专题 “科学数据资源、技术与政策”联合专刊

全球科学数据仓储平台的建设实践现状与展望

姜璐璐[1]，张泽钰[1]，李宗闻[1]，盖虹羽[1]，王鹏尧[1]，李成赞[1]，周园春[1*]

1. 中国科学院计算机网络信息中心，北京 100083

摘要：科学数据仓储平台是支撑开放数据实践的重要基础设施。科学数据仓储平台的建设在全球范围内已有较为广泛的实践，并探索形成了日臻完善的指导原则体系。本文整理了国际主要的数据仓储平台建设原则，并基于 re3data 的注册平台开展了针对性的调研，分析了现阶段国际科学数据仓储平台的建设实践情况，并重点分析了典型领域专业型数据仓储平台和全学科通用型仓储平台的实践特色。最后，本文重点分析了国际科学数据仓储在可信化、开放化、生态化方面的发展趋势，为我国科学数据仓储平台的建设和发展提供有益参考。

关键词：科学数据仓储平台；科学数据共享；建设原则；re3data；实践分析

ISSN 2096-2223

CN 11-6035/N

文献 CSTR：

32001.14.11-6035.csd.2023.0027.zh

文献 DOI：

10.11922/11-6035.csd.2023.0027.zh

文献分类：信息科学

收稿日期：2022-12-01

开放同评：2023-02-02

录用日期：2023-03-07

发表日期：2023-03-31

* 论文通信作者

周园春：zyc@cnic.cn

引 言

在数据密集型的科研范式下，科学数据既是科研活动的必要输入，也是科研活动的重要输出。科学数据的开放共享在维护科学自纠错能力、提升研究透明度、节约政府财政投入等方面[1]的诸多益处，已得到科学共同体的广泛认同。2021 年 11 月，联合国教科文组织大会第 41 届会议审议通过《开放科学建议书》[2]，标志着开放科学迈入全球共识的新阶段，开放数据是其中的重要组成部分。

国际各界一直以来都是十分关注解决科学数据共享中的关键问题。在政策要求层面，欧盟“地平线 2020”要求研究数据默认开放[3]，美国国立卫生研究院（National Institutes of Health，NIH）的新版数据管理计划与共享政策将于 2023 年 1 月 25 日生效[4]，德国[5][6]、英国[7]等国家的主要基金资助机构都不同程度地支持科学数据管理与共享。在我国，国务院办公厅于 2018 年印发《科学数据管理办法》[8]，明确了公共资金资助科学数据管理的职责、原则、方式和机制。在数据管理的最佳实践上，众多科学家和组织联合提出了科学数据的 FAIR（Findability、Accessibility、Interoperability、Reusability）指导原则[9]，并迅速在全球范围内得到普遍认同[10]。

作为支撑开放数据实践的重要基础设施，科学数据仓储平台扮演着十分重要的作用。它不仅是联接政策制定方（包括基金资助方、机构、期刊出版商等）与科研人员的重要纽带，也是数据管理与数据共享最佳实践的必要保障。标准化的、安全可信的科学数据仓储平台，除了可以帮助科研人员满足政府、项目资助方、机构以及学术出版商等对数据管理与数据共享的要求，更重要的是在确保数据 FAIR 化、数据共享生态可持续发展等关键问题上发挥效用。

国际范围内，科学数据仓储平台的建设发展备受关注，在标准化体系的建设上，国际科研数据联盟（Research Data Alliance，RDA）基于 FAIR 原则基础，提出了科学数据仓储平台建设的 TRUST（Transparency、Responsibility、User Focus、Sustainability、Technology）原则[11]、世界数据系统（World Data System，WDS）与 Data Seal of Approval（DSA）组织联合推出 CoreTrustSeal 认证[12]，等等。在科学数据仓储的建设上，越来越多的平台涌现。据 2022 年 11 月 23 日 re3data 注册数据显示，全球已有 3024 个科学数据仓储平台，其中，我国建设运维的平台占比仅为 3.17%。本文从科学数据仓储平台的建设原则调研，全球科学数据仓储平台建设实践及典型平台实践分析几个维度，揭示全球科学数据仓储平台的建设现状和发展趋势展望，为我国科学数据仓储平台的建设实践和平台发展提供参考。

1 科学数据仓储平台的建设原则

科学数据仓储平台作为开放科学体系的重要基础设施在数字资源长期保存、数据共享等方面发挥着重要作用。但是，如何确保数据的长期可访问、推动数据的可发现和可重用是科学数据仓储平台在实践层面上需要解答的重要问题。因此，开放科学事业的各方参与者从不同角度对数据仓储平台应具备的能力提出了各种建议与要求。

1.1 政府机构

联合国教科文组织（United Nations Educational Scientific and Cultural Organization，UNESCO）2021 年底发布《开放科学建议书》[2]，旨在为开放科学政策和实践提供一个国际框架，助力于缩小数字、技术和知识鸿沟。该建议书概述国际层面关于开放科学的通用定义、共同价值观、原则和标准，其中，对开放科学基础设施的建立完善提出相应建议和要求。建议书认为，开放科学基础设施应通过持久唯一的标识符明确识别科学对象，采用可互操作的标准和最佳实践确保数据可发现可重用；其次，应为管理、获取、分析、集成数据和科研文献等提供必不可少的开放和标准化服务，同时，不同存储库应根据所保存对象的特殊性、当地情况、用户需求和研究社群的要求做出相应改变

和优化；最后，建议书认为开放科学基础设施应不以营利为目的，并最大限度地保证所有公众可不受限制地永久使用。

NIH 在其数据管理与共享政策中对理想数据存储库的特征做出陈述，共包含 12 项内容，分别是：唯一且持久的标识符，长期可持续、元数据、数据治理和质量保证、免费便捷的访问、广泛且可衡量的重用、明确的使用指南、安全性和完整性、保密性、通用格式和标准、可溯源以及数据保留期政策。“元数据”部分强调确保数据附带元数据信息，以便更好地实现数据的发现、重用和引用。“广泛且可衡量的重用”涉及两方面，一是数据的广泛重用，这要求存储库尽可能使其数据集和元数据信息置于最有利于数据重用的条款之下；二是强调存储库应有能力追踪和评估数据的分发、引用及重用情况[23]。此外，NIH 还针对特殊领域的数据，例如人类数据提出特殊要求。

1.2 国际组织

作为推动科学数据共享 2020 年 5 月，RDA 旗下工作组提出 TRUST 原则，该原则为维护数字存储库可信度，尤其是负责研究数据管理的数据存储库提供了指引[11]。该原则由 5 部分组成，分别是：透明度（Transparency）、责任（Responsibility）、用户需求（User focus）、可持续（Sustainability）和技术（Technology）。其中，“透明度”建议通过公开可获取的证据验证数据存储库的功能和数据保存能力，例如存储库的数据政策、使用条款以及数据存储周期等；“责任”提出存储库应负责确保数据保存的真实性、完整性以及平台服务的可靠性、持久性；“用户关注”提到存储库要确保满足目标用户群体的数据管理规范标准和期待；“可持续”强调平台服务的可持续和数据资源的长期保存；“技术”要求平台具有足够能力和基础设施以支持安全、持久且可靠的服务。目前，TRUST 原则已获得来自全球不同国家各类组织的遵循和认可。国际出版商 Springer Nature，行业学协会美国地球物理学会（AGU），主流数据存储库 figshare、dryad，以及我国自主建设的科学数据存储平台科学数据银行（Science Data Bank，简称 ScienceDB）等均宣布认可并遵守 TRUST 原则。

