

基于黄河流域专题数据的国家科学数据中心 关联模型构建研究*

孙叶华¹ 刘桂锋^{1,2} 陈帅印¹

(1. 江苏大学科技信息研究所, 镇江 212013; 2. 江苏大学图书馆, 镇江 212013)

摘要: 国家科学数据中心是科学数据管理的重要载体, 是科学数据汇集管理、开放共享的重要基础设施。基于黄河流域专题数据, 设计关联元数据, 研究国家科学数据中心科学数据集元数据的关联关系, 构建国家科学数据中心关联模型, 尝试从完善自然资源监测数据、提升流域内灾害防控与应急能力、助力生态修复工程实施、建立一体化生态环境智能监测系统4个方面推动黄河流域的生态保护与修复工作。关联模型的构建研究有助于实现不同领域科学数据之间的互操作, 增强数据流通, 促进科学数据共享, 推动科技创新发展。

关键词: 科学数据; 元数据; 黄河流域; 关联模型; 数据管理

中图分类号: G250 **DOI:** 10.3772/j.issn.1673-2286.2022.08.001

引文格式: 孙叶华, 刘桂锋, 陈帅印. 基于黄河流域专题数据的国家科学数据中心关联模型构建研究[J]. 数字图书馆论坛, 2022 (8) : 2-10.

科学数据是指人类社会科技活动所产生的基本数据, 以及按照不同需求而系统加工的数据产品和相关信息^[1], 其价值的实现主要体现在对科学数据资源的深度挖掘和广泛应用上。随着科研活动的多样性以及科学数据的重要性发展, 科技创新越来越依赖数据, 科学数据的开放、共享也逐渐成为国内外相关学者及机构的研究重点。例如, Science Europe在面对海量的科学数据时, 选择将FAIR原则作为管理和共享科研数据的基础^[2]。

我国自2001年底正式启动“科学数据共享工程”以来, 为了充分发挥我国科技数据资源的作用, 在资源环境、农业、人口与健康等领域的24个部门开展科学数据共享工作, 启动9个科学数据共享试点^[3]。2011年, 23个科技平台被认定为国家首批科技基础条件平台, 进一步推动我国科技资源的整合共享与高效利用。2014年, 正式发布《科技平台标准化工作指南》《科技平台元数据标准化基本原则与方法》《科技平台元数据注册与管理》《科技平台资源核心元数据》, 从元数据等方面

对科技平台的数据进行统一的管理, 使得科学数据在开放共享以及应用的过程中更加规范化、标准化。2015年, “十三五规划”中首次提出“实施国家大数据战略, 推进数据资源开放共享”^[4]。2018年, 科技部、财政部正式发布《国家科技资源共享服务平台管理办法》, 强调利用科学数据等科技资源在国家层面设立专业化、综合性公共服务平台的必要性。同年, 国务院办公厅颁布的《科学数据管理办法》确立了科学数据“开放为常态、不开放为例外”的共享原则, 明确了要加强对科学数据的开放共享和管理使用^[5]。

科学数据共享通过提供可靠且可信赖的数据来显著改善科研过程, 已成为数据驱动科学的基本服务^[6]。目前, 我国科学数据共享平台存在以分散、凌乱、孤立为特点的“数据孤岛”现实问题。其中, 国家科学数据中心是由国家按照统一标准和规划设立的大型数据管理设施, 承担着我国科学数据管理共享、应用和支撑关联学科发展的重任, 具有领域代表性、行业领导性、技术先进性和数据公开性, 能够体现出我国科学数据

* 本研究得到江苏大学科研课题立项项目“国家科技基础条件平台关联模型构建研究——以农业平台为例”(编号: 21C310)资助。

共享平台的较高水平, 因此选择国家科学数据中心进行关联研究, 以期促进科学数据的开放与共享。

1 国内外研究现状

1.1 科学数据元数据研究现状

国内外不同学者主要从科研领域的元数据标准、元数据体系、元数据应用、元数据关联及互操作等方面开展科学数据元数据的相关研究。Qin等^[7]探讨科学数据元数据的功能和架构需求, 以支持数据管理、数据质量控制、数据发现和数据使用; 刘峰等^[8]分析了当前科研领域6种典型的元数据标准, 在统计的基础上设计并构建科学数据通用元数据规范; 司莉等^[9]以美国开放政府数据网站Data.gov中的元数据标准为例, 分析其元数据的体系及具体标准, 归纳总结美国开放政府数据元数据标准的体系结构; Si等^[10]调查研究在USA News排位世界前100名的大学图书馆科学数据服务内容, 发现28.7%的高校图书馆提供了科学数据的元数据相关服务, 包括元数据的简介以及创建等; 黄文碧^[11]详细阐述元数据关联在馆藏资源聚合研究中的重要性, 从元数据仓库构建、元数据映射和元数据关联等方面进行分析与应用; 贾欢^[12]选取地球科学相关领域的科学数据元数据标准, 研究元数据之间的互操作方法, 通过本体实现科学数据元数据之间的语义互操作。

