资源整合共享，2004年7月，科技部、国家发改委、财政部、教育部启动了国家科技基础条件平台建设^[9]，并经过多年的建设取得了积极进展和成效^[10]，相继在多个领域内建成了一批国家科技基础条件平台，例如：国家地球系统科学数据共享平台、气象科学数据共享网、国家农业科学数据共享中心、林业科学数据平台、国家人口与健康科学数据共享平台、标本资源共享平台、国家标准物质资源共享平台、国家标准文献共享服务平台等。这些平台或面向单一学科领域，或服务于单一数据资源类型，缺乏数据汇交管理功能。虽然其他一些依托

各类科学计划建立的平台包含了数据汇交管理的功能，但仅面向单一领域，例如：“973”计划资源环境领域项目数据汇交服务网^[11]仅面向资源环境领域的数据展开汇交共享。然而，科技基础性工作专项数据资源以项目形式组织汇交，其内容涵盖多个学科领域、数据类型复杂，对汇交共享平台跨领域项目、跨资源类型的数据获取能力以及面向管理人员的项目汇交管理功能都提出了较高的要求。因此，已有平台并不能直接运用于科技基础性工作专项数据资源的汇交共享。应根据科技基础性工作专项数据资源特点，研发一个综合的系统平台，力求为用户提供“一站式”的用户体验。鉴于此，本文拟对科技基础性工作专项数据汇交共享平台进行初步探讨。

2 科技基础性工作专项数据资源的特征

科技基础性工作专项产生的数据资源包括：数据、图集、志书典籍、标本资源、标准规范、论文专著、研究报告。数据是指在基础性工作中通过考察、观测、探测、监测、调查、试验、试验以及编撰等方式获取到的各类科学数据。图集是按照一定规则编制的图片的集合。志书是以地区为主，综合记录该地自然和社会方面有关历史与现状的著作，又称地志或地方志。典籍是古代重要文献的总称。标本资源是动物、植物、矿物等实物，经过各种处理后可以长久保存，并尽量保持原貌，藉以提供作为展览、示范、教育、鉴定、考证及其他各种研究之用。标准规范是领域内共同遵守的准则和依据。论文专著是指对各学术领域进行研究和描述学术研究成果的文章。研究报告是指在从事一种重要活动或决策之前，对相关各种因素进行具体调查、研究、分析，评估项目可行性、效果效益程度，提出建设性意见、

建议、对策等,为决策者和主管机关审批的上报文件。其中,图集、志书典籍、标准规范、论文专著、研究报告的数据资源需要将内容进行数字化,以电子格式文件为载体作为项目数据汇交的内容。标本资源分为8大类:植物种质资源、动物种质资源、微生物菌种资源、人类遗传资源、生物标本资源、岩矿化石资源、实验材料资源以及标准物质。实物不在项目数据汇交的范围,由项目承担单位或科技部指定的单位保存,但需要汇交内容为实物的基本信息描述,这些基本信息描述通过每类标本资源对应的描述规范进行约束。

基础性工作专项数据资源主要具有以下几个方面特点。

(1) 跨领域性。科技基础性工作专项数据资源涉及领域众多,不仅涵盖自然科学领域,而且包含社会科学领域。即使同一个基础性专项产生的数据资源有时也会跨越多个领域,例如:中国北方及其毗邻地区综合科学考察项目产生了气候、人口、社会经济等多个领域的的数据。

(2) 数据类型复杂。科技基础性工作专项数据资源的数据类型包括文档、表格、图片、数据库、矢量文件等类型,并且同一种数据类型又常常包含多种数据格式,如图片数据类型包含了jpg、tiff、geotiff、png等数据格式。

(3) 分散性。科技基础性工作专项数据资源以项目的形式组织,不仅具有地域分散性,而且具有内容分散性。地域分散性是指数据资源具有不同的空间位置属性,而内容分散性是指同一要素内容的数据资源可能分散在不同的项目中,例如:多个项目中含有降水数据。

(4) 异构性。科技基础性工作专项数据资源的异构性是指同一要素的数据资源采用的描述规范不一致。例如:采用不同分类体系的土地覆被数据。

3 科技基础性工作专项汇交数据的组织

目前,常用的一种做法是以元数据为核心实现数据检索。该方法通过元数据对资源的属性特

征进行描述并将描述结果进行发布,使用户可以通过关键词与元数据进行匹配,进而定位和获取相应资源^[12]。本文从数据检索和资源集成的角度,设计了自下而上、逐步集成的分层数据组织框架(图1)。

