地球科学数据仓储通用框架设计

蒋 涵[1,2] 王卷乐[2,3] 袁月蕾[2]
（1. 江苏海洋大学海洋技术与测绘学院，江苏连云港 222005；
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；
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）

摘要：随着“大数据”理念的普及、数据驱动科学研究“第四范式”的兴起，海量数据资源的有序存储和规范化管理对于促进科学数据共享和驱动科学发现将发挥越来越重要的作用。目前我国科学数据仓储刚刚起步，在地球科学领域还与国际有很大差距，迫切需要研究地球科学数据仓储的通用构建框架。本文首先概述Figshare等国内外数据仓储的进展，然后结合地球科学数据特点，分析地球科学领域数据仓储的结构和需求，设计出数据仓储构建的流程和框架，最后以CC协议为例探讨数据作者权利，构建合理灵活的版权制度，解决著作权限制规则过度的问题，为实现地球科学数据有序、合理、永久的存储、使用和共享提供框架参考。

关键词：地球科学；数据仓储；框架设计；数据中心；CC协议
DOI：10.3772/j.issn.1674-1544.2022.02.002 **CSTR：**15994.14.issn.1674.1544.2022.02.002
中图分类号：TP302.1；G321；G203 **文献标识码：**A

General Framework Design for Geoscience Data Repository

JIANG Han[1,2], WANG Juanle[2,3], YUAN Yuelei[2]
(1.School of Marine Technology and Geomatics, Jiangsu Ocean University, Lianyungang 222005; 2.Institute of Geographic Science and Natural Resources Research, Chinese Academy of Sciences, State Key Laboratory of Resources and Environmental Information Systems, Beijing 100101; 3.Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023)

Abstract: With the popularization of the concept of “big data”and the rise of the“fourth paradigm” of data-driven scientific research, the orderly storage and standardized management of massive data resources will play an increasingly important role in promoting scientific data sharing and driving scientific discovery. At present, scientific data repository has just started in China, and there is still a big gap with the world in the field of Geoscience. It is urgent to study the general construction framework of geoscience data repository. This paper summarizes the progress of data repository at home and abroad, such as Figshare, combined with the characteristics of geoscience data, analyzes the structure and requirements of data warehousing in the field of Geoscience, designs the process and framework of data repository, takes CC licenses as an example to discuss the rights of data authors, constructs a reasonable and flexible copyright system, and solves the problem of excessive copyright restriction rules. This study is expected to provide a framework reference for the orderly,

作者简介：蒋涵（1996—），男，江苏海洋大学硕士生，研究方向为地理信息共享；**王卷乐**（1976—），男，中国科学院地理科学与资源研究所研究员，博士生导师，研究方向为科学数据集成与共享（通信作者）；**袁月蕾**（1987—），女，中国科学院地理科学与资源研究所工程师，研究方向为地球科学数据共享。

基金项目：中国科学院战略先导专项（A类）子课题“面向‘全景美丽中国’的地球大数据综合集成”（XDA19040501）。
收稿时间：2021 年 12 月 1 日。

reasonable and permanent storage, use and sharing of geoscience data.
Keywords: geoscience, data repository, frame design, data center, CC licenses

0　引言

随着数据密集型科学研究迅速发展，科学研究第四范式的到来，大数据已经成为新时代战略型数字化资源。第四范式是继第一范式的经验科学、第二范式的理论科学、第三范式的计算科学之后，以数据密集型科学研究为特征的新的研究范式。科学数据已成为驱动创新发现的重要因素，也是科学研究数字化基础设施的核心内容[1]。学术界一直非常重视科研数据对于科学研究的支撑作用[2]。2016年3月提出的FAIR（可发现、可访问、可互操作和可重用）原则被称为“科学数据管理的指导原则”[3]，其对科学数据的有序管理与共享具有广泛的科学意义与价值。

海量地球科学数据的汇聚和共享离不开数据仓储。数据仓储（Date Repository），即数据存储库，是为研究人员、学术期刊、机构提供数据存储、数据管理、数据保存、数据共享和数据出版及数据在线获取服务的基础设施。确保数据的真实、可靠、完整和可用是科学数据仓储的核心目标。国际上将数据仓储具备的这4种特性称之为“TRUST”（可信任数据仓储）[4]。数据仓储不仅有利于用户便捷准确地获取免费的学术信息，而且可以长期保存、有效管理学术成果，向他人展示自己科学研究内容，加速学术传播与交流，提升科研成果的引用率，促进学术的良性竞争等。数据仓储的研究重点不仅仅是信息系统建设，还涉及数据政策、数据标准、数据的权益保护等问题。数据仓储在数据全生命周期管理中扮演着关键角色，为研究成果提供了一个稳定的平台。清楚地了解如何处理数据仓储存储的数据，可以促进数据的有效治理，从而使数据仓储管理人员、数据作者、数据用户以及更广泛的学术团体能够从数据仓储中获得最大的收益[5-7]。