1.2 科学数据关联与融合研究现状

科学数据的关联与融合是科学数据在科研全过程中达到互操作、互利用的必要过程。目前科学数据关联与融合相关研究主要围绕关联数据、文献数据与科学数据的融合等方向进行。Kauppinen^[13]在2011年发起建立linked-science.org, 提出关联科学(Linked Science), 将科学数据进行关联, 实现数据的透明化; 游毅等^[14]构建基于关联数据的科研数据资源共享模式, 包括科研关联数据的创建发布和科研共享数据网络的构建; 庄倩等^[15]在关联数据概念模型的基础上, 提出了科学数据组织模式, 该模式利用RDF描述特征, 通过本体映射和互操作来实现语义互联关系的搭建, 从而进一步确立了科学数据关联融合的新方向; 陆颖等^[16]从图书馆服务研究的角度, 基于多个层面对文献数据与科学数据融合进行调研, 针对各文献数据与科

学数据融合服务中的用户与学科服务之间关系, 提出了合作、支撑、服务以及共享互动的关系网络; 张晓霞^[17]对科学文献中生物本体间的情感关系表达、潜在关系抽取以及异源异构数据融合三方面进行研究, 使异源异构数据能实现集成信息服务, 达到跨异构库知识发现的研究目的。

1.3 科学数据共享平台研究现状

科学数据的关联融合研究推动了科学数据的开放、共享进展, 而科学数据共享中心的出现是科学数据相关研究不断推进的结果。许多国家都颁布了多项科学数据共享政策法规, 用以促进与规范科学数据的共享^[18]。目前有许多学者对科学数据共享平台进行调研, 其研究内容侧重于科学数据共享平台建设、科学数据组织、科学数据管理与共享等方面。Plantin等^[19]讲述数据存储、共享和管理平台案例研究, 科学数据共享平台正在逐步取代和重新配置传统学术交流的基础规范与技术; 黄铭瑞等^[20]梳理美、英及中国的国家科学数据中心的建设布局, 归纳并提出其管理模式与评价方式, 通过对比分析提出我国科学数据中心的发展演化模式; 卢逸航等^[21]调研各种科学数据互操作模式与技术适用条件, 结合我国科学数据中心的特点, 提出7种科学数据中心间互操作的模式。Curtis等^[22]研究制定一个系统框架, 评估科学数据共享平台的社会影响; 司莉等^[23]调查6个国家科学数据中心, 了解各平台数据的学科分布与数据量, 从数据组织方式、数据描述、数据检索3个方面分析平台数据组织现状, 发现其中不足, 从而提出平台优化建议; 刘小宇等^[24]以国家科技基础条件平台为例, 分析科学数据共享平台的被引用状况, 发现在平台建设和科学数据引用方面存在的问题, 以期能规范化建设平台和标准化引用科学数据; Thomas等^[25]详细介绍用户如何使用SPARK平台共享数据, 以及如何利用平台数据来建立假设; 李晓萍^[26]以农业科学数据共享平台为例, 对农业科学数据共享平台现状进行阐述, 并结合多种研究方法对影响用户持续使用农业科学数据共享平台的主要因素进行了研究, 从而基于用户的使用影响因素提出完善农业科学数据共享平台的建议。

1.4 研究现状述评

目前, 科学数据元数据、科学数据的管理与关联

融合、科学数据共享平台的理论研究较为丰富,但是实践应用方面还较为薄弱。由于我国国家层面的科学数据共享平台标准颁布较晚,而在此之前各数据平台已经结合自身数据特征制定了一系列标准,导致缺乏数据标准的通用性,因此出现大量的独立“数据烟囱”的状况。与此同时,不同领域的科学数据共享平台之间缺乏协调,难以实现不同科学数据之间的互操作,如各平台采用各自的学科元数据标准、平台间元数据字段不尽相同、通用的核心字段较少等。这不仅会影响数据质量,也会给不同平台之间数据的流通与整合造成不便。

因此,本文将关联数据技术的思想引入科学数据研究中,基于黄河流域生态保护和高质量发展跨平台专题数据,设计关联元数据,对国家科学数据中心的科学数据集元数据进行关联研究,从而构建国家科学数据中心关联模型,探究不同领域科学数据的互操作性、互相关性,有助于促进科学数据之间的协调性、增进数据流通的通用性,从而达到科学数据在科研过程中的应用性、在科研成果中的可转化性以及在不同科学数据中心之间的可流通性。

2 数据来源

2.1 平台概况

本研究的数据来源于国家科技基础条件平台上的国家科学数据中心。前期采用网络调研法,对20个国家科学数据中心的平台建设、数据资源、数据服务等基本情况进行调查。调查发现,物理、生态、农业、林业、气象等多个领域建设了国家科学数据中心,覆盖范围广泛,数据资源丰富,数据量庞大,中心还提供数据检索、数据专题、数据汇交、数据共享等多种数据服务。但这些数据中心成立时间不一,发展成熟度也有区别,同时存在数据分散、数据标准格式不一致、数据通用性差、数据利用率较低等问题。