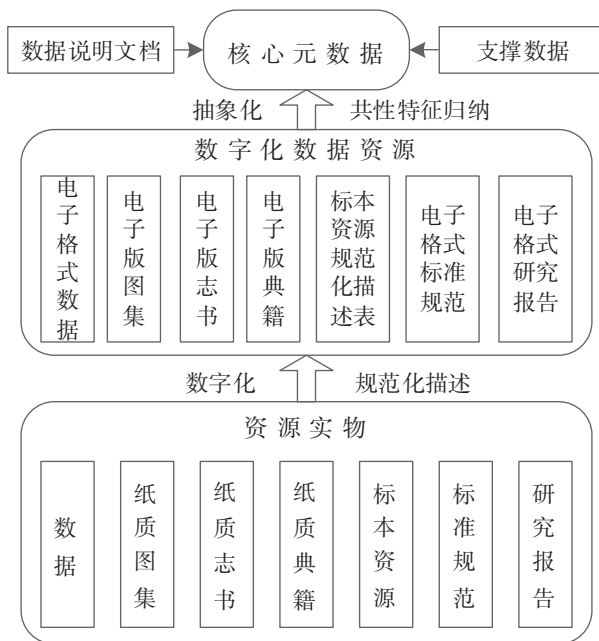


图1 数据组织框架

从图1可以看出,核心元数据的形成经过了两次集成的过程。第一次集成过程针对的是资源实体本身,包括数字化和规范化描述两种处理方式。数据、标准规范和研究报告数据资源的实物是指其包含的内容信息,需要经过数字化后以相应的电子格式文件为载体进行保存,这一步通常已经在项目实行过程中完成。图集、志书典籍数据资源的实物是指纸质版的图集及纸质版的志书典籍,同样需要经过数字化得到电子版图集和电子版志书典籍作为专项项目数据汇交的内容。标本资源实物即为八大类资源实物本身,这类数据资源需要汇交采用相关描述标准得到的规范化描述信息表,这种信息表实质上是一种面向标本资源的元数据。第二次集成过程针对的是第一次集成过程的结果,即数字化后的数据、图集、志书典籍、标准规范及标本资源规范化描述信息表,最终得到科技基础性工作专项汇交数据资源核心

元数据。核心元数据包括各类数据资源共性特征的简要描述和数据资源项目来源信息。同时，为了填补核心元数据内容中有关于数据资源个性特征的缺失，数据组织框架采用将数据与相关的数据文档和支撑数据进行关联的方式，辅助用户快速获取更为详细的数据信息。表 1 为核心元数据的内容项，一共包含 19 项，其中 16 项必填，3 项可选。

该数据组织框架中的核心元数据为用户提供了一个对各类数据资源访问的统一接口。通过统一的接口，用户首先可以访问到数据的基本描述信息，进而定位到特定类型资源的详细信息和相关项目信息，从而为汇交共享平台实现跨资源类型的数据检索提供了强大的支撑，为实现跨项目的数据检索奠定了基础。

4 汇交共享平台的设计

科技基础性工作专项数据汇交共享平台逻辑上可以将平台划分为 5 个层次，自下而上分别是基础设施层、数据层、功能层、服务层和用户层

(图 2)。

基础设施层是汇交共享平台正常运行的基本保障，主要包括硬件基础设施和软件基础设施。硬件基础设施首先是网络、服务器以及大量的存储。软件基础设施是保障共享平台运行的软件环境，包括一切软件正常运行依赖的操作系统和各类应用服务软件，例如数据库管理软件。

数据层是对平台中所有数据的逻辑抽象，包含了原始数字化文件、规范化描述数据、核心元数据以及平台业务数据。原始数字化文件是未经修改的通过项目汇交得到的原始数据。规范化描述数据是通过相关描述规范对标本资源数据类型描述的结果。核心元数据是对原始文件和规范化描述数据按照图 1 的数据组织框架进行集成得到的元数据。业务数据是平台在运行过程中产生的各类数据，如用户注册信息、数据下载订单数据等。