值得说明的是，科学数据仓储与用于数据存储和计算的云平台是有区别的。云概念是基于“云计算”技术，实现各种终端设备之间的相互联通。云存储技术是基于传统媒体系统发展的一种全新的信息存储管理方式。这种方式将计算机系统的软硬件优势进行整合应用，即将计算、存储、网络资源封装成服务的形式提供给用户，用户以自己所需的方式通过互联网获取所需的服务[8]。而数据仓储则是一个与期刊论文数据关联的共享应用系统，它既可以借助于现有的云存储平台，也可以独立建立自身的存储系统。

然而，如何科学有效地建立本领域的科学数据仓储？本文首先调研对比国内外已建的典型数据仓储在功能结构、服务内容、服务形式等方面的情况，再以地球科学数据仓储为实例，结合地球科学数据的特点，对地球科学数据仓储的架构和功能进行总体设计，最后提出地球科学数据仓储通用框架设计，以期对地球我国科学数据仓储的建设起到参考和借鉴的作用。

1　国内外数据仓储的发展情况

本节以国内外通用型科学仓储和地球科学领域的学科型仓储在使用流程、服务内容和特点等方面进行概述，以期找到数据仓储构建的服务切入点，探索地球科学领域的数据服务模式。

1.1　Figshare

Figshare是英国Digital Science公司旗下的一个数据仓储，为科研人员提供发布各类研究成果的平台，使研究成果可以更好地被引用、共享和发现。Figshare允许用户上传包括文本、图片、多媒体等多种数据类型的数据，使各种形式的研究成果都能被更好地存储、使用和分享。它为所有数据分配了数字对象唯一标识DOI（Digital Object Identifier），以便于数据的引用。该平台对所有用户开放，并可以对上传数据进行统计，了解每个文件被浏览、共享和下载的次数。Figshare是一个基于云计算技术的数据仓储，采用Creative Commons（CC）许可协议共享数据，

以云平台和云数据为基础，来保证数据存储的可靠性和安全性，同时也满足了出版商的出版结构[9-11]。

Figshare平台涉及地学、社会学、生物学、化学、工程学等近30个学科的数据。用户首先需要注册一个账户，然后登录到系统中，根据自己的需求上传存储自己的研究成果。在用户空间内可对数据进行编辑，并根据自身需要设置访问权限，选择可公开或者可不公开方式。如果选择公开方式则可以在此平台上发表出版，如果选择不公开方式则数据仅为用户本人可见，但用户在后期仍可以根据自身情况决定是否公开出版[12-13]。

Figshare平台为科研人员、科研机构和出版商三类用户提供多种服务。

（1）为个人提供服务。个人是Figshare平台的主要服务对象，为其提供了许多免费的服务。平台接收上传最大5 GB的文件，给个人提供20 GB的免费私人空间。不仅可以上传包括图片、表格等任何文件格式的数据，而且可以与同事一起建立私人文件夹并进行访问。数据上传后可以拥有DOI，用户个人选择是否公开数据。

（2）为机构提供服务。Figshare平台为机构提供科研数据管理、科研数据传播和可定制的门户展示服务。Figshare还为机构提供统计和报告的服务，管理人员可以查看机构人员或者数据下载量、引用量等统计数据。

（3）为出版商提供服务。Figshare与PLOS One、Wiley等出版商合作，涉及学科类型众多。期刊文献的每一个附带数据都有一个唯一的DOI与之匹配。平台会创建一个包含数字媒体和数据文件的库，促进数据的开放使用从而增加文章的流量。

1.2　ScienceDB

Science Data Bank（ScienceDB）由中国科学院计算机网络信息中心建设维护，致力于打造中国的数据长期共享和数据发布资源库。ScienceDB是一个公共的通用型科学数据仓储，主要面向科研人员、科研项目/团队、科研期刊、科研机构及高校等利益相关者，提供科学数据汇交、保存、出版、共享和获取等服务，支持多种数据获取与使用许可，在充分保障数据拥有者权益的基础上，促进数据的共享与使用。ScienceDB确保出版数据的持续访问、长期管理，并面向国际学术界、学术期刊和出版商以及其他利益相关者提供配套的数据发布和获取服务。ScienceDB致力于出版数据符合主流数据标准或惯例的科学数据，旨在促进科学数据的可发现性、可访问性、互操作性和可重用性（FAIR原则），并推动数据共享文化氛围在中国的培育及良性发展[14]。在ScienceDB上发表一个数据集需要4个步骤，即注册与登录、数据提交、数据评审和数据发布。ScienceDB的服务具有以下特性。