CODATA 作为国际科学理事会（International Science Council，ISC）的数据委员会，其认为，可信赖的数字存储库是对被管理的数字资源提供长期可信赖访问的基础设施，它们存储、管理并治理数字对象，并在收到请求时做出反馈。可信赖的存储库需要经过一系列评估，例如 DSA、TRAC（ISO 16363）等[13]。CODATA 认为，虽然仓储平台的评估标准及方式有所不同，但必须满足一些质量标准以与其他存储方式区分开来。CODATA 致力于促进全球合作，提高所有按研究领域数据的价值和可用性，并支持研究产生的、易用于研究的数据应尽可能开放并在必要时尽可能封闭[14]。

1.3 学术出版商

国际出版商 Springer Nature 在其数据政策中提出优秀的数据存储库应具备的五大条件[15]：第一，确保数据长期存储，时间至少为出版后五年；第二，获得科研社群或科研机构的支持和认可；第三，为所存储数据提供稳定且永久的资源标识符；第四，允许数据访问且无非必要限制；第五，数据详情页提供明确的数据使用和访问条款，例如数据使用许可证等。此外，为方便作者提交数据至合适存储库，Springer Nature 还提供了学科领域型数据库和通用型数据库的推荐列表，并提到数据应尽可能提交给特定学科、社区认可的存储库，若无合适的学科特定存储库，数据可以提交给通用数据存储库[16]。

剑桥大学出版社（Cambridge University Press）鼓励作者提供支持其研究结果的各类证据，以提高研究透明度和可再现性。剑桥大学出版社建议作者选择符合以下五个条件的数据库：第一，社区支持并认可其所存储的内容；第二，为其所保存的内容分配永久唯一标识符；第三，致力于内容的长期保存和可访问；第四，允许为所存储数据的重用设置数据使用许可协议；第五，对其所存储内容的访问不向公众收取费用，但允许合理例外。此外，作者还应根据实际需求选择允许访问控制的存储库、支持提供盲审链接的存储库等[18]。

1.4 学术索引平台

DCI（Data Citation Index）致力于将各类学术资源建立关联，提供从众多全球跨学科资源库中获取优质研究资料的单一访问点[19]。DCI 提出数据存储库的基本出版标准，即持久且稳定、资助信息说明、数据同行评议、内容长期存续以及数据与论文建立关联。“资助信息声明”部分 DCI 强调，将会重点关注那些可以溯源数据关联文献，以及显示数据资助信息的存储库；“数据同行评议”部分强调，存储库需要关注数据总体质量和引用参考文献的完整度。此外，DCI 还会优先考虑提供数据引用或被数据引用的数据库，并表示会关注存储库的国际化及多样性，持续收录来自不同地理范围、学科领域的专有型或通用型数据库[20]。

BASE 是世界上最庞大的学术网站资源搜索引擎之一，基于 OAI-PMH（开放存档计划元数据收割协议）标准实现对论文、数据等学术资源的元数据收割，现已索引涵盖一万多个内容提供商的 3 亿多份文档资源[21]。BASE 对其内容提供者提出以下筛选标准：来源必须包含学术内容，至少部分文件/文档提供开放获取，文档的元数据通过有效的 OAI-PMH 界面提供。除此以外，BASE 还会定期查阅存储库目录（repository directories），例如 OpenArchives、ROAR、OpenDOAR 等平台，收集

和索引具有适当来源的内容[22]。

总结而言，各类主体普遍认为科学数据仓储平台应具备以下几个基本特征：为数据分配永久资源标识符，可确保数据的长期保存，可确保数据的开放获取，以及开展数据治理工作。此外，各类主体也关注平台的互操作性、数据政策情况、是否响应并满足用户需求、是否安全稳定可信赖，以及是否能建立数据与论文的关联等问题。简言之，促进实现数据的 FAIR 化，并确保数据长期存储依然是科学数据仓储平台的首要目标。

2 全球科学数据仓储平台的建设情况

全球科学数据仓储平台的建设已有相当丰富的实践，也积累了较为全面的建设情况统计分析。夏姚磺[24]基于 re3data 对中美科学数据仓储进行对比研究，从科学数据仓储的建设国家、学科分类、数据内容类型、数据开放获取程度、元数据标准等方面进行统计。Heinz Pampel[25]和王辉[26]对科学数据仓储进行了全面的分析，包括类型、学科、政策、内容、访问、技术等仓储基本情况。马瀚青[27]则从科技期刊的视角，对国际通用型数据仓储进行研究，包括基本功能、数据格式和规范、数据保存和访问，以及在数据仓储注册库的收录情况等。

基于第 1 章总结归纳的政府机构、国际组织、学术出版商、学术索引库等普遍关注的科学数据仓储平台的基本特征，本文选取国际科学数据注册仓储平台 re3data（https://www.re3data.org/）对全球数据仓储平台的建设情况进行调研、统计和分析。调研发现，截至 2022 年 11 月 23 日，re3data 上共注册仓储平台 3024 个，描述仓储平台的指标项 27 个，涵盖平台是否分配了永久资源标识、是否开放获取、是否开展数据治理、是否具有可信的资质等，与第 1 章总结的特征具有较强一致性和关联性。因此，本文将 re3data 中仓储平台 27 个指标进行筛选、分类，选择了 12 个指标，按照平台基本信息、平台开放性、平台标准化建设、平台条款和合规性四大类分别进行统计分析，详见表 1。需要说明的是，除“数据质量管理”外，其他指标项均为多选项，因此各指标存在数量相加之和大于仓储平台数量的情况。比如，仓储平台所属国家可以为多个国家，因此各国家平台数量相加大于 3024。

表 1　仓储平台建设指标项及分类

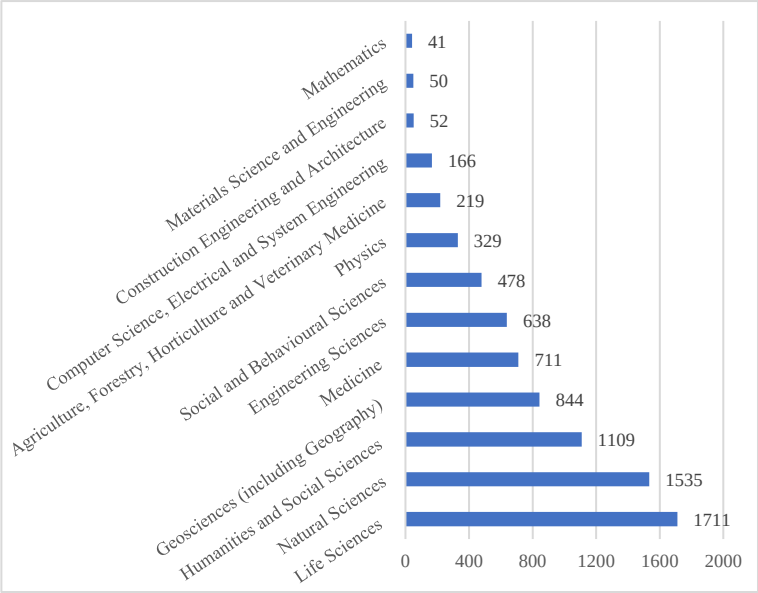
Table 1　Index items and classification of repository construction