在研究各个国家科学数据中心科学数据时,发现国家地球系统科学数据中心、国家农业科学数据中心、国家林业和草原科学数据中心、国家气象科学数据中心合作共建了黄河流域生态保护和高质量发展跨平台专题(以下简称“黄河流域专题”),其元数据通用性强,可扩展性好,故选择该专题数据为突破口进行国家科学数据中心关联研究。

2.2 黄河流域生态保护和高质量发展跨平台专题

党的十八大以来,以习近平同志为核心的党中央,对于人类文明在不断发展的过程中形成的规律进行了深刻总结,在我们国家的社会主义建设的总体布局中,加入生态文明建设^[27]。习近平总书记多次实地考察黄河流域生态保护情况,并作出重要批示。国家地球系统科学数据中心、国家农业科学数据中心、国家林业和草原科学数据中心、国家气象科学数据中心积极响应并落实《中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议》,联合开展黄河流域“三生”(生态-生产-生活)协同发展研究,建立黄河流域生态保护与高质量发展专题库,为促进黄河流域生态保护与高质量发展国家战略实施、保障黄河长治久安提供全方位科技支撑具有重要意义^[28]。截至2022年7月,黄河流域专题共拥有3 884个科学数据集,可按主题分为林业、农业、气象气候、地理、环境、生态、对地观测、资源、地质九大类型,主要集中在林业和地理两大主题(数据集数量分别为1 882个和1 083个);从数据集所属数据中心看,绝大部分的科学数据集由国家地球系统科学数据中心、国家林业和草原科学数据中心提供,来源较为单一,且大部分的气象气候数据以及地理、环境、生态、对地观测、资源、地质6种主题科学数据均由国家地球系统科学数据中心提供,究其原因,可能是由于国家地球系统科学数据中心涵盖大气圈、水圈、岩石圈、陆地表层等18个一级学科资源,涉及领域广泛。

经调查研究发现,国家生态科学数据中心和国家对地观测数据中心的科学数据集里可供关联的元数据较多,且元数据具备典型特色,可以为黄河流域专题提供良好的数据服务。因此,本研究基于黄河流域专题数据,以其元数据为切入点,增加国家生态科学数据中心和国家对地观测数据中心相关数据,研究国家地球系统科学数据中心、国家农业科学数据中心、国家林业和草原科学数据中心、国家气象科学数据中心、国家生态科学数据中心和国家对地观测数据中心之间的关联关系,共同构建国家科学数据中心关联模型,扩大科学数据的流通与共享,丰富黄河流域专题数据的来源与类型,有助于从不同视角完善黄河流域相关科学数据,有利于促进黄河流域生态保护和高质量发展。

3 国家科学数据中心关联模型构建

3.1 关联元数据元素

构建国家科学数据中心关联模型,其核心在于实现数据中心之间的数据关联和共享,首先基于黄河流域专题数据的元数据确定关联元数据元素,然后利用关联元数据元素实现6个国家科学数据中心科学数据的关联。

经调查,黄河流域专题数据元数据分别包括数据集名称、CSTR、主题词、数据覆盖范围、数据贡献者、所属国家科学数据中心、数据摘要、质量控制信息。元数据元素,是元数据最基本的信息单元,采用一定的元数据标准来描述数据的组织结构和内容特征,从而实现对大量数据的高效管理。元数据元素设计应遵循复用性、可扩展性、互操作性等元数据设计的一般原

则,以保证元数据的结构合理、元素可用,保证经描述的数据资源可共享。本研究根据这些元数据转化设计关联元数据元素依次为题名、资源标识、主题、覆盖范围、创作者、资源来源、资源描述、数据质量。以关联元数据元素为媒介,将6个国家科学数据中心的元数据进行关联得到6个国家科学数据中心的元数据的关联关系(见表1)。其中,题名、资源标识、主题、覆盖范围、创作者、资源来源、资源描述复用的是DC元数据标准,其元数据简单、通用、可扩展性强;数据质量是参考《林业科学数据元数据标准》中的数据质量信息与《气象数据集核心元数据》的数据质量描述后自定义的元数据。

同时,将这6个国家科学数据中心涵盖的元数据元素以此划分为通用元数据元素和专门元数据元素(见表2)。

表1 6个国家科学数据中心的元数据关联关系

序 号	关联元数据元素	黄河流域专题元数据	所属国家科学数据中心						复用标准
			国家地球系统科学数据中心	国家农业科学数据中心	国家林业和草原科学数据中心	国家气象科学数据中心	国家生态科学数据中心	国家对地观测科学数据中心	
1	题名	数据集名称	名称	数据库名称	名称	数据名称	中文标题	名称	DC
2	资源标识	CSTR	CSTR标识	科技资源标识	CSTR标识	-	CSTR标识符	CSTR	DC
3	主题	主题词	主题词	关键词	关键词	关键词	关键词	关键词	DC
4	覆盖范围	数据覆盖范围	空间位置	涉及区域	-	空间范围	空间范围	空间位置	DC
5	创作者	数据贡献者	数据贡献者	生产者	-	-	贡献者	-	DC
6	资源来源	所属国家科学数据中心	-	-	-	-	-	-	DC
7	资源描述	数据摘要	数据摘要	数据描述	数据描述	-	摘要描述	数据摘要	DC
8	数据质量	质量控制信息	-	-	-	数据质量描述	-	数据质量信息	自定义