（1）开放与共享性。①永久可访问：保障上传数据与出版资源的永久可访问。②出版数据可发现：ScienceDB对所有发布数据资源进行了搜索引擎发现优化。③开放共享：在尊重数据作者知识产权的前提下，ScienceDB上的数据提倡并支持开放共享，并推荐选择CC0协议出版用户数据成果。④OPEN API：出版数据集及ScienceDB的公开服务提供OPEN API为程序或第三方服务使用。⑤数据资源可获取：用户可在线获取在ScienceDB上发布的数据集元数据和数据文件。

（2）数据管理方式和模式可信任性。①通用兼容：ScienceDB不限制数据所涉及的学科领域，并支持所有格式的数据文件上传。②数据资源可引用：提供标准化的元数据采集过程和自动化的数据资源唯一标识注册，推荐每个发表数据资源的引用格式，确保发表的数据成果可规范引用。③出版资源唯一标识：配套唯一标识自动注册及管理服务，确保所有出版数据集可唯一认证标识，且支持DOI和CSTR（Chinese Science and Technology Resource）标识体系注册及认证。④数据更新可追溯：ScienceDB跟踪数据资源的每一次更新，记录、发布并标识注册发布数据集的历史版本信息。⑤数据可管理：提供数据质量审核、访问权限控制、数据版本控制等管理服务。⑥资源互联与国际化推广：提供自动化的中

英双语服务，助力数据成果传播；与第三方数据服务资源对接，达成数据资源的全网互联与数据流通。

1.3　PANGAEA

PANGAEA数据仓储是由德国阿尔弗雷德韦格纳研究所、赫尔姆霍兹极地和海洋研究中心及不来梅大学海洋环境科学中心主办的对任何组织和个人开放并保证长期运行的地球科学数据库，旨在归档、发布和分发地球系统研究的相关数据。它是国际科学理事会世界数据系统（WDS）的成员，是经CoreTrustSeal认证的公共仓储。PANGAEA发展起步早，在数据标准和政策、运行方式和模式等方面经验丰富[15−16]。

（1）PANGAEA中的数据具有良好的可发现性。大多数数据是免费提供的，可以根据数据集描述中提到的许可条款进行使用，可以使用数字对象标识符DOI来识别、共享、发布和引用该数据库中的数据集。

（2）PANGAEA允许将数据作为论文的附件进行发布，或者与Scientific Data、ESSD、Geoscience Data Journal以及其他数据期刊相结合进行数据集的发布。数据领域涵盖地球化学、海洋、岩石圈、生物分类、大气、古生物、生态学、生物圈、地表环境、地球物理、冰冻圈、湖泊与河流、人类活动等专题，并支持根据作者、发布时间、具体项目、测试方法、地理位置等参数对数据进行筛选。

（3）PANGAEA有一个完善的互操作性框架，从而能够向数据注册、数据门户和其他服务提供者传播元数据和数据。PANGAEA提供了广泛的Web服务（SOAP/REST），包括用于元数据获取的OAI−PMH，对于选定的应用程序还可提供数据仓库Web服务，API允许检索任何一组数字和文本数据。所有PANGAEA数据集也遵从Schema.org/DataSet元数据，以便于数据的管理、更新和使用。

（4）PANGAEA的数据政策有明确的表述。数据库内容定义为地球系统研究数据。数据可以在时间和空间上进行地理参照。可以设置数据保护期对数据进行保护。数据由来源项目或机构负责。数据（元数据）的格式和描述必须确保其最广泛和最容易使用。此外，在使用来自PANGAEA的数据时，需要用户正确引用这些数据。

1.4　Dryad

Dryad由美国国家进化分析中心等机构在美国自然科学基金会的资助下建立，其最初由进化生物学和生态学的主要期刊和科学团体提出，鼓励与数据一同提交手稿，进行存储。目前已有451种期刊与之合作进行存储数据。Dryad具有以下特性。

（1）Dryad接收大多数类型的提交文件。如文本、电子表格、视频、照片、代码等，也接收多个文件的压缩文件。

（2）每个通过Web接口上传的数据发布都有300 GB的限制。Dyrad可以接受更大数据的提交，但提交者需要先与管理员联系。建议单个文件不应超过10 GB，这样可以确保Dryad用户轻松访问和下载文件。