分类	指标项
平台基本信息	学科分类
	所属国家

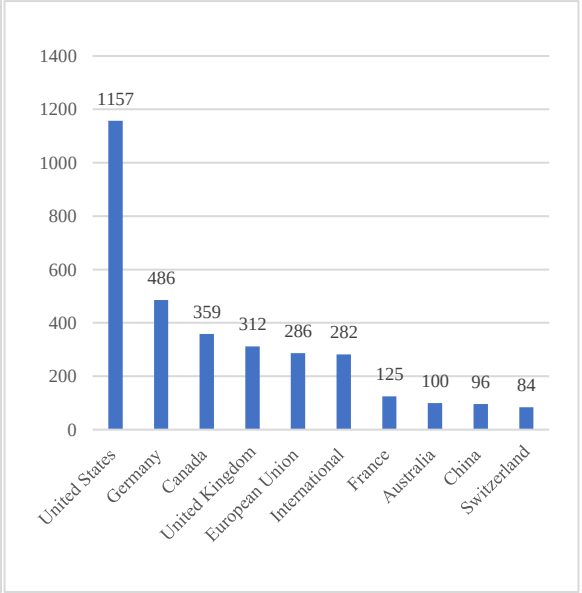
分类	指标项
	所属机构及责任类型
平台开放性	应用程序接口
	数据访问和限制情况
	数据上传和限制情况
平台标准化建设	平台的认证资质
	元数据标准
	数据永久标识
平台条款和合规性	数据许可协议
	数据库许可协议
	数据质量管理

2.1 平台基本信息

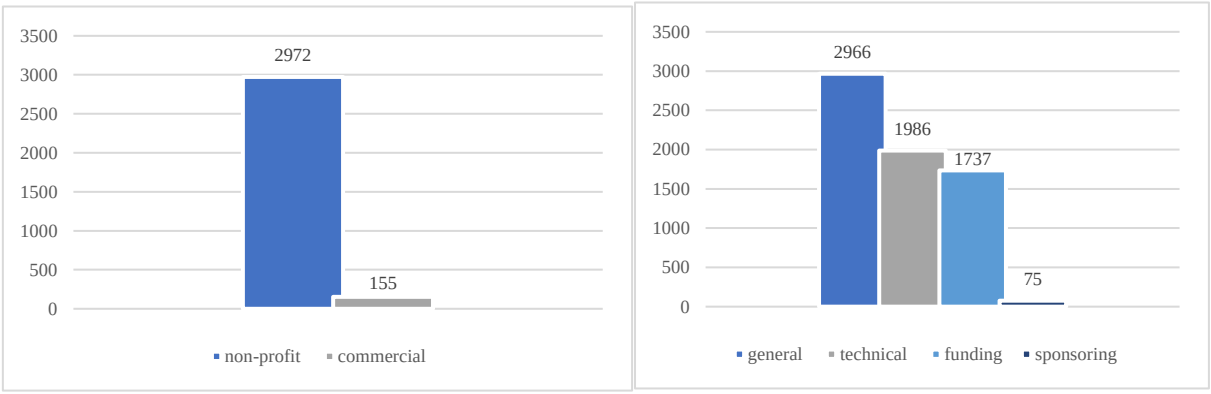
仓储平台的基本信息有助于我们了解全球仓储平台的建设基本情况。本文关注的基本信息包括仓储平台的学科分类（Subjects）、所属国家（Countries）、所属机构类型（Institution type）、所属机构责任类型（Institution responsibility type）。



(a) 平台学科分类情况统计



(b) 平台所属国家分布统计



(c) 平台建设机构类型统计　　(d) 平台建设机构责任类型统计

图 1　re3data 注册仓储平台基本信息的统计分布图

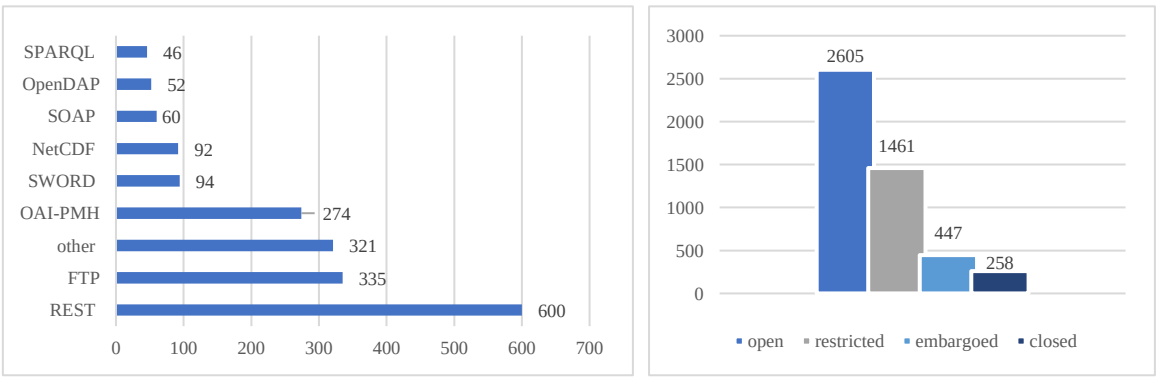
Figure 1　Statistical distribution of basic information of registered repositories in re3data

re3data 将注册的仓储平台分为 13 个一级学科，其中，生命科学（Life Sciences）、自然科学（Natural Sciences）、人文与社会科学（Humanities and Social Sciences）学科数据仓储平台数量位列前三名（**平台建设机构责任类型统计**

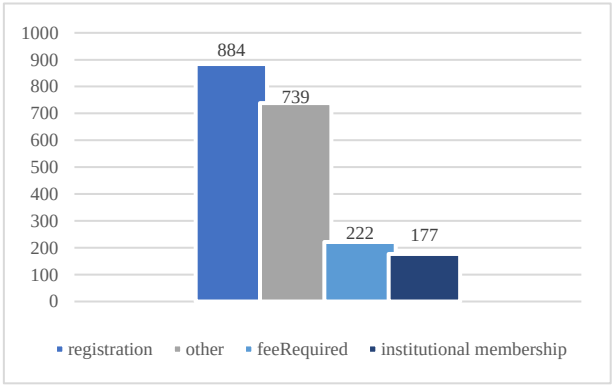
图 1-a）。3024 个仓储平台来自 84 个国家，美国、德国和加拿大的仓储平台注册数量位列前三名，共占全部注册平台的过半数量，中国排名第九，是前十名中唯一的亚洲国家（图 1-b）。此外，仓储平台的建设大多依靠非盈利的政府机构、研究机构、行业组织、高校等（图 1-c），并且它们除了常规管理事项，还主导平台建设的资金资助、技术支持等主要建设工作（图 1-d）。

2.2　平台开放性

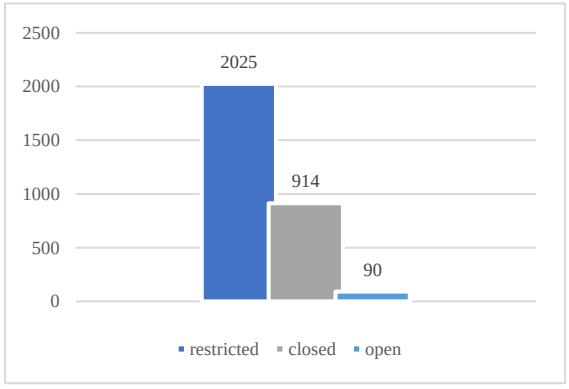
在开放科学（Open Science）和大数据时代的大背景下，开放数据（Open Data）也成为开放科学最为关键的环节[27]。因此，仓储平台的开放性，是仓储平台建设实践的重要方面。re3data 中能够体现仓储平台开放的元数据包括，应用程序接口（API）、数据访问情况（Data access）、数据访问限制（Data access restrictions）、数据上传情况（Data upload）、数据上传限制（Data upload restrictions）。