*注:“-”表示无此项

表2 6个国家科学数据中心的元数据元素

中 心	通用元数据元素	专门元数据元素
国家农业科学数据中心	题名、资源标识、主题、覆盖范围、创作者、资源描述	数据共享方式、数据类型
国家林业和草原科学数据中心	题名、资源标识、主题、资源描述	元数据提交方、属性表字段数据、属性表记录数
国家气象科学数据中心	题名、主题、覆盖范围、数据质量	更新频率、共享级别、数据源、数据ID
国家地球系统科学数据中心	题名、资源标识、主题、覆盖范围、创作者、资源描述	学科分类、主题分类、DOI标识
国家生态科学数据中心	题名、资源标识、主题、覆盖范围、创作者、资源描述	内部管理标识符、PID标识符、空间分辨率、时间分辨率、共享服务、关联出版论文
国家对地观测科学数据中心	题名、资源标识、主题、覆盖范围、资源描述、数据质量	平台名称、仪器名称、产品级别、数据投影信息

3.2 关联模型构建

根据设计的元数据元素,首先确定核心类为黄河流域科学数据集与国家科学数据中心,其元数据为子类,同时增加流域范围和主题类型为核心类,丰富关联模型,其分类类型为自己的子类;其次,以元数据之间的关联关系确定类与类、类与子类、子类与子类之间的对象属性,关联元数据中的覆盖范围子类与流域范围类关联,资源来源子类与国家科学数据中心类关联,9种主题科学数据与其所属的国家科学数据中心关联;最后,利用Protégé本体编辑工具对其进行关联模型的可视化呈现。

图1是基于黄河流域专题数据构建的6个国家科学

数据中心关联模型。其中,虚线框内是原先实际联合建设黄河流域专题的4家平台,即国家地球系统科学数据中心、国家农业科学数据中心、国家林业和草原科学数据中心、国家气象科学数据中心。流域范围数据集包括黄河全流域、黄河上游、黄河中上游、黄河中游、黄河中下游的数据以及分省数据。主题类型中农业科学数据关联至国家农业科学数据中心,林业科学数据关联至国家林业和草原科学数据中心,气象气候数据关联国家气象科学数据中心和国家地球系统科学数据中心,对地观测数据关联至国家对地观测科学数据中心,生态科学数据关联至国家生态科学数据中心,其余地理、地质、资源、环境等科学数据关联至国家地球系统科学数据中心。