（3）Dryad中的所有数据都遵循知识共享协议CC0。CC0（又名CC Zero）专门用于减少对数据重用的法律和技术障碍，但是CC0并不能免除研究人员在某种程度上重复使用这些数据的局限性，以及要求引用原始数据作者的权利。CC0有助于发现、重用和引用该数据。

（4）提交数据会获得一个唯一的DOI，并以https://doi.org/10.5061/dryad.XXXX.this格式进行保存，这样就可以方便数据的查找与引用[17−20]。

1.5　数据仓储总结与对比

表1是对以上几类典型数据仓储情况的对比结果。

近年来，除了ScienceDB，国内科学数据仓储呈现快速发展态势。2019年，国家地球系统科学数据中心的WDS可再生资源与环境数据中心入选美国地球物理学会（AGU）发布的“领域−学科仓储库推荐名单”。2020年，国家青藏高原科学数据中心通过《自然》（Nature）数据期刊Scientific Data认证，成为Nature及其子刊文章投

表 1　数据仓储对比

科学数据仓储	支持的机构	国家	永久标识符	支持的学科	数据许可协议	数据仓储目的	数据规模
Figshare	Digital Science公司	英国	DOI	多学科	CC BY 4.0、CCO、MIT、GPL、GPL 2.0+、GPL3.O+、Apache 2.0	为科研人员提供发布各类研究成果的平台，以便研究的产出可以更好地被引用、共享和发现	截至2021年10月，Figshare数据仓储存储的数据为55 665 036个
ScienceDB	中国科学院计算机网络信息中心	中国	DOI、CSTR	多学科	CC0、CC BY 4.0	提供科学数据汇交、长期保存、出版、共享和获取等服务，支持多种的数据获取与使用许可，在保障数据所有人权益的基础上，促进数据的可发现、可引用、可重用	截至2021年10月，ScienceDB存储数据集756 320个，数据体量为111 705+ GB，数据集访问量为410 305次，数据集下载量为13 583 062次
PANGAEA	德国阿尔弗雷德韦格纳研究所、赫尔姆霍兹极地和海洋研究中心、不来梅大学海洋环境科学中心	德国	DOI	地球科学	CC0、CC BY 4.0、CC−BY−SA	对任何组织和个人开放并保证长期运行的地球科学数据库，旨在归档、发布和分发地球系统研究的相关数据	截至2021年8月，PANGAEA收录有来自595个研究计划、405 097个数据集、超过190亿条数据
Dryad	美国国家进化分析中心等机构	美国	DOI	多学科	CC0	促进科研数据公开获得，与学术期刊相结合使科研成果得以重用	截至2021年10月，Dryad共收集数据43 080个

稿时可选的数据仓储中心。2021年，国家空间科学数据中心（NSSDC）成为AGU旗下期刊当年推荐的21个仓储库之一。

对上述国内外数据仓储的梳理发现，当前国内外数据仓储在功能、服务等方面有以下共同的特点：一是数据都拥有可唯一识别的DOI，使数据易于发现和引用；二是数据都遵循一定的协议许可，使数据权益问题得到保障等；三是用户可以根据自己的需求决定数据是否出版发布。这些功能对于地球科学数据仓储的框架设计具有非常重要的借鉴意义，但还不能直接满足地球科学数据仓储的需求。如有的地球科学数据年代久远，数据的时间范围难以直观掌握；数据的经纬度信息不明确，不便于在一定空间范围内数据的获取。因此，在借鉴以上国内外成果数据仓储经验基础上，本文提出地球科学数据仓储的总体设计框架。

2　地球科学数据仓储的总体设计

2.1　需求和背景

地球科学涵盖学科种类繁多，领域广泛，涉及包括空间科学、地质学、海洋科学等领域在内的多学科及其交叉融合。地球科学数据指的是在地球科学研究过程中，研究人员通过实地勘测、空间探测、实验测试、计算机模拟等手段采集的各类地球科学数据的总称。地球科学数据形式各异、种类繁多、来源广泛，具有大数据的体积大（volume）、速度快（velocity）、模态多（variety）、真伪难辨（veracity）和价值高密度低（value）的5V特性，还具有高时空性、高可视化、高相关性和高（多）维度的“四高”特征[21]。以固体地球科学为例，如表2所示。