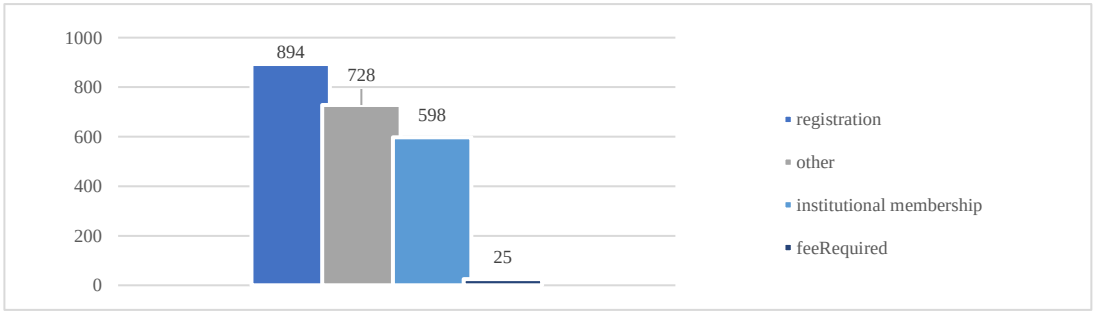
(a) 平台使用的应用程序接口统计　　(b) 平台数据访问限制情况统计



(c) 平台访问受限原因统计



(d) 平台数据上传限制统计



(e) 平台数据上传的限制原因统计

图 2　re3data 注册仓储平台的开放性统计图

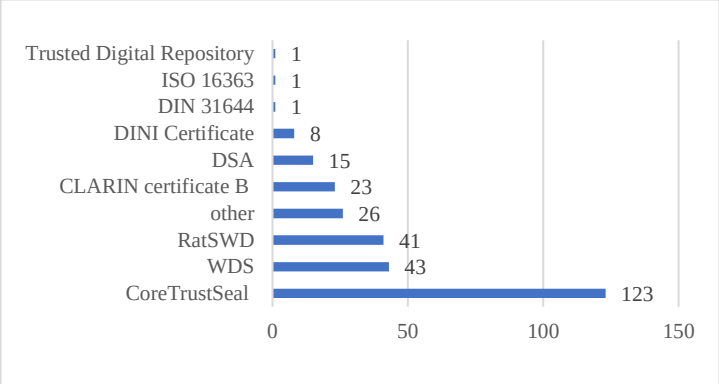
Figure 2　Statistical distribution of the openness of registered repositories in re3data

在 re3data 的 3024 个注册仓储平台中，小部分平台采用了应用程序接口。其中，使用最多的三个接口为 REST、FTP 和 OAI-PMH，分别占 19.84%、11.08%和 10.62%（图 2-a）。提供开放（open）数据访问服务的平台占比 86.14%，存在受限访问情况（restricted）的平台占比 48.31%，14.78%平台提供了数据访问的禁运保护（embargoed），8.53%的平台关闭（closed）数据访问服务（图 2-b）。在存在受限访问情况的平台里，主要的受限原因是仅注册用户可访问数据（60.51%），其次是机构会员资格（institution membership）和收费（feeRequired）限制，其他限制原因比例也较高（图 2-c）。

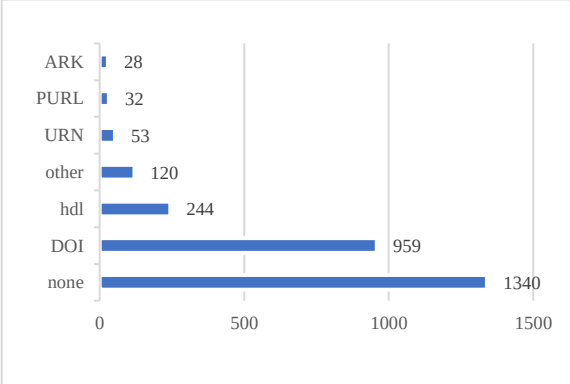
据统计，大部分数据仓储平台免费提供了开放的数据访问服务，但近半数平台通过用户注册、收费和会员资格等方式限制了数据访问。同时，这些也是仓储平台限制数据上传的方式。

2.3　平台标准化建设

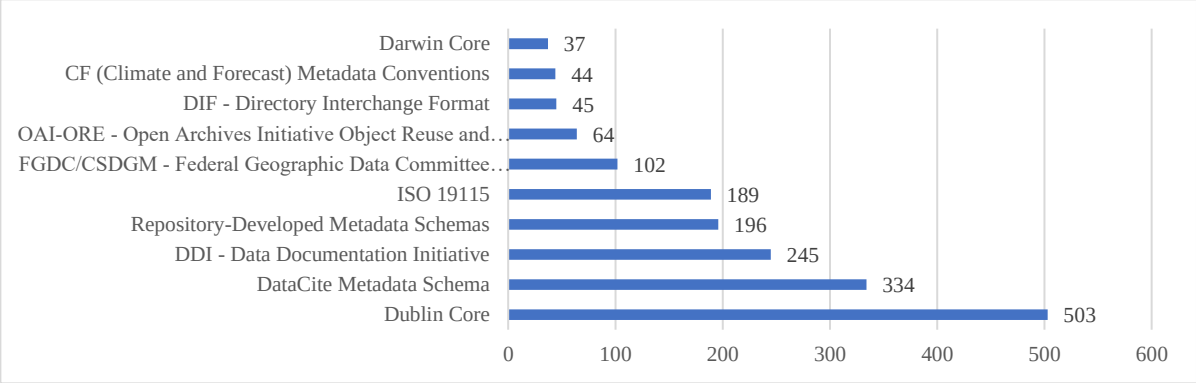
re3data 中平台标准化建设的指标包括平台获得的认证资质（Certificates）、元数据标准（Metadata standards）、永久标识（PID systems）。



（a）平台获得资质认证情况统计



（b）平台使用永久标识情况统计



（c）平台使用元数据标准情况统计

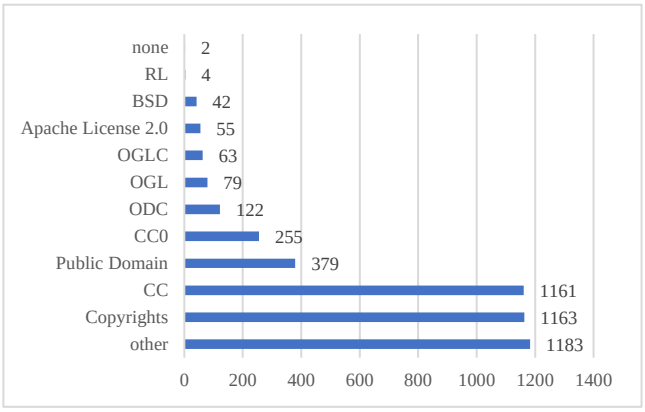
图 3　re3data 注册仓储平台标准化建设情况统计

Figure 3　Statistics on the standardization construction of registered repositories in re3data