功能层是对平台所有功能等逻辑抽象，可按模块划分为检索、数据下载、项目汇交管理、用户中心和系统管理等部分。每个模块又包含了若

表 1 科技基础性工作专项汇交数据资源核心元数据

属性名称	属性描述	是否必填
资源标识	资源唯一标识符，由系统自动产生	必填
资源学科分类	资源所属的学科分类	必填
中文名称	反映资源内容及主要特征的中文名称	必填
英文名称	中文名称对应的英文名称	选填
资源描述摘要	资源内容、特征的简要描述	必填
关键词	反映数据集主要特征，便于检索的主题词	必填
资源类型	资源所属的类型（如图集、志书典籍等）	必填
资源格式	资源存储格式	必填
资源时间	资源内容的时间点或时间范围	必填
资源地点	资源内容采集的地理位置	必填
最新修订时间	资源的最新更新时间	必填
共享方式	完全开放共享、协议共享、暂不共享其中的一种	必填
资源质量描述	对资源质量相关属性的描述	必填
在线链接地址	在线获取或访问资源的网络地址	选填
缩略图	反映资源概貌、内容或特征的图片	必填
来源项目	产生数据资源的基础性专项项目信息	必填
资源负责方	产生资源或对资源拥有处置权并对资源质量负责的个人及单位信息	必填
资源管理方	保藏、管理和对外提供资源服务的单位及联系人信息	选填
元数据管理信息	项目数据汇交联络人相关信息	必填

干的功能,功能之间相互配合完成对平台内业务流程的支撑。平台的详细功能设计将在下一小节进行详细阐述。

服务层是对功能层的封装。将平台功能按照一定的粒度包装并发布成服务,为平台内部其他模块或平台外部其他系统进行调用,可实现平台功能的重用,降低平台代码冗余,提高平台开发速度。这些服务包括检索服务、数据服务、汇交管理服务、用户验证服务以及资源管理服务等。

用户层是对平台使用者的逻辑抽象。基础性共享平台将用户划分为游客用户、普通用户和管理员用户3类。游客用户不需要在平台中完成注册,只能对项目和数据信息进行查询浏览,不能下载数据。普通用户需要在平台中完成注册,登录后不仅可以对数据资源进行浏览,而且可以下载。管理员用户是平台管理者,可以对平台的资源进行日常管理,例如数据的发布和下架、新闻公告的发布等。

科技基础性工作数据汇交共享平台的功能可以分为五大模块,即检索功能模块、数据下载功能模块、项目汇交管理模块、用户中心模块以及系统管理模块(图3)。

检索功能包括项目检索、数据检索、资料检索和数据订单检索。用户可以通过项目检索功能根据输入的项目名称、项目时间、项目承担单位等信息查询到需要的项目。同时,项目检索的

结果中包含有项目包括数据的链接,用户可以进一步导航至具体数据资源,从而下载得到需要的数据。数据检索功能允许用户直接根据数据的名称、学科名称、数据资源类型等信息对数据进行检索,同时检索结果中还含有数据所属项目信息等链接,可以方便用户利用导航查询相关项目,并进一步浏览或下载同一个项目的相关数据。平台中的资料是指在基础性工作数据汇交过程中制定的标准规范、培训讲义等文件,数据订单是用户在数据下载过程中产生的业务数据,用户也可以通过检索功能对这两类信息进行检索。

数据下载功能模块是科技基础性工作数据汇交共享平台的核心模块,是实现数据资源共享的关键环节。模块包括数据下载、资料下载、软件下载、论文下载等功能。数据下载是指对数据、图集、志书典籍、标本资源规范化描述信息、标准规范、论文专著、研究报告的数字化数据资源进行下载,但需要用户提前在平台中注册并登录。数据下载的流程是:用户先选择需要的数据加入到数据推车中,再将数据推车中的数据生成数据订单,平台将根据用户的订单内容将数据进行压缩打包,用户可在数据订单列表中查询到订单处理信息,待压缩打包完成后用户可下载结果数据。软件和论文是指在基础性项目中辅助项目进行的软件工具和产生的论文成果,这两类数据不需要用户在平台中进行注册和登录,而是用户