随着大数据的快速发展，地球科学数据在研究陆地、海洋和大气等圈层在内的地球形成、生命演化、地球物质组成等重大课题中越来越重要[22]。地球科学数据大多是反映一定时间和空间范围内物质状态和物质性质的数据，因此地球科学数据具有高时空性的特点。这正是地球科学数据仓储构建中的关键特点。

2.2　总体架构

地球科学数据仓储的基本使用流程如图1所

示。首先，用户需要注册账户登录系统，然后根据需求填写数据的注册信息，包括是否有DOI、是否需要上传数据等，再提交给管理员进行数据注册信息的审核。如果需要上传数据，数据可设

表 2　固体地球学科分类

序号	中文名称	英文名称	序号	中文名称	英文名称
1	地层学	Stratigraphy	15	矿物	mineral
2	古生物学	Paleontology	16	地球物理学	geophysics
3	地质年代表	Geochronology	17	地球化学	geochemistry
4	沉积学	Sedimentology	18	遥感技术	remote sensing
5	火成岩	Igneous rocks	19	地貌学	Geomorphology
6	变质岩	metamorphic rock	20	古地磁学	paleomagnetic
7	构造	tectonics	21	钻探和钻孔	drill and borehole
8	地质绘图	geologic mapping	22	地球资源	Earth resources
9	地质图编制	geological map compilation	23	地热	geothermal
10	水文地质学	hydrogeology	24	地质勘探和管理	geo−exploration and administration
11	地质灾害	geohazard	25	古气候学	paleoclimatology
12	石油地质学	petroleum geology	26	古地理学	paleogeography
13	数学地球科学	mathematical geoscience	27	地质文献	geological literature
14	海洋地质学和海岸区	marine geology and coast zone			