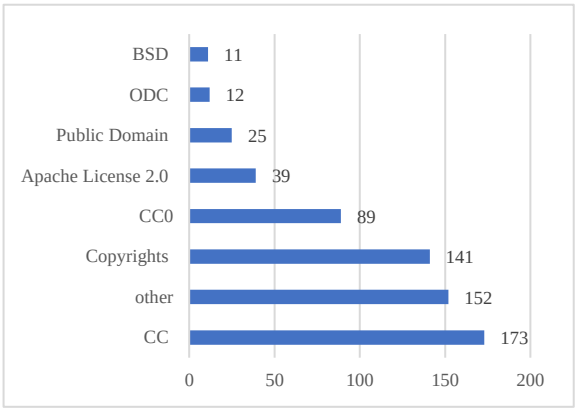
其中，获国际认证资质的平台占比较低（图 3-a），占比较多的是 CoreTrustSeal、WDS 认证和 RatSWD 认证。在数据永久标识的使用上，有 44.31%的平台没有为数据分配注册永久标识（图 3-b）。在元数据的标准化建设上，65.08%的平台有所遵循，其中采用占比最高的标准包括都柏林核心（Dublin Core）、DataCite Matadata Schema、DDI - Data Document Initiative（图 3-c）。统计显示，全球仓储平台的标准化建设方面还有很多工作亟待开展，绝大部分平台没有获得可信仓储的认证资质，并且在数据永久标识分配还存在近一半的缺口。

2.4　平台条款和合规性

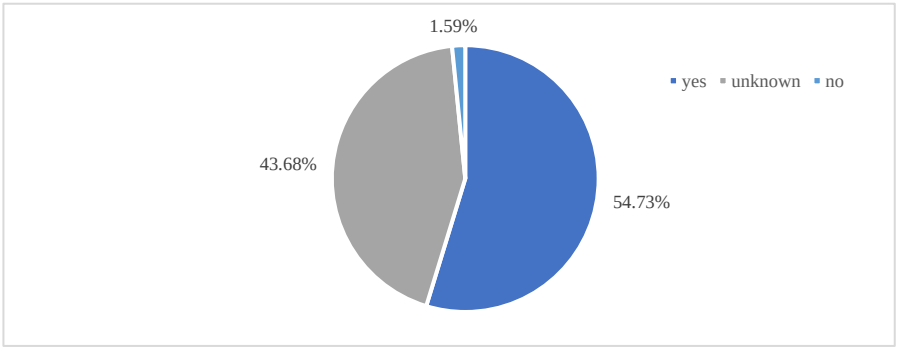
为了规范数据上传和数据使用行为，保障利益相关方的合法权益，仓储平台通常会提供许可协议供提交者选择，包括数据许可协议、数据库许可协议，并且可能对上传数据的质量进行管理。



(a) 平台提供的数据许可协议情况统计



(b) 平台提供的数据库许可协议情况统计



(c) 平台的数据质量管理情况统计

图 4　re3data 注册仓储平台的服务条款与合规性建设统计图

Figure 4　Statistical chart of service terms and compliance construction of registered repositories in re3data

统计显示，re3data 中大多数平台提供了版权协议（Copyright）、知识共享许可协议（Creative Commens License）和其他协议，其他平台还提供了共有领域（Public Domain）、CC0、开放政府许可协议（Open Government License，OGL）等协议（图 4-a）；只有非常小部分的平台提供数据库许可协议（图 4-b）。在数据质量管理上，54.73%的平台会对数据质量进行控制，1.59%的平台没有进行质量管理，43.68%的平台在这方面未明确说明（图 4-c）。可见，平台的数据质量管理需投入更多关注。

综上可知，re3data 注册的仓储平台在开放性、条款和合规性方面具有较好的建设实践（见表 2）。但在以下方面仍需开展优化建设：

平台开放性建设方面，大部分平台还应提供更加丰富的应用程序接口，提高平台的开放性和标准化建设能力，在提升共享数据的机器可读性和互操作能力上加强建设。

平台标准化建设上，应更加广泛地采用 DOI 等国际标准的数据永久标识，帮助数据溯源和引用；应遵循规范化的元数据标准，形成高质量元数据，帮助数据传播和重用。

平台合规性建设方面，应对数据形式、质量进行把控治理，提高数据价值，更好发挥验证科研

成果、复现科学研究、促进交流创新等作用。

表 2　仓储平台建设指标统计结果

Table 2　Statistical results of repository construction indicators

分类	总体评价	指标项	评价	统计结果
平台基本信息	-	学科分类	-	生命科学和自然科学领域的数据仓储占比最多
		所属国家	-	美国、德国、加拿大位列前三名
		所属机构及责任类型	-	98.28%的平台为非营利机构建设
平台开放性	优	应用程序接口	中	采用较少，且选择情况分散
		数据访问和限制情况	优	86.14%的平台提供数据开放获取
		数据上传和限制情况	优	大部分通过用户登录等方式限制上传
平台标准化建设	中	平台的认证资质	中	获得认证资质较少
		元数据标准	中	采用较少，且选择情况分散
		数据永久标识	良	三分之一采用 DOI 标识，近一半未分配永久标识
平台条款和合规性	中	数据许可协议	优	大多数平台提供了数据许可协议
		数据库许可协议	中	大多数平台未提供数据库许可协议
		数据质量管理	良	过半平台进行数据质量管理

3　典型科学数据仓储平台实践分析

依据平台发布数据的学科方向，科学数据仓储平台可分为领域专用型数据仓储平台和全学科通用型数据仓储平台。基于国际主流的科学数据仓储平台的建设原则，本文分别选取了两类数据仓储平台中的典型实践情况进行了分析整理。

3.1　领域专用型数据仓储平台案例

本文选取了国际上广受认可、学科特点明显的 5 个学科型仓储平台，作为重点分析对象，从平台定位、运营方式、学科特点、数据标识、数据政策、数据提交方式、审核制度、是否开放获取、所获认证等维度进行了调研（表 3）。

表 3　调研领域专用型数据仓储平台信息一览表

Table 3　List of information of investigated specialized data repositories

数据库名称	国家	学科	数据标识	数据提交方式	是否审核数据	是否免费
PANGAEA	德国	地理	DOI	在线提交、外部认证提交	是	大部分免费
GenBank	美国	基因序列	序列号	专门的提交门户	是	是
Nomad	欧洲	计算材料科学	DOI	在线提交，下载客户端提交	否	是
CCDC	国际组织	材料和生命科学	DOI、ISSN	在线提交	否	是
TPDC	中国	青藏高原科学数据	DOI	在线提交	是	是

5 个平台均为开放存取库，有较好的数据引用性、全面的元数据描述、良好的数据与元数据互操作性、优化的数据库存储结构，且保证其内容的长期可用性（通常为 10 年），大多数数据是免费提供的，依据许可条款使用，部分受密码保护的数据集，可在一定时长的保护期后被公开访问。每个数据集都可以通过一个 DOI 来识别、共享、发布和引用（例如，PANGAEA、NOMAD、CCDC），而 GenBank 采用特殊的序列号，CCDC 除 DOI 外，还可通过 ISSN 号引用。下文将逐个分析这些平台的建设特点。

3.1.1　PANGAEA

PANGAEA（https://www.pangaea.de/）由归档、发布和分发来自地球系统研究的地理参考数据，以及实验数据和模型/模拟。资源主要包含化学、岩石圈、生物分类法、大气、古生物学、海洋、生态学等 15 个小类。

在数据治理上，PANGAEA 聘请了覆盖地球和环境科学所有方向的科学家作为数据编辑。在数据服务上，PANGAEA 基于 Google Earth 提供了地理参考数据的可视化服务；支持元数据按图层描述，例如水深测量网络地图服务，提供了不同图层的摘要等元数据信息，便利用户使用平台数据。