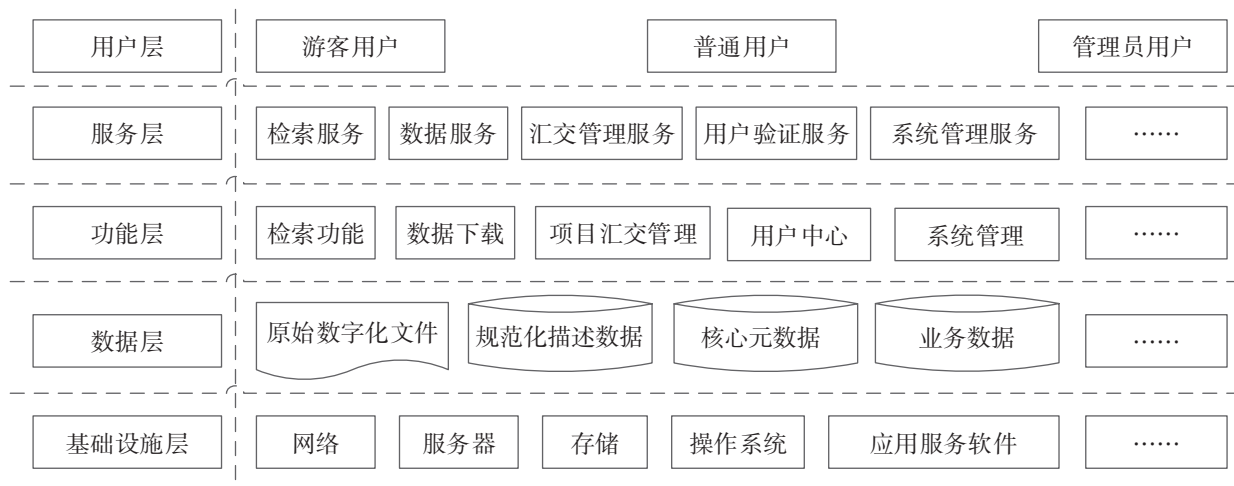


图2 科技基础性工作专项汇交共享平台框架

通过在项目检索结果中含有这两类数据资源的下载链接直接进行下载。资料下载也不要求用户在平台中注册和登录，用户可在资料下载页面进行下载。

用户中心功能是平台中跟用户相关功能的集合，包括用户注册、用户登录、用户信息修改、用户密码找回等功能。在注册过程中，用户需要填写用户名、用户真实姓名、密码、电子邮箱、联系电话、工作单位等信息，平台会根据用户的输入实时判断用户名和电子邮箱是否已经被其他用户注册，并提示用户。注册完成后，用户可以根据用户名和密码进行登录。通过用户信息修改功能，用户可以对自己的注册信息进行更新。利用用户密码找回功能，通过官方邮件向用户注册的邮箱发送密码重置链接，用户通过该重置链接可实现密码重置，从而应对密码遗忘的情况。

项目汇交管理模块面向管理员用户，是协助管理员管理项目数据资源汇交过程的功能模块，包括数据汇交审核、汇交进度查询、汇交进度管理。基础性项目数据汇交过程包括汇交方案和元数据提交、实体数据提交和验收三大阶段，汇交内容在每个阶段中只有被审核通过后才能进入下一阶段。管理员通过数据汇交审核功能决定数据汇交过程能否进入下一阶段。通过汇交进度查询功能，管理员可查询指定项目的汇交进度和处于指定汇交阶段的所有项目。如汇交内容通过审

核，则管理员用户可通过汇交进度管理功能修改项目汇交的状态信息，更新项目汇交进度。

系统管理功能模块面向管理员用户，包含数据管理、新闻公告发布和数据统计功能。管理员用户通过数据管理功能可以实现数据、资料、软件、论文等资源的发布和取消发布。新闻公告发布功能负责对平台新闻公告版面内容的管理。数据统计功能主要对平台中的项目数、数据量、平台访问量等数据进行统计展示。

5 平台构建与关键技术

构建科技基础性工作专项数据汇交共享平台采用了B/S（浏览器/服务器模式）结构，基于Java语言进行功能模块的开发。平台整体采用当前优秀的MVC（模型、视图、控制器）框架SpringMVC，便于其他技术框架的集成，简化平台维护和功能扩展流程。该平台的前端页面是直接和用户交互的接口，美观大方的页面设计有利于用户体验的提升。平台在前端页面的搭建采用美国著名社交网站Tiwttter设计师开发的Bootstrap框架。该框架的组建不仅功能强大而且在不同浏览器上兼容性良好，即减少了平台开发的时间也增加了平台的稳定性。数据持久化是将内存中的数据保存到磁盘上加以固化，这一过程大多是通过数据库来完成。平台采用支持普通SQL查询、存储过程和高级映射到优秀持久层框架MyBatis