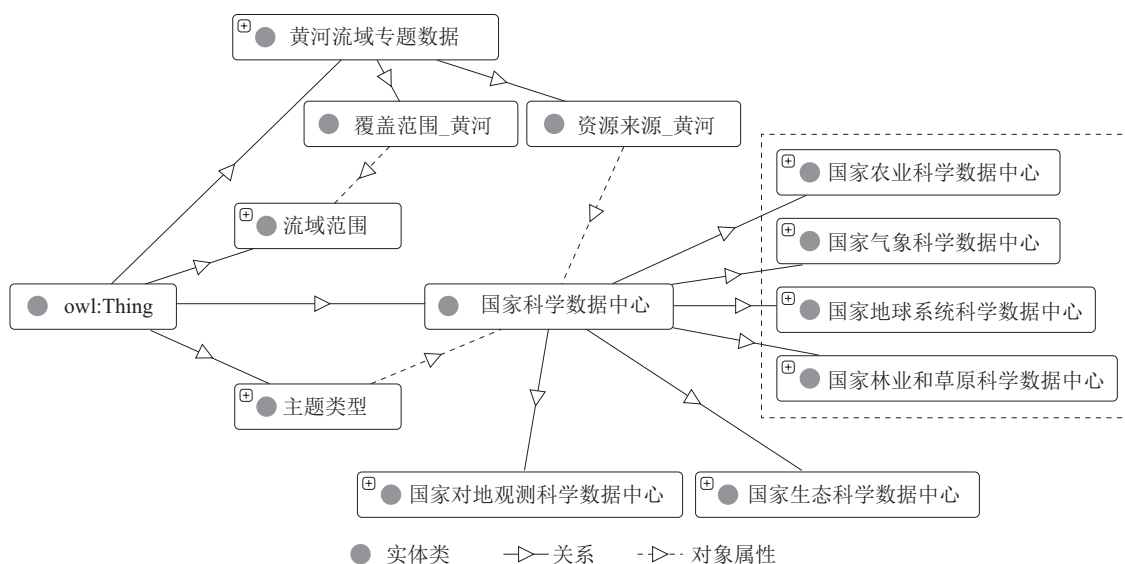


图1 6个国家科学数据中心关联模型

原先国家地球系统科学数据中心对黄河流域专题的数据支撑压力较大,各国家科学数据中心的专门元数据也较少。在增加2个国家科学数据中心后构建的关联模型中,黄河流域专题的数据来源与涉及的学科领域更加丰富,能够调动更多数据资源,扩充了时间和空间分辨率、仪器名称、数据投影信息等特色鲜明的专门元数据,可以有效减轻国家地球系统科学数据中心的数据压力。此外,国家对地观测科学数据中心汇集大量遥感卫星数据,支持遥感数据治理和服务;国家生态科学数据中心拥有实时同化预测系统、国家生态系统评估综合集成平台、中国陆地生态系统碳源汇模拟平台等一系列生态系统科学数据公共服务与分析平台,以及独有的全国野外站监测数字系统,能够展现更多实

时数据,有助于从不同视角完善黄河流域相关科学数据资源,更好地实现跨平台、跨学科的数据流通,提供更多数据应用与分析服务。

4 国家科学数据中心关联模型应用

黄河流域生态保护与治理作为推进我国生态文明建设的重要方面,事关沿黄人民的切身利益和我国经济社会发展全局。基于元数据的国家科学数据中心关联模型,充分汇聚整合多源数据资源,可实际应用于黄河流域的生态保护与修复,通过完善自然资源监测数据、提升流域内灾害防控与应急能力、助力生态修复工程实施、建立一体化生态环境智能监测系统4个方面,为

其生态治理与修复提供数据支持,从治理与防控两个角度综合解决黄河流域的生态问题,提高黄河流域治理的科学性和有效性。

4.1 完善自然资源监测数据

黄河是我国的第二长河,流经青海、四川、甘肃、宁夏、内蒙古、山西、陕西、河南、山东9个省(自治区),覆盖面积广阔,流域内资源丰富,若要进行流域生态保护和治理先要了解流域范围内自然资源的分布情况。利

用国家地球系统科学数据中心的资源科学数据(如水资源、农业资源、生物资源、基础地理数据、地形地貌数据等),国家林业和草原科学数据中心的湿地分布、土地资源、植被资源等数据集,叠加国家对地观测科学数据中心的全球生态环境遥感监测系统实时数据,进一步明确黄河流域滩区边界,摸清流域内林草、湿地、荒漠等自然资源空间分布情况,完善黄河流域自然资源三维立体“一张图”,精准确定林草等自然资源保护和监测范围边界(见图2)。

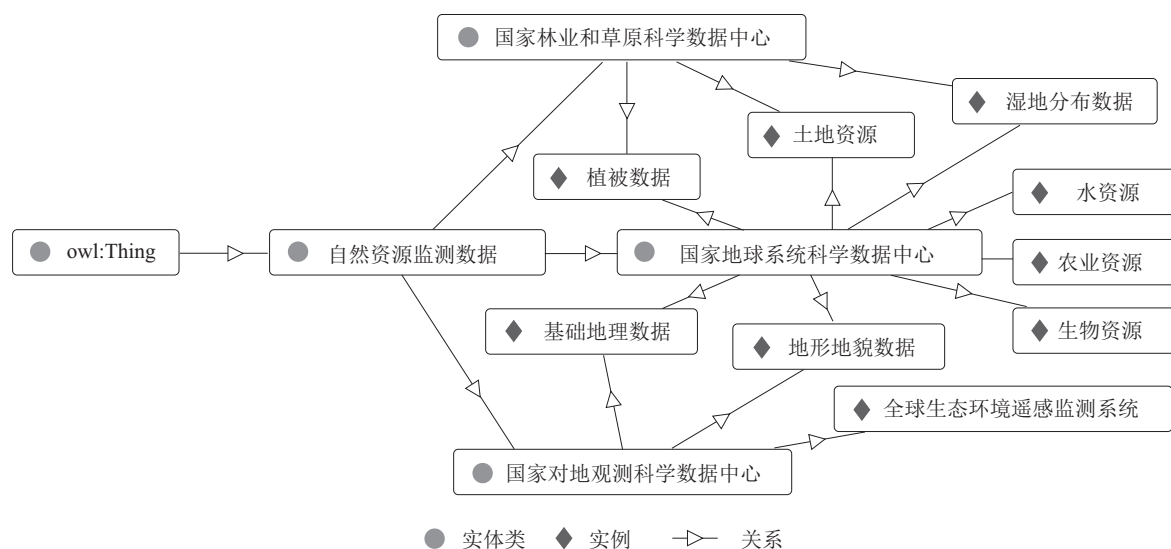


图2 自然资源监测数据关联可视化

以基础地理数据为例,完善该类数据可关联至国家地球系统科学数据中心,如黄河下游河道1:10万基础地理数据集,该数据集包含黄河下游区域的基础地理要素,包括铁路、公路和大车路等道路要素,河流、渠道、大堤、生产堤、工程险工、大断面等线状地物,水域和居民地等面状要素,省界与县市界,闸门、虹吸、提灌站、公里桩等水利建筑设施。国家对地观测科学数据中心也提供基础地理数据,如2020年黄河流域矢量范围数据,通过流域边界的提取,可以划定土壤侵蚀及非点源污染的迁移范围,可用于黄河流域土壤风蚀时空动态评价、生态环境质量评价等方面研究。此外,可针对林草、水文等重要资源持续开展专题监测,形成季度监测成果,及时掌握自然资源动态变化情况,完善黄河流域自然资源综合监测体系,为黄河流域的资源空间规划、生态治理和修复等工程提供基础数据支持。