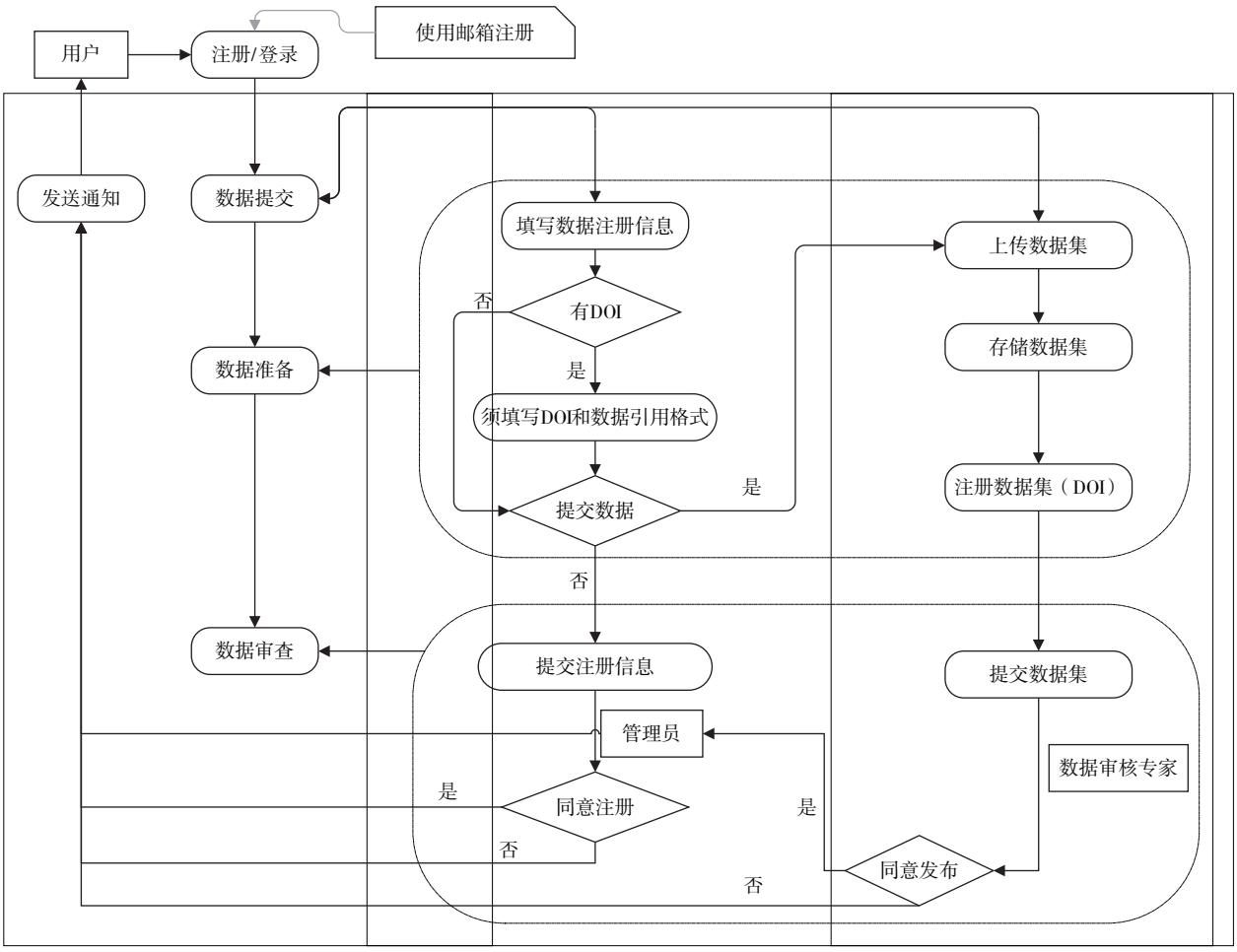

图 1　地球科学数据仓储总体流程

置为仅对本人可见，不进行公开发布，也可以选择发布数据，但对数据有一定的保护期限制，限制期过后自动出版。数据上传完成后，对于需要出版的数据提交数据注册，系统会生成一个唯一的DOI号，然后由系统根据关键词自动分配或管理员指定数据审核专家对数据进行审核，直至数据注册信息和数据集审核通过后，注册成功，数据才可以公开出版。

2.3　功能设计

地球科学数据仓储是一个针对地球科学数据的集注册、存储、管理以及应用于一体的系统。在计算机架构的基础上，结合地球科学数据特点，形成包括数据注册分系统、仓储前端应用分系统、数据管理分系统和用户管理分系统在内的4个分系统。

2.3.1　数据注册分系统

（1）用户中心。用户中心主要包括用户的个人账户注册登录以及个人信息维护。①注册：注册信息支持输入邮箱获取验证码、设置密码、确认密码进行个人账户的注册。②登录：支持输入用户名或者邮箱、密码、验证码进行登录认证；支持退出登录。③找回密码：用户如果忘记了密码，可以通过邮箱或找回密码。④用户信息：用户可以查看自己的基本信息，内容包括用户名、用户真实姓名、邮箱、研究领域、所属国家等信息；支持修改用户个人信息。

（2）数据注册。数据注册主要是填写相关元数据信息，将数据作者及对数据自身的描述信息进行登记。数据注册主要包括三大部分：数据集标识信息、数据时空信息、数据的权利责任信息。数据集标识信息是要求用户将数据集的名称、作者、关键词、摘要、学科分类等信息的填写。数据时空信息是对数据集的时间分辨率、高度和深度、数据集的坐标信息等时空信息的描述，以便于对数据的时间空间信息进行准确定位。数据的权利责任信息是责任方姓名、国家、电子邮件等信息。

（3）数据存储。安全可靠的存储环境是科学数据仓储稳定持续为用户服务的前提。数据生产者针对自己所要提交的数据，填写科学数据的描述信息和分类等注册信息，以便后期使该数据的检索、使用、发布、引用以及分析提供支撑。数据注册成功后，如果需要对数据进行发布，则需要上传数据。数据上传成功后，会自动生成内部标识符和DOI。对于首次发表的数据，还需要选择数据的许可协议，以便明确数据的权利责任。

2.3.2　仓储前端应用分系统

（1）个人数据中心。个人数据中心是一个包括我的数据、项目、收藏夹、购物车等对数据进行管理的模块。①我的数据：支持数据的列表展示，包括数据名、注册信息审核状态、数据集审核状态、状态更新日期、数据量大小、操作；支持将数据移入到项目数据中；支持将数据移入到收藏夹中；支持未提交注册的数据保存草稿，且支持编辑修改。②项目：支持项目的管理，包括创建、查看、编辑、删除；支持在项目中移出、移入数据集；支持项目成员管理（创建项目的人可管理项目成员，普通用户不可以）；支持项目中所有成员共享项目。③收藏夹：支持收藏夹的管理，包括创建、查看、编辑、删除；支持在收藏夹中移入移出平台数据集和个人数据集；支持按数据量、添加日期、修改日期进行排序；支持按数据集状态查看；支持按关键字进行搜索。④购物车：支持购物车的管理，包括查看、删除；支持购物车中的数据移入收藏夹，并可进行下载；支持按数据量、添加日期、修改日期进行排序；支持按数据集状态查看；支持按关键字进行搜索。