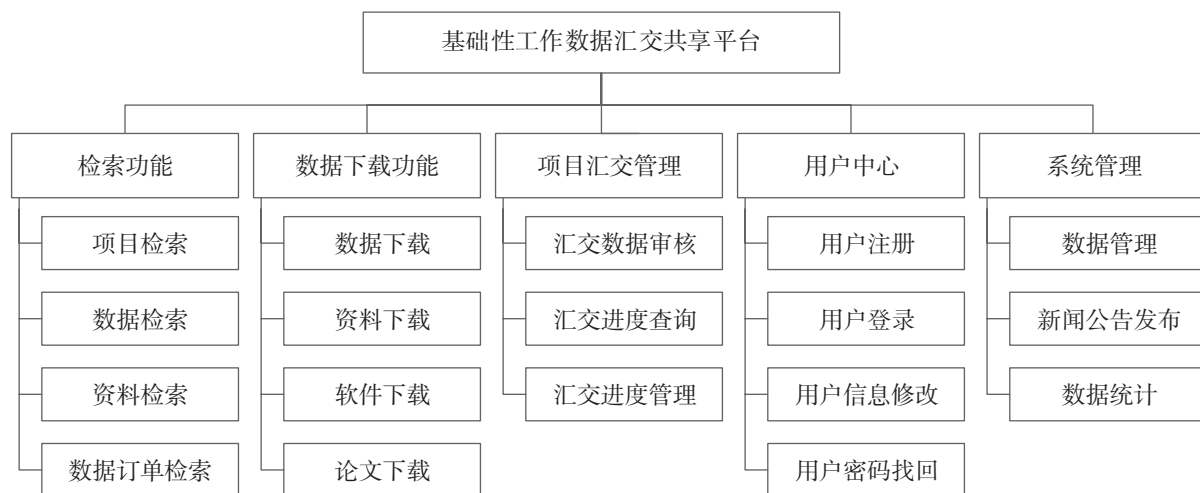


图3 科技基础性工作专项汇交共享平台功能图

完成系统业务数据的持久化。在数据库方面,平台同时采用了关系型数据库MySQL和非关系型数据库MongoDB,两者相互配合支撑平台的数据存储。关系型数据库主要用于存储结构规则的规范化描述数据、核心元数据和平台业务数据;非关系数据库用于存储复杂的原始数字化文件结构信息,以便于快速的文件浏览。

在构建平台中采用了以下几项关键技术。

(1) 支持跨领域项目、跨资源类型、递进式检索的数据关联方法

本文根据科技基础性工作专项数据资源特点,设计了从资源实体到数字化数据资源再到核心元数据的自下而上、逐步集成的数据组织框架。在该框架中核心元数据的结构设计对实现跨领域项目、跨资源类型、递进式数据检索功能至关重要。在核心元数据内容中不仅包含了各种数据资源类型的共性特征,而且包含了项目来源信息。其中,数据资源共性特征实现了核心元数据与数据资源的关联,项目来源信息是核心元数据与项目信息连接的纽带。基于该数据组织框架,在纵向上,用户可以先检索项目信息,再检索与项目相关的元数据信息,进而检索到数据资源及其相关信息,从而实现了递进式数据检索;在横向上,用户可以检索相同项目下的不同数据以及不同项目下的相同或相似数据,实现了跨项目、跨资源类型的数据检索。

(2) 高准确度的数据检索方法

在数据的检索过程中,用户通常是在平台数据搜索页面的搜索框中输入与目标数据资源相关的关键词,平台则通过关键词与数据库中的核心元数据内容进行匹配来寻找用户需要的数据。这一做法的弊端是查询结果的质量高度依赖于关键词与元数据分词时所基于的词库,由于本文所涉及的数据资源为科技资源,专业术语众多,致使这个问题尤为突出。因此,本文提出的解决方案是先抽取核心元数据库中所有数据的关键词内容,将其加入到IKAnalyzer中文分词工具的扩展词库中,再根据该词库将用户输入的关键词和核心元数据内容进行分词,进而进行匹配。除此之