4.2 提升流域内灾害防控与应急能力

黄河流域生态环境脆弱,灾害频发。做好黄河流域内洪水、森林火灾等灾害的防控应急措施工作,可以有效遏制流域生态的急剧恶化。充分调动国家地球系统科学数据中心的自然灾害数据、大气质量数据、山洪风险专题等数据,同时可依据灾害主题关联专题相关数据集,如国家林业和草原科学数据中心拥有各地森林火灾数据库、国家对地观测科学数据中心提供夏季中国洪涝灾害数据专题服务等。同时诸多数据中心提供“防灾减灾”数据服务,如国家农业科学数据中心的抗旱减灾保产实验数据、新品种抗旱试验数据库等,可为抗旱减灾工作提供数据支持。国家气象科学数据中心提供一系列气象专题,如暴雨、干旱、台风等气象灾害专题数据,可以了解黄河流域历年来多发的灾害类型、时间、地点、原因及影响等,以此为依据有针对性地制

定灾害防控与应急措施。国家气象科学数据中心拥有的气象预警专题,基于WebGIS,发布大风、雷电、寒潮、干旱、冰雹等近20种自然灾害,同时,预警展示中结合了实况预报数据,对预警相关气象要素的动态进行综合展示,便于有关部门提前布置灾害防控措施,有效降低灾害产生的恶劣影响,减少损失。国家对地观测科学数据中心具有全生命周期的遥感数据治理和服务技术体系,可充分利用其建设的对地观测科学数据共享资源库中环境减灾遥感卫星资源,提升灾害风险预警能力,加强灾害风险评估和隐患排查治理,提高灾害应急能力。

以预防洪涝灾害为例,国家地球系统科学数据中心设立了山洪风险专题、中国2020年夏季洪涝救灾专题,国家对地观测科学数据中心收集了2000—2020年黄河流域1千米降水数据、2021年河南特大暴雨灾害数据,同时提供2022年夏季中国洪涝灾害数据专题服务,国家气象科学数据中心集齐暴雨灾害专题数据、中国近百年降水专题数据集,可以此分析降水年际变化,确定洪涝多发地区与时间,再结合其中心可视化平台的降水实况资料和气象卫星遥感数据,精准开展高风险、高影响地区的综合监测和灾害预警,提升洪涝灾害预警的准确性以及时效性。

4.3 助力生态修复工程实施

随着国家生态文明建设的发展,一系列重大生态保护修复工程陆续实施。科学实施生态修复,有助于提高我国生态系统的质量与稳定性,筑牢国家生态安全。可结合国家生态科学数据中心建设的国家生态系统评估综合集成平台、生态水文数据、生态系统功能数据等,国家对地观测科学数据中心的土壤风蚀、水土流失、土地利用等数据,国家林业和草原科学数据中心的植被数据和退耕还林工程数据,排查重要水系阻隔、水土流失、湿地退化等问题,梳理当前黄河流域生态治理的现状。还可利用国家生态科学数据中心的全国野外站监测数字大屏展示系统来实时浏览各生态站监测活动、获取监测指标数据,及时发现污染水质、破坏岸线、侵占河道等行为。又能借鉴国家地球系统科学数据中心的环境退化及保护数据、环境污染及治理数据、黄河口湿地微地形修复实验数据等对生态修复有启示作用的数据集可分流域、分省市有针对性地制定生态保护和修复方案,并及时跟踪后续修复进程及效果。

4.4 建立一体化生态环境智能监测系统

生态环境保护的数字化转型需要综合管理数据资源,推动跨平台、跨学科、跨领域协同合作,全面提升生态资源服务数字化、智能化水平,充分释放数据要素价值。国家地球系统科学数据中心整合了卫星影像、雷达影像、地物波谱等遥感数据;国家气象科学数据中心的气象卫星遥感数据和气象预警专题可提供气象实况资料;国家对地观测科学数据中心具有全生命周期的遥感数据治理和服务技术体系、中国遥感卫星数据统一检索平台;国家生态科学数据中心拥有全国野外站监测数字大屏展示系统,同时建成先进的生态系统科学数据公共服务与分析平台,如实时同化预测系统、国家生态系统评估综合集成平台、中国陆地生态系统碳源汇模拟平台;国家农业科学数据中心建设有国家农业科学观测数据可视化平台;国家林业和草原科学数据中心提供空间数据展示服务、生态站定位观测数据服务系统。国家科学数据中心关联模型整合多源数据资源,支持构建一体化生态环境智能监测平台,打造生态环境综合管理数字平台,致力于相关数据资源的全局规划、智能检索、高效共享和协同服务建设,可实现数据监测一体化、数据服务一站式,提供生态系统科学数据共享、分析应用等服务,充分发挥我国长期生态监测数据价值。