3.1.2　GenBank

GenBank（https://submit.ncbi.nlm.nih.gov/subs/genbank/）美国国家生物技术信息中心（NCBI）建立，是世界上最大的核苷酸档案库，包含来自生命所有分支的序列，以及与它们相关的文献著作和生物学注释。GenBank 的数据来源主要是科研人员直接提供或大规模基因组测序计划。目前 GenBank 与 EMBL（欧洲 EMBL-DNA 数据库）、DDBJ（日本 DNA 数据库）建立了相互交换数据的合作关系，3 个数据库同步。

除提供免费的数据获取服务，GenBank 支持基因序列分析和比对，可结合 DNA Star 软件进行基因序列分析和比对；并提供基于 BLAST 的序列相似性检索服务。为便利信息索引，GenBank 内的每个序列名称紧跟其数据文件专区和主要的登录号；为便利信息分类与垂直服务，NCBI 为不同序列数据提供提交工具，如 GenBank 用于提交特定的数据类型（SARS CoV-2，流感，诺沃克病毒和登革热；原核 rRNA；细胞器 rRNA；真核 rRNA-ITS；后生动物 COX1），作者在提交后，可获得一个 GenBank number[29]。

3.1.3 Novel Materials Discovery（NOMAD）

NOMAD（https://nomad-lab.eu）主要由 NOMAD 卓越中心（CoE）开发，是一个保存并共享不同材料计算软件输入输出文件的领域专业大数据平台。平台致力于整合大量材料计算结果，并建成了全球最大的计算材料科学知识库，包含了超过 1 亿个高质量计算的输入和输出文件，可用于辅助大数据分析和材料设计[30]。

同时，NOMAD 提供了各类工具服务，如 NOMAD Encyclopedia 可帮助用户探索和理解使用各种工具和不同方法获得的计算；可视化工具可利用虚拟现实（VR）进行交互式索引，以可视化形式帮助人类理解材料结构；NOMAD Artificial Intelligence Toolkit 可以对所有可用的材料数据进行分类，识别相关性和结构情况，并检测趋势和异常等[31]。

3.1.4 剑桥晶体数据库中心（CCDC）

剑桥晶体数据中心（CCDC，https://ccdc.cam.ac.uk）是世界领先的材料和生命科学研究和发展相关结构化学数据、软件、知识的专家，是一家非盈利慈善机构，主要接收来自 x 射线、中子和电子衍射研究以及粉末研究的 CIF 格式的晶体结构数据。

其剑桥结构数据库 CSD（The Cambridge Structural Database）内每个结构都基于广泛的验证和交叉检查，通过自动化工作流程，结合化学家、晶体学家的手工管理进行质量控制。CSD 提供分子结构的 3D 可视化呈现服务，现已在全世界 70 个国家，1200 多家科研机构，150 多家制药及生物公司内得到了广泛的应用[32]。此外，平台提供小分子晶体数据的集合 CSD Communications 服务**错误！未找到引用源。**，该服务已为全球科学家提供了 40 000 多种可能从未公开的结晶结构。

3.1.5 国家青藏高原科学数据中心（TPDC）

国家青藏高原科学数据中心（http://www.tpdc.ac.cn/zh-hans/）是我国针对青藏高原及周边地区科

学数据门类最全、最权威的数据中心，依托中国科学院青藏高原研究所建设，共建单位包括兰州大学、北京师范大学和中国科学院计算机网络信息中心。TPDC 目前已整合的数据资源领域包括：大气、冰冻圈、水文、生态、地质、地球物理、自然资源、基础地理、社会经济等。

TPDC 提供特色专题数据集快速检索，并构建大数据质量控制、自动建模与分析、数据挖掘及交互式可视化的方法库、4 个模型资源库，实现青藏高原科学数据、方法、模型与服务的广泛集成。支持常规数据提交、未发表论文提交、项目数据提交 3 种提交方式，后者主要提交政府预算资金资助的各级科技计划（专项、基金等）项目所形成的科学数据。

典型领域专用数据库的建设实践，在遵循行业标准和建设原则之外，呈现出了以下发展特色：首先，配备精细化的分级分类能力。其中 GenBank、NOMAD 为典型代表。其次，纵深化的数据治理与加工能力，进一步提升了数据可重用性。其中，PANGAEA 与 CCDC 通过广泛的领域专家支持，实现了数据的深层治理水平。另外，具备领域垂直服务能力，在深度激发数据再利用机制上开展实践。领域垂直服务的典型模式是数据与服务工具相结合，如 GenBank、NOMAD、PANGAEA 和 TPDC。

3.2 全学科通用型数据仓储平台

全学科通用型存储库通常可接受存储和长期保存各种类型、各种领域方向的科学数据，且开放公共用户注册使用，对数据提交者所处的国家、地域、隶属机构情况不做限制。本文对 Springer Nature 数据政策中推荐的 6 个通用型数据存储库[17]，即 dryad、figshare、Harvard Dataverse（简称 HD）、Open Science Framework（简称 OSF）、ScienceDB、zenodo 进行了调研比较，依照前文对 re3data 的指标分析梳理形成了表 4。

表 4 典型通用型科学数据仓储平台情况一览表

Table 4 List of typical general scientific data repositories

平台名称	国家/性质	标准化数据接口/元数据标准	数据资源标识	数据质量机制控制	存储收费
dryad	美国/非盈利	OAI-PMH/ Dublin Core(DC)、OAI-ORE、DataCite Metadata Schema(DCMS)、RDF Data Cube Vocabulary	DOI、ORCID	平台提供同行评议	收费
figshare	英国/商业	OAI-PMH/ DC、DCMS、Schemar.org	DOI、ORCID	平台无审核服务；支持合作机构潜入审核工作流	20 GB 免费空间，超额收费
HD	美国/非盈利	OAI-PMH/ DC、Schema.org	UUID	提供数据审核服务	免费
OSF	美国/非盈利	OAI-PMH/ DC、Schema.org	OSF 编码	平台无审核服务	免费

平台名称	国家/性质	标准化数据接口/元数据标准	数据资源标识	数据质量机制控制	存储收费
ScienceDB	中国/非盈利	OAI-PMH/ DC、DCMS、Schema.org	DOI、CSTR、 ORCID、ROR	平台提供形式审核；支持共建机构学术审核	免费
zenodo	欧盟/非盈利	OAI-PMH/ DC、DCMS	DOI、ORCID	平台无审核服务；合作社区提供审核	免费

除表中列举的各项内容外，Springer Nature 遴选推荐的通用型存储平台在标准化、安全可信建设上都有良好的实践。例如平台都配备有明确的数据许可协议，以确保数据在明晰的框架下被使用。其中，除 dryad 仅支持严格的 CC0 协议外，其他 5 个平台提供了更多的可选项，其中适用于软件与代码的协议也包括在内；在提升数据可重用上，各平台均提供了版本控制机制。

在遵循行业标准和国际上数据共享和数据仓储建设的主流原则之外，各平台逐步探索发展出一些特色服务：

（1）数据文件可视化服务能力。figshare、OSF、ScienceDB 和 zenodo 提供了不同程度的数据文件在线可视化服务，其中，figshare 的服务特色最为突出。据统计，figshare 可支持上百种文件的在线可视化服务[34]，并公开提供可嵌入式的文件在线阅读 widget 服务，供第三方站点集成使用。Springer Nature、ACS 等国际出版商，学术资源索引库 Dimensions 等已在其论文浏览页面实现对其文件可视化 widget 集成。