（2）数据检索。数据检索是对数据的查询与检索：支持全文检索；支持筛选条件检索，包括关键词、作者、DOI号等；支持分类筛选检索，包括地质时代（如以10−50 Ma为形式的数字时间）、空间范围（输入四角经纬度）、学科类型、发布时间；检索结果支持列表及缩略图展示；地图检索（二维世界地图）支持输入四角坐标和地图框选进行数据搜索。

2.3.3　数据管理分系统

（1）元数据管理。支持多种类型字段的添

加、修改、删除，支持对学科类别的管理，支持对关键词管理，支持对协议许可的管理，支持数据格式管理等。

（2）数据内容管理。数据状态包括未审核、审核中、无需审核、审核不通过、审核通过、已发布；专家系统审核包括数据提交学科数据专家审查，支持随机推送专家、指定专家，专家填写意见；管理员审核包括支持管理员对用户提交的注册信息进行审核，支持管理员对用户提交的数据内容进行审核、回访跟踪、回退、发布等操作。

2.3.4　用户管理分系统

用户管理分系统包括普通用户、管理员、数据审核专家三类用户。普通用户支持搜索下载数据集。管理员可对普通用户和数据审核专家进行管理和授权，支持显示用户列表；支持根据用户名称、邮箱进行查找，按照用户状态进行过滤；支持按照用户最近登录时间进行排序；支持查看普通用户和数据审核专家的角色全选；支持增加、修改、删除数据审核专家角色的功能。数据审核专家由管理员进行管理，对所接收到的数据进行审核。

3　数据仓储中的权益保护策略

科学数据是数字化内容的重要组成部分，为了保护数据作者的合法权益，且使数据能够得到充分的共享，Creative Commons能够为构建灵活合理的版权制度，解决著作权限制规则过度提供帮助[23]。CC协议是一个全球性的非营利组织，它通过提供免费的法律工具来实现创造力和知识的共享和再利用。知识共享协议是一个相对宽松的版权协议。它是通过作者对4种权利（署名、非商业用途、禁止演绎、相同方式共享）的选择和组合，同时让使用者可以明确知道所有者的权利，从而达到不容易侵犯对方版权以及作品可以得到有效传播的目的[24]。

知识共享协议实质上是一系列许可协议的总称，其核心理念是让协议许可人（通常为作品的著作权人）自愿保留一定的权利而放弃一些权利。CC0是一个完全开放给公众的许可（public dedication tool），它允许创作者放弃他们的版权并将他们的作品放入全球公共领域。CC0无条件地允许重用者以任何媒介或格式分发（distribute）、混合（remix）、改编（adapt）和重建（build upon）相应内容。该理念在实践中则通过四大授权要素来实现，即署名（Attribution by）、非商业性使用（Non-commercial）、禁止演绎（No Derivative Works）、相同方式共享（Share Alike），见表3。在具体的使用过程中，著作权人可以根据自身实际需求对这4种要素进行组合，即署名（CC BY）、署名—非商业性使用（CC BY-NC）、署名—相同方式共享（CC BY-SA）、署名—禁止演绎（CC BY-ND）、署名—非商业性使用—相同方式共享（CC BY-NC-SA）与署名—非商业性使用—禁止演绎（CC BY-NC-ND），最终得到一个相对订制化的知识共享许可协议[25]。

CC协议在仓储系统对于数据共享所起到的作用是提交者必须做出声明并保证。其本人是内容的创建者和拥有者，或拥有足够的权利使内容被公开。在CC协议下，他人可以在任何媒介以任何形式复制、发布本协议下的数据集，也可以在任何用途下，甚至出于商业目的对数据进行修改、转换或以本数据集为基础进行创作。这对于数据的共享使用是极其重要的。

表3　CC协议的授权方式

授权方式中文名称	英文名称	简称	主要内容
署名	Attribution By	BY	使用者须保留原作者的署名
非商业使用	Non-commercial	NC	限于非商业性（即非营利性）目的
禁止演绎	No Derivative Works	NC	不得进行演绎创作(即不得对原作进行修改和二次创作)
相同方式共享	Share Alike	SA	允许进行二次创作，但须使用相同的CC许可协议

4　结语

在大数据时代，数据成为科学研究的重要资本，科学数据的价值越来越大。针对国内数据仓储建设的紧迫需求，本文总结了国内外数据仓储的建设经验，以及它们在开放共享方面的功能结构和特性特点，发现目前全球科学数据仓储建设发展迅速，数据的存储、管理、维护、共享、引用等遵从全生命周期的数据管理要求，为数据分配DOI可以使数据得以快速检索和复用，使用CC协议等可以保护著作人的合法权益。结合地球科学数据高时空性的需求和特点，设计了地球科学数据仓储的总体框架和功能，形成了地球科学数据仓储的基本架构，包括数据注册分系统、仓储前端应用分系统、数据管理分系统、用户管理分系统四大分系统。整体架构的设计解决了地学数据在特定时间和空间范围内获取问题，通过对CC协议的描述与解读，分析了数据仓储对于CC协议的现实需求，为地球科学数据的获取与共享提供了知识产权方面的支撑。本文在地球科学数据通用仓储建设框架方面的设计，及其CC协议方面的考虑，预期能够为不同学科数据仓储提供借鉴参考。