5 总结与展望

科学数据作为国家战略科技资源,在支撑创新驱动发展战略实施,推动经济体系优化升级和促进产业融合,抗击疫情保障民生和推进生态文明建设等方面的作用更加凸显^[29]。国家科学数据中心作为我国科学数据管理与应用、开放与共享的重要载体,为相关学科领域发展、团体研究、项目实施等提供科学数据资源与服务。

黄河流域是我国重要的生态屏障和重要的经济地带,在我国经济社会发展和生态安全方面占据十分重要的地位。本文以黄河流域专题数据为基础,以元数据为切入点,探寻各国家科学数据中心间的关联关系,通过构建国家科学数据中心关联模型,为黄河流域生态保护与高质量发展提供数据支持。最后将理论研究应用到实践探索中,运用关联模型从完善自然资源监测数据、提升流域内灾害防控与应急能力、助力生态修复工

程实施、建立一体化生态环境智能监测系统4个方面解决黄河流域的生态保护与治理修复问题,有助于保障黄河长治久安、促进全流域高质量发展。同时,为科研人员高效利用国家科学数据共享平台提供理论范式和实践参考,着力解决当前各国家科学数据中心间数据资源存在的通用性低、互操作差等问题。

本文在国家科学数据中心关联模型构建研究过程中,以黄河流域专题数据为基础,只选取了6个与其密切相关的国家科学数据中心,通过专题已有的8个元数据进行关联模型构建,或许存在一定不足,但也有进一步提升的可能性。国家科学数据中心关联模型研究未来可以从以下几个方面进一步深入。第一,国家科技基础条件平台建设有20个国家科学数据中心,数据资源丰富,涉及学科领域广泛,可以继续关联更多数据中心,如国家冰川冻土沙漠科学数据中心、国家微生物科学数据中心等。除此以外,一些科研机构、高校建立的数据中心或数据库也可纳入统一考虑的范畴,进一步拓展数据资源。第二,还要扩展更多的关联元数据,丰富元数据种类,编制核心元数据目录,区分通用核心元数据和专题扩展元数据,规范数据中心元数据标准,提高关联模型的适用性和扩展性,更好地解决科学数据中心间数据资源存在的通用性低、互操作差的问题。第三,国家科学数据中心关联模型,需要创新数据管理机制、深化数据高效共享、促进数据有序开发利用,不断完善开放共享的数据资源体系,真正实现跨机构、跨学科、跨领域的互通共享,充分挖掘科学数据的潜在价值,推动科技创新发展。

参考文献

- [1] 黄鼎成, 郭增艳. 科学数据共享管理研究 [M]. 北京: 中国科学技术出版社, 2002.
- [2] 邢文明, 肖嘉丽, 陈继丽. 面向FAIR原则的《科学数据管理办法》优化研究 [J]. 图书馆论坛, 2022, 42 (4): 86-95.
- [3] 诸云强, 潘鹏, 石蕾, 等. 科学大数据集成共享进展及面临的挑战 [J]. 中国科技资源导刊, 2017, 49 (5): 2-11.
- [4] 中共中央关于制定十三五年规划的建议 (全文) [EB/OL]. [2022-07-17]. <http://politics.people.com.cn/n/2015/1103/c1001-27772701.html>.
- [5] 国务院办公厅印发《科学数据管理办法》[EB/OL]. [2022-07-17]. http://www.gov.cn/xinwen/2018-04/02/content_5279295.htm.
- [6] ARDESTANI S B, HAKANSSON C J, LAURE E, et al. B2SHARE: An Open eScience Data Sharing Platform [C] // 2015 IEEE 11th International Conference on e-Science. 2015: 448-453.
- [7] QIN J, BALL A, GREENBERG J. Functional and Architectural Requirements for Metadata: Supporting Discovery and Management of Scientific Data [C] // International Conference on Dublin Core and Metadata Applications. 2012: 62-71.
- [8] 刘峰, 张晓林. 科学数据元数据标准述评及其通用化设计研究 [J]. 现代图书情报技术, 2015 (12): 3-12.
- [9] 司莉, 赵洁. 美国开放政府数据元数据标准及启示 [J]. 图书情报工作, 2018, 62 (3): 86-93.
- [10] SI L, XING W W, ZHUANG X Z, et al. Investigation and analysis of research data service in university libraries [J]. The Electronic Library, 2015, 33 (3): 417-449.
- [11] 黄文碧. 基于元数据关联的馆藏资源聚合研究 [J]. 情报理论与实践, 2015, 38 (4): 74-79.
- [12] 贾欢. 科学数据元数据互操作研究 [D]. 武汉: 武汉大学, 2017.
- [13] KAUPPINEN T. Why to Manage and Share Research Data? [EB/OL]. [2022-07-06]. <http://linkedscience.org>.
- [14] 游毅, 成全. 基于关联数据的科研数据资源共享 [J]. 情报杂志, 2012, 31 (10): 146-151.
- [15] 庄倩, 何琳. 科学数据共享中科研人员共享行为的演化博弈分析 [J]. 情报杂志, 2015 (8): 152-157, 168.
- [16] 陆颖, 张雅婷, 胡佳琪. 基于文献数据与科学数据融合的学科服务用户分析与发展建议 [J]. 图书馆学研究, 2021 (7): 75-82.
- [17] 张晓霞. 基于情感和异源异构数据融合的潜在关系发现模型研究 [D]. 厦门: 厦门大学, 2014.
- [18] ARIAS J J, PHAM-KANTER G, CAMPBELL E G. The growth and gaps of genetic data sharing policies in the United States [J]. Journal of Law and the Biosciences, 2014, 2 (1): 56-68.
- [19] PLANTIN J C, LAGOZE C, EDWARDS P N. Re-integrating scholarly infrastructure: the ambiguous role of data sharing platforms [J/OL]. Big Data & Society, 2018, 5 (1): 14 [2022-07-01]. <https://doi.org/10.1177/2053951718756683>.
- [20] 黄铭瑞, 李国庆, 李静, 等. 国家科学数据中心管理模式的国际对比研究 [J]. 农业大数据学报, 2019, 1 (4): 14-29.
- [21] 卢逸航, 李国庆, 陈祖刚. 科学数据中心间互操作模式研究 [J]. 数据与计算发展前沿, 2022, 4 (1): 69-83.
- [22] CURTIS S K, SINGH J, MONT O, et al. Systematic framework to assess social impacts of sharing platforms: Synthesising literature and stakeholder perspectives to arrive at