（2）第三方资源集成导入服务。figshare、OSF、ScienceDB 与 zenodo 在不同程度上集成了 GitHub 的用户认证服务和代码库导入服务。此外，在链接外部资源上 figshare 支持了 GitLab、BitBucket 的资源库导入，OSF 连接了 Dropbox、Google Drive 的资源导入服务，为便利国际数据传输与共享开展了更广泛的实践。

（3）融入科研活动过程，接入科研成果发布流程。figshare 和 OSF 在融入科学数据管理过程中展开了深入尝试，2 个平台均提供了工作组内协同管理数据的服务支持；此外，OSF 支持使用 R 语言进行数据分析等服务。在论文成果发布的流程接入上，figshare、dryad 先后实现了与 Springer Nature 的投稿平台集成，ScienceDB 与科学出版社的 SciEngine 平台、方正鸿云系统实现集成接入。

（4）深入实践数据的防篡改机制建设。figshare、zenodo、HD、OSF 和 ScienceDB 提供了数据文件的数字文摘计算服务，均采用计算 MD5 值的方式以确保数据文件的完整性和一致性。此外，ScienceDB 集成了第三方区块链——科学数据链（Science Data Chain）服务，使得所有数据记账上链，覆盖数据集粒度和数据文件粒度的确权，并在数据防篡改上开展了更进一步的实践。

（5）实践科学数据评价计量追踪。科学数据的使用情况追踪可以更好地激励数据治理与数据共享[37]，调研平台在不同程度上都开展了数据使用情况的追踪计量。其中，ScienceDB 提供了数据访

问、数据下载的地理分布统计服务；figshare 和 ScienceDB 在站点集成了 Altmetrics、Dimensions 服务。此外，为推动数据使用追踪指标计量的标准化，各平台相继参与到 Make Data Count（MDC）计划[35]中，实现了符合 COUNTER（Counting Online Usage of NeT-worked Electronic Resources）[36]规范的数据计量与标准化推送服务。

4 发展趋势与展望

全球科学数据仓储平台的建设实践在国际组织、学术界、出版商、平台建设运维方等利益相关方的共同努力推动下取得了长足的发展，也逐步形成了一大批具有显著国际影响力的领域仓储平台和通用仓储平台。通过深入调研现阶段国际科学数据仓储平台的建设实践情况、重点分析典型领域专业型数据仓储平台和全学科通用型仓储平台的实践特色，可以发现，国际科学数据仓储平台的建设实践呈现出了可信化、开放化、生态化等发展趋势。

科学数据仓储平台的可信能力建设是仓储平台的最基本要求。国际出版商和学术索引库在推荐和收录数据仓储平台时，均以科学数据仓储平台服务的安全稳定可靠性作为重要评判依据。其中，TRUST 原则成为国际广泛遵循和认可的科学数据仓储平台可信能力建设指导性原则，从透明度、责任、用户需求、可持续和技术维度为科学数据仓储平台建设运行提供了指引。随着数据共享实践的广泛开展，科学数据仓储平台会继续在标准化、规范化上加强建设，除了在元数据规范、标识、许可协议、引用规范等方面上继续开展深入实践，并持续形成更加健全的行业标准。

科学数据仓储平台的开放性是保障仓储平台影响力和资源汇聚的有效举措。目前，全球知名科学数据仓储平台均提供了基于 FAIR 原则的数据开放获取服务，并提供了国际化的具有良好用户体验的服务和交互界面。在共享边界上，“应开放时尽可能开放，须保护时尽可能保护（As open as possible, as closed as necessary）”是重要的指导原则，科学数据仓储平台的分级开放服务能力逐步增强，在保护利益相关方的前提下，推动更广泛的数据共享实践。在 FAIR 原则的实践上，数据的可互操作性能力建设仍会是提升数据利用率、释放数据价值的重要建设方向。另外，具有显著资源规模和全球影响力的科学数据仓储平台均支持全球化开放性提交。

生态化和可持续发展模式是科学数据仓储平台的发展的生命线。目前，科学数据仓储平台大多采用公益性的非盈利服务模式，而科学数据仓储平台的建设和运行需要稳定、庞大和持续不断的人力、物力、财力保障，稳定的经费投入和良好的可持续发展模式是科学数据仓储平台建设面临的最大挑战。因此，数据仓储平台会在可持续化发展模式探索上开展实践。另外，随着用户规模和资源规模的增长，数据治理也将成为数据仓储平台面临的问题，采用云服务社区治理模式，依托数据社

区开展数据治理工作是目前主要大型的数据仓储平台的发展思路。此外，科学数据仓储平台的良好发展需要构建稳定可靠的上下游生态。无缝支持科学数据的生产与发布，无缝集成期刊论文的采编流程，持续开展科学数据的影响力追踪与评价等也是重要的实践趋势。

致　谢

感谢中国科学院科学传播专项“数字化平台建设”项目（No. 2022-qkcb-07），中国科学院“十四五”网信专项工程建设项目“科学大数据工程（三期）”（No. CAS-WX2022GC-02）资金资助。

作者分工与职责

姜璐璐（1989—），女，河南省郑州市人，硕士，工程师，长期从事数据出版与科学数据管理工作。主要承担工作：论文撰写，案例调研与分析。

张泽钰（1994—），女，河南省开封市人，硕士，工程师，研究方向为知识产权法。主要承担工作：论文撰写，案例调研。

李宗闻（1990—），女，北京市人，硕士，工程师，研究方向为数据出版。主要承担工作：论文撰写，数据整理。

盖虹羽（1993—），女，吉林省白山市人，硕士，助理工程师，研究方向为数据挖掘。主要承担工作：论文撰写，案例调研。

王鹏尧（1989—），男，山东省临沂市人，硕士，工程师，研究方向为科学大数据开发、应用及数据工程。主要承担工作：数据收集与清洗，数据整理。

李成赞（1982—），男，河南省邓州市人，博士，高级工程师，主要从事科学数据管理、出版与服务工作，重点开展科学数据出版模式、机制、关键技术等的研究探索与实践。主要承担工作：论文撰写，案例分析。

周园春（1975—），男，江西人，博士，研究员，博士生导师，研究领域为科学大数据、大数据分析与挖掘、知识图谱等。主要承担工作：论文研究思路、分析方法的指导。

参考文献

[1] BOULTON G, CAMPBELL P, FRENG B C, et al. Science as An Open Enterprise[M]. London: The Royal Society Science Policy Centre, 2012.

[2] UNESCO. UNESCO Recommendation on Open Science[R/OL]. (2021) [2022-11-22].

https://en.unesco.org/science-sustainable-future/open-science/recommendation.

[3] 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. DOI:10.13266/j.issn.0252-3116.2019.13.014. [QIN S, WANG Q L, XING W M. Research and enlightenment on the services of open access publishing platform for scientific data in Europe and America[J]. Library and Information Service, 2019, 63(13): 129–136. DOI: 10.13266/j.issn.0252-3116.2019.13.014.]

[4] NIH Sharing Policies and Related Guidance on NIH-Funded Research Resources. https://grants.nih.gov/policy/sharing.htm.

[5] Deutsche Forschungsgemeinschaft. (2022). Guidelines for Safeguarding Good Research Practice. Code of Conduct [EB/OL]. [2022-11-22]. https://doi.org/10.5281/zenodo.6472827.

[6] Recommendations for Policies of the Helmholtz Centers on Research Data Management. https://gfzpublic.gfz-potsdam.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_5010062

[7] Overview of funders' data policies. https://www.dcc.ac.uk/guidance/policy/overview-funders-data-policies

[8] 国务院. 国务院办公厅关于印发科学数据管理办法的通知 [EB/OL]. [2022-11-22]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm. [The State Council. Issued by the State Council general office on the measures for the management of scientific data to inform [EB/OL]. [2022-11-22]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.]