外,平台还对用户的搜索关键词进行记录和统计,并将用户搜索频率较高的词同时加入到词库中,从而大大增加了数据检索的准确度,提升了用户体验。

(3) 大数据量文件信息的流畅展示方法

基础性工作专项汇交的原始数据组织结构复杂,不仅存在文件夹多层嵌套的情况,而且数据文件个数从几个到几万个不等,浮动很大。在大数据量的情况下,首先要在用户的浏览器客户端一次性展示上万个文件信息,不仅对系统数据读取是一个巨大的挑战,而且会严重降低数据浏览的用户体验。如果采用关系型数据库对这些文件夹和文件的相对关系进行保存,再通过数据库的关联查询实现数据组织结构的重组并分页展示,上万级别的频繁关联查询势必降低了平台的服务性能。本文利用这些数据资源原始文件的稳定性(数据组织结构一般不会变动),采用非关系型数据库MongoDB对这些文件夹和文件的相对关系以json文件进行存储,并进行了分页组织。该方法实质上是利用一种静态化处理的思想,避免了关系型数据库的频繁关联查询,从而提高了客户端的数据浏览速度,保证了平台的性能。

本文基于上述技术选型开发了科技基础性工作专项数据汇交共享平台的原型系统,如图4所示。

基于汇交共享平台,项目主管部门可以对专项项目数据汇交进展情况进行监督管理,项目单位及其他科研人员可以灵活高效地进行跨项目、跨类型的专项项目数据检索和访问。

表2内容是在平台中使用“气温”检索的部分结果对应核心元数据记录的部分属性。需要补充说明的是,限于文章篇幅,表2并未列出检索结果的核心元数据“关键词”属性,而在表2中的第3条结果正是由于“关键词”属性含有“气温”一词,故被平台检索到。根据检索结果,3条检索结果虽均为气温相关数据,但是来源于不同的项目,所属学科跨越了地理学、大气科学和海洋科学,资源类型既有数据又有图集,资源格式也各不相同。基于检索结果,用户可进一步获



图 4 科技基础性工作专项数据汇交共享平台首页

表 2 平台中使用“气温”关键字检索的部分结果

中文名称	来源项目	资源学科分类	资源类型	资源格式
青藏高原1公里格网气温数据库（1960、1970、1980、1990、2000年）	格网化资源环境综合科学调查规范	地理学	数据	grid
中蒙俄考察区1980—2010年气温数据集	中国北方及其毗邻地区综合科学考察	大气科学	数据	coverage
2009—2012年南海海洋科学断面考察AWS气象观测图集	南海海洋断面科学考察	海洋科学	图集	bmp

取对应的数据资源。因此，平台能够为用户提供良好的跨项目领域、跨资源类型共享服务。

6 结语

本文首先分析了科技基础性工作专项数据资源的特点，然后结合相关数据资源描述规范，阐述了平台数据层的组织框架，进而从平台的逻辑层次和功能模块角度详述了平台的设计过程，接着从技术选型和关键技术方面对平台的实现过程进行了剖析，最终对平台的不足进行了总结和展望。

（1）在分析科技基础性工作数据汇交共享平台各类数据资源特点的基础上，通过分层数据组织框架实现不同领域项目、不同类型数据资源的集成，为用户提供“一站式”的服务体验。

（2）实践证明，平台可有效实现多领域、多资源类型的数据资源汇交共享，突破单一领域、单一类型数据资源共享平台的局限性，促进科技基础性工作专项数据汇交共享的发展。

（3）平台使用传统的数据关联方法，没有实现数据资源间的语义关联，无法向用户提供智能检索体验，也未能展开深度的数据挖掘并向用户提供潜在的信息，有待今后进一步改进和完善。

参考文献

- [1] 中华人民共和国科学技术部. 国家科技基础性工作专项“十二五”专项规划[Z]. 2012.
- [2] 中华人民共和国科学技术部. 国家“十五”科技基础性工作专项实施意见[J]. 中国基础科学, 2001(8): 31-34.

（下转第67页）

基本信息、数据实体、数据文档、论文专著和辅助软件工具等进行全面的审查。最后, 基于本模型开发了科技基础性项目汇交数据质量审查系统。除了数据审查功能之外, 还开发了用于约束规则配置的工具界面, 添加了审查日志和审核意见导出等功能, 方便对同一数据集的持续跟踪审查。

通过基于约束规则的科技基础性数据质量审查模型的研究与实现, 能够提高科技基础性工作项目汇交数据质量审查的质量和效率, 节约数据审核人员的时间, 使数据审核人更集中精力去审核一些更深层次的数据质量问题。科技基础性项目数据汇交是一项长期持续性的工作。目前, 数据质量审查模型还更多地侧重于形式方面的审查, 随着数据汇交工作的开展而不断深入, 数据质量审查模型将深入数据内容, 建立经验模型, 逐步实现数据质量的自动审查和意见反馈。