- a framework and practice-oriented tool [J]. PloS One, 2020, 15 (10): e0240373.
- [23] 司莉, 王雨娃. 我国科学数据共享平台数据组织的现状及改进建议——基于国家科技基础条件平台的分析 [J]. 图书馆建设, 2018 (10): 52-58.
- [24] 刘小宇, 纪雪梅, 王芳. 我国科学数据共享平台的被引用状况调查与分析——以国家科技基础条件平台为例 [J]. 情报杂志, 2020, 39 (9): 168-175.
- [25] THOMAS J, NAVRE M, RUBIO A, et al. Shared platform for antibiotic research and knowledge: a collaborative tool to SPARK antibiotic discovery [J]. Acs Infectious Diseases, 2018, 4 (11): 1536-1539.
- [26] 李晓萍. 农业科学数据共享平台用户持续使用的影响因素研究 [D]. 北京: 北京邮电大学, 2018.
- [27] 侯超菲. 协同治理视角下黄河流域生态保护与治理的问题及对策研究 [D]. 太原: 山西大学, 2021.
- [28] 国家地球系统科学数据中心. “黄河流域生态保护和高质量发展”跨平台专题 [EB/OL]. [2022-07-15]. <http://www.geodata.cn/apps/yriver/index.html>.
- [29] 苏靖. 国家科学数据中心联合专刊总序 [J]. 中国科技资源导刊, 2022, 54 (1): 1.

作者简介

孙叶华, 女, 1999年生, 硕士研究生, 研究方向: 科学数据管理。

刘桂锋, 男, 1980年生, 博士, 研究馆员, 通信作者, 研究方向: 科研数据管理、大数据分析、专利分析, E-mail: liuguifeng29@163.com。

陈帅印, 女, 1998年生, 硕士研究生, 研究方向: 科学数据管理。

Research on the Construction of Association Model of National Science Data Center Based on the Thematic Data of Yellow River Basin

SUN YeHua¹ LIU GuiFeng^{1,2} CHEN ShuaiYin¹

(1. Institute of Science and Technology Information, Jiangsu University, Zhenjiang 212013, P. R. China;

2. Jiangsu University Library, Zhenjiang 212013, P. R. China)

Abstract: The National Science Data Center is an important carrier of scientific data management and an important infrastructure for scientific data collection, management and open sharing. Based on the thematic data of Yellow River Basin. Based on the thematic data of the Yellow River Basin, this paper designs the association metadata, studies the association relationship of the scientific data set of the National Science Data Center, and builds the association model of the National Science Data Center. This paper attempts to promote the ecological protection and restoration of the Yellow River Basin from four aspects: improving the monitoring data of natural resources, improving the disaster prevention and emergency response capabilities in the basin, assisting the implementation of ecological restoration projects, and establishing an integrated ecological environment intelligent monitoring system. The research on the construction of the association model is helpful to realize the interoperability of scientific data in different fields, enhance the data circulation, promote the sharing of scientific data, and promote the development of scientific and technological innovation.

Keywords: Scientific Data; Metadata; Yellow River Basin; Association Model; Data Management

(收稿日期: 2022-07-19)