[9] WILKINSON M D, DUMONTIER M, AALBERSBERG I J, et al. The FAIR Guiding Principles for scientific data management and stewardship[J]. Nature Scientific Data, 2016, 3: 167-172.

[10] KHODIYAR V, LAINEH, O'BRIEN D, et al. Research Data: The Future of FAIR White paper [R/OL]. figshare, 2021. https://doi.org/10.6084/m9.figshare.14393552.v1

[11] RDA. The TRUST Principles: An RDA Community Effort [EB/OL]. (2020-05-18) [2022-11-22]. https://www.rd-alliance.org/trust-principles-rda-community-effort-0.

[12] CoreTrustSeal [EB/OL]. [2022-11-22]. https://www.coretrustseal.org/.

[13] CODATA. Trusted Digital Repository[EB/OL]. [2022-11-22]. https://codata.org/rdm-terminology/trusted-digital-repository/.

[14] CODATA. About. [EB/OL]. [2022-11-22]. https://codata.org/about-codata/.

[15] Springer Nature. Research Data Policies FAQs. Part 2: Questions about data repositories [EB/OL].

[2022-11-22]. https://www.springernature.com/gp/authors/research-data-policy/data-policy-faqs.

[16] Springer Nature, Data repository guidance [EB/OL]. [2022-11-22]. https://www.springernature.com/gp/authors/research-data-policy/recommended-repositories.

[17] Springer Nature. Generalist repositories [EB/OL]. [2022-11-22]. https://www.springernature.com/gp/authors/research-data-policy/repositories-general/12327166.

[18] Cambridge Core. Where to share your data [EB/OL]. [2022-11-22]. https://www.cambridge.org/core/services/authors/open-data/where-to-share-your-data.

[19] Data Citation Index. About us [EB/OL]. [2022-11-22]. https://clarivate.com.cn/solutions/data-citation-index/.

[20] Data Citation Index. Basic repository publishing standards [EB/OL]. [2022-11-22]. https://clarivate.com/webofsciencegroup/essays/the-repository-selection-process/.

[21] BASE. What is BASE [EB/OL]. [2022-11-22]. https://www.base-search.net/about/en/index.php.

[22] BASE. Become a content provider [EB/OL]. [2022-11-22]. https://www.base-search.net/about/en/faq.php#chap02.

[23] NIH. Selecting a Data Repository [EB/OL]. [2022-11-22]. https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/selecting-a-data-repository.

[24] 夏姚璜. 基于 re3data 的中美科学数据仓储对比研究[J]. 图书馆学研究, 2018(6): 17–26.DOI:10.15941/j.cnki.issn1001-0424.2018.06.003. [XIA Y H. A comparative study on the scientific data repositories based on Re3data between China and United States[J]. Research on Library Science, 2018(6): 17–26. DOI: 10.15941/j.cnki.issn1001-0424.2018.06.003.]

[25] PAMPEL H, VIERKANT P, SCHOLZE F, et al.呈现科研数据知识库:re3data.org 注册机制[J].现代图书情报技术,2014(03):26-34. [PAMPEL H, VIERKANT P, SCHOLZE F, et al. Making Research Data Repositories Visible: The re3data.org Registry[J]. Modern Technology of Library and Information Service,2014(03):26-34.]

[26] 王辉, WITT M. 基于 re3data 的科研数据仓储全景分析[J].图书情报工作, 2017, 61(22): 69–76.DOI:10.13266/j.issn.0252-3116.2017.22.009. [WANG H, WITT M. Panoramic analysis of research data repositories based on Re3data[J]. Library and Information Service, 2017, 61(22): 69–76. DOI: 10.13266/j.issn.0252-3116.2017.22.009.]

[27] 马瀚青, 关琳琳, 孔丽华, 等. 数据仓储该如何助推中国科技期刊开放数据?——基于国际科技

期刊数据仓储的对比分析[J]. 中国科技期刊研究, 2022, 33(4): 470–477. DOI: 10.11946/cjstp.202108220669. [MA H Q, GUAN L L, KONG L H, et al. Data repositories promote the data sharing of Chinese scientific journals: comparison of data repositories of international scientific journals[J]. Chinese Journal of Scientific and Technical Periodicals, 2022, 33(4): 470–477. DOI: 10.11946/cjstp.202108220669.]

[28] Pangaea. About [EB/OL]. [2022-11-22]. https://www.pangaea.de/about/.

[29] GenBank. Submission Portal [EB/OL].[2022-11-22]. https://submit.ncbi.nlm.nih.gov/subs/genbank/.

[30] DRAXL C, SCHEFFLER M. The NOMAD laboratory: from data sharing to artificial intelligence[J]. Journal of Physics: Materials, 2019, 2(3): 036001.

[31] NOMAD Center of Excellence[EB/OL].[2022-11-22]. https://nomad-coe.eu.

[32] CCDC[EB/OL].[2022-11-22]. https://ccdc.cam.ac.uk.

[33] 甘豫. 剑桥结构数据库-全球小分子晶体结构的集散中心[C]//中国晶体学会.中国晶体学会第五届全国会员代表大会暨学术大会（小分子分会场）论文摘要集, 2012:1.[GAN Y. Cambridge Structural Database - Global Hub for Small Molecular Crystal Structure[C]// Abstracts from the 5th National Membership Congress and Academic Conference of Chinese Crystallographic Society (Small Molecular Branch), 2012:1.]

[34] File formats supported for in-browser viewing. [EB/OL]. [2022-11-22]. https://help.figshare.com/article/file-formats-supported-for-in-browser-viewing.

[35] Make Data Count. [EB/OL]. [2022-11-22]. https://makedatacount.org/.

[36] Project COUNTER. [EB/OL]. [2022-11-22]. https://www.projectcounter.org/.

[37] PIERCE H H, DEV A, STATHAM E, et al. Credit data generators for data reuse[J]. Nature,2019: 30-32.DOI:10.1038/d41586-019-01715-4.

论文引用格式

姜璐璐, 张泽钰, 李宗闻, 等. 全球科学数据仓储平台的建设实践现状与展望[J/OL]. 中国科学数据, 2023, 8(1). (2023-03-27). DOI: 10.11922/11-6035.csd.2023.0027.zh.

The construction practice and prospects of global scientific data repository platform

JIANG Lulu[1], ZHANG Zeyu[1], LI Zongwen[1], GAI Hongyu[1],

WANG Pengyao[1], LI Chengzan[1], ZHOU Yuanchun[1*]

1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100083, P.R. China

ABSTRACT: As an essential infrastructure, scientific data repositories play an important role in promoting the practice of open research data. As a bridge between polices and researchers, it makes the data sharing possible. However, it is a challenge to build a repository to be trustworthy and in compliant with FAIR principles. More and more guidelines on how to select a repository and what a trustworthy repository should be like have been brought up. The research elaborates several popular principles and extracts their common requirements. Upon those, the paper analyzes the repositories registered on the website of re3data and summarizes their current development. Moreover, the research focuses on the featured practices in some typical repositories, including domain-specific and generalist data repositories. At last, the research analyzes the development trend of international scientific data repositories in terms of trustworthiness, openness, and ecologization, which can be used as an instructive reference for the construction and development of similar platforms in China.

KEYWORDS: research data repository; open research data; practice guidelines; re3data; case studies