参考文献

- [1] SADIQ Shazia, INDULSKA Marta. Open data: quality over quantity[J]. International Journal of Information Management, 2017, 37(3): 150–154.
- [2] 陈远, 罗琳, 沈祥兴. 信息系统中的么据质童问题研究[J]. 中国图书馆学报, 2004(1): 48–50.
- [3] 蒋萍, 田成诗. 全方位、立体性数据质量概念的建立与实施[J]. 统计研究, 2010, 27(12): 8–15.
- [4] 许涤龙, 龙海跃. 欧盟数据质量评估框架及其对我国的启示[J]. 统计与决策, 2013(8): 4–7.
- [5] TAGGARTA Jane, LIAWA Siaw–Teng, YU Hairong. Structured data quality reports to improve EHR data quality[J]. International Journal of Medical Informatics, 2015, 84(12): 1094–1098.
- [6] PRESSER Karl, HINTERBERGER Hans, WEBER David, et al. A scope classification of data quality requirements for food composition data[J]. Food Chemistry, 2016, 193: 166–172.
- [7] 徐启恒, 张新长, 张兴飞. GIS数据检查与质量控制系统的设计与实现[J]. 测绘通报, 2012(5): 38–40.
- [8] 王兴, 朱定真, 苗春生. 基于规则引擎的多元大气信息数据质量检查方法[J]. 南京信息工程大学学报(自然科学版), 2011, 3(3): 238–243.
- [9] 杨家芳. 基于规则引擎的基本农田划定内业数据质量检查方法研究[D]. 杭州: 浙江大学, 2014.
- [10] MERINO Jorge, CABALLERO Ismael, RIVAS Bibiano, et al. A data quality in use model for big data[J]. Future Generation Computer Systems, 2016, 63: 123–130.
- [11] 曹一化, 刘旭, 许增泰, 等. 自然科技资源共性描述规范[M]. 北京: 中国科学技术出版社, 2006: 1–86
- [12] HELLMANN D. The python standard library by example[M]. Indianapolis, Indiana: Addison Wesley, 2011: 1–10.
- (上接第59页)
- [3] 胡光晓. 提升我国地层研究知名度展现我国地层工作最新成果:《中国岩石地层名称辞典》[J]. 科技成果管理与研究, 2015(8): 79–80.DOI: 10.3772/j.issn. 1673–6516.2015.08.029.
- [4] 王训练, 徐均涛. 古生物学研究的新成果: 中国古生物志与中国各门类化石编研[J]. 中国基础科学, 2002(5): 18–23.DOI: 10.3969/j.issn.1009–2412.2002.05.004.
- [5] 吴小红. 京族医药调查报告[J]. 中国民族医药杂志, 2016, 22(3): 57–59.DOI: 10.16041/j.cnki.cn15–1175.2016.03.037.
- [6] 徐福荣, 戴陆园, 韩龙植. 21世纪初云南稻作地方品种图志[M]. 北京: 科学出版社, 2016.
- [7] 张芳, 王思. 中国农业古籍目录[M]. 北京: 北京图书馆出版社, 2003.
- [8] 徐冠华. 加强科技资源研究促进科技资源共享[J]. 中国科技资源导刊, 2008, 40(3): 3–5.DOI: 10.3772/j.issn.1674–1544.2008.03.001.
- [9] 叶玉江. 加强科技平台工作推进科技资源管理[J]. 中国科技资源导刊, 2015, 47(2): 1–6.DOI: 10.3772/j.issn. 1674–1544.2015.02.001.
- [10] 国家科技基础条件平台中心. 国家科技基础条件平台发展报告: 2011–2012[M]. 北京: 科学技术文献出版社, 2013.
- [11] 王卷乐, 杨雅萍, 诸云强, 等. “973”计划资源环境领域数据汇交进展与数据分析[J]. 地球科学进展, 2009, 24(8): 947–953.DOI: 10.3321/j.issn: 1001–8166.2009.08.013.
- [12] 王建涛, 朱龙文. 基于XML元数据描述的空间数据共享管理平台的实现与应用[J]. 测绘工程, 2007, 16(1): 12–15.DOI: 10.19349/j.cnki.issn1006–7949. 2007. 01. 003.