致谢：感谢深时数字地球大科学计划（DDE）、国家地球系统科学数据中心为本文研究提供了科研条件。

参考文献

[1] 王卷乐, 王玉洁, 张敏, 等. 2020年地球数据科学与共享热点回眸[J]. 科技导报, 2021, 39(1): 105−114.

[2] 司莉, 曾粤亮. 国外机构科研数据知识库研究进展[J]. 情报学报, 2017, 36(8): 859−870.

[3] WILKINSON M D, DUMONTIER M, AALBERSBERG I J, et al. The FAIR guiding principles for scientific data management and stewardship[J]. Scientific data, 2016, 3(1) : e1002295−D42.

[4] LIN D, CRABTREE J, DILLO I, et al. The TRUST principles for digital repositories[J]. Scientific data, 2020, 7(1) : IFPDA−15.

[5] 胡芳. 国外典型科学数据仓储实施的元数据方案及启示[J]. 图书与情报, 2015(1): 117−121.

[6] 李芳薇, 程瑾, 张群, 等. 国外图书馆生物医学科研数据管理服务及启示[J]. 中华医学图书情报杂志, 2015, 24(8): 5−10.

[7] 孔丽华, 习妍, 张晓林. 数据出版的趋势、机制与挑战[J]. 中国科学基金, 2019, 33(3): 237−245.

[8] 刘源. 大数据场景下的云存储技术及其应用研究[J]. 信息与电脑(理论版), 2020, 32(9): 13−14.

[9] 马瀚青, 杨小梅, 侯春梅, 等. 数据论文联合出版模式及数据论文出版[J]. 中国科技期刊研究, 2018, 29(7): 698−703.

[10] 刘晶晶, 顾立平, 范少萍. 国外通用型数据知识库的政策调研与分析[J]. 现代图书情报技术, 2015(11): 4−11.

[11] 张瑶, 吕俊生. 国外科研数据管理与共享政策研究综述[J]. 图书馆理论与实践, 2015(11): 47−52.

[12] 田丽, 李佳翼. 英国科研数据共享服务的经验与启示：以Figshare平台为例[J]. 图书馆学研究, 2018(23): 76−84.

[13] Figshare[EB/OL].[2021−10−05].https: //figshare.com/.

[14] SCIENCE DATA BANK[EB/OL].[2021−10−05].https: //www.scidb.cn/.

[15] PANGAEA[EB/OL].[2021−10−05].https: //www.pangaea.de/.

[16] 许义江, 李成龙, 谈昊林, 等. 表生地球化学数据库及大数据研究进展[J]. 高校地质学报, 2021, 27(1): 58−72.

[17] 邹丽雪, 欧阳峥峥, 王辉, 等. 生命科学领域科研数据仓储特点及服务分析[J]. 图书情报工作, 2016, 60(7): 59−66.

[18] 秦顺, 汪全莉, 邢文明. 欧美科学数据开放存取出版平台服务调研及启示[J]. 图书情报工作, 2019, 63(13): 129−136.

[19] 余文婷. 开放科学数据仓储资源开发模式比较分析：以SRDA、eCrystals和Dryad为例[J]. 图书馆学研究, 2014(11): 58−62, 92.

[20] Dryad[EB/OL].[2021−10−05].http: //datadryad.org.

[21] 郭华东. 地球大数据科学工程[J]. 中国科学院院刊, 2018, 33(8): 818−824.

[22] 董少春, 齐浩, 胡欢. 地球科学大数据的现状与发展[J]. 科学技术与工程, 2019, 19(20): 1−11.

[23] 于水婧. 知识共享协议与开放存取期刊出版[J]. 出版广角, 2016(22): 50−52.

[24] 杨静, 武晓耕. 灵活取用CC协议促进我国科技期刊国际化发展：以《西北工业大学学报》申请DOAJ为例[J]. 科技与出版, 2020(6): 108−113.

[25] 程铭, 潘云涛, 马峥, 等. 国内外学术期刊出版数据政策研究[J]. 科技与出版, 2021(4): 17−22.