

基于FAIR原则的中国科学数据资源 现状分析及启示

李骐安¹ 孟宪飞² 张书华¹ 张璐¹ 张蓓¹ 窦天芳¹

(1. 清华大学图书馆, 北京 100084; 2. 清华大学科研院, 北京 100084)

摘要: 为推动开放科学实践, 促进科学数据共享与重用, 选取re3data、FAIRsharing和Data Citation Index这3个国际主要科学数据仓储 (SDR) 注册和发现平台作为主要数据源, 开展中国科学数据资源现状与国外的对比分析, 发现我国有国际影响力的SDR数量与科学数据集的体量不匹配, 包括我国在内的各国科学数据集的传播共享主要依靠欧美国家的SDR; 基于FAIR原则分析我国典型SDR面临的挑战, 结果表明, 我国的国家科学数据中心在数据和元数据的可发现性、可访问性方面受到数据标识符分配不足和无法正常解析的影响。在数据和元数据的可交互性和可重用性方面, 存在引用信息和使用说明不完整、不清晰等问题, 阻碍数据的传播和共享。国际组织和机构的典型经验与举措为我国的科学数据管理提供了启示。基于此, 未来我国可以通过开展对FAIR原则的政策支持和规范实施, 提高我国科学数据资源的开放共享水平。

关键词: 开放科学; 科学数据; 科学数据仓储; FAIR原则; 中国

中图分类号: G239.1; G203 **DOI:** 10.3772/j.issn.1673-2286.2023.01.007

引文格式: 李骐安, 孟宪飞, 张书华, 等. 基于FAIR原则的中国科学数据资源现状分析及启示[J]. 数字图书馆论坛, 2023 (1): 50-57.

随着当今科学技术全球化的不断深入, 开放、共享正成为科学研究的核心内涵与鲜明特征, 开放科学 (Open Science) 的实践得到越来越多人的认可和采用^[1]。科学数据是开放科学的重要物质基础, 科学数据资源的管理与共享水平是衡量一个国家整体科技水平和综合国力的一项重要标志^[2]。

我国非常重视科学数据的管理与共享。自1984年正式加入国际数据委员会 (CODATA) 并成立中国委员会以来, 我国陆续启动气象、林业、农业等科学数据共享中心的建设与服务试点。随着全球科技创新能力和投入的不断增强, 科学数据采集能力持续提升, 但同时也给科学数据的存储、管理和共享带来了挑战。在此背景下, 科学数据仓储 (Scientific Data Repository, SDR) 迅速发展, 成为促进科学数据开放共享的重要载体, 例如, 新冠疫情暴发以来汇聚疫情数据的权威平台

GISAID、美国政府的数据门户以及我国的20个国家科学数据中心^[3]。

为提高科学数据资源的显示度和使用率, 克服数据发现与重用的障碍, 国际学术界提出面向科学数据管理的FAIR原则^[4], 从可发现 (Findable)、可访问 (Accessible)、可交互 (Interoperable)、可重用 (Reusable) 四方面指导和评估科学数据管理实践。基于这一原则, 国内外学者对科学数据资源管理的现状、趋势及存在问题进行了研究。王辉等^[5]基于re3data探讨了全球范围内SDR在数据访问权限、数据标识符使用等方面的建设特点。Cho^[6]对亚洲地区SDR的数据使用许可、数据标识符等进行了调研。Zhang等^[7]发现越来越多的开放科学数据研究者开始关注数据的所有权、使用许可等相关问题。李春秋等^[8]基于FAIR原则调研了我国医学领域的SDR, 指出在数据标识符使用和数据

收稿日期: 2022-11-25

溯源描述方面的不足。翟军等^[9]对我国政府数据的开放情况进行调研,发现各地政府数据存在元数据标准不统一、数据标识符缺失、交互性差、使用许可不清晰等问题。邢文明等^[10]从政策制定视角分析了我国《科学数据管理办法》对FAIR原则的支持度,结果表明只有少部分FAIR原则关注的要素得到了完全支持。虽然已有研究中探讨的数据访问权限、数据使用许可和标识符是评估科学数据管理水平的重要因素,但SDR的显示度、数据标识符的可解析性、数据的引用方式及使用说明等要素对于促进科学数据的发现与重用具有同样的重要性。

1 研究方法

本文首先将我国的科学数据资源现状从SDR层面和科学数据集层面与国外进行比较,选取re3data、FAIRsharing、Data Citation Index (DCI)等3个国际主要SDR注册和发现平台作为数据统计和分析的来源。本文进一步从FAIR原则的四个方面对我国典型SDR面临的挑战进行分析,选取我国的20个国家科学数据中心^[3]作为数据统计和分析的来源。本文中的SDR指一类用于长期存储和访问科学数据的信息基础设施^[11],是与知识库类似的一类数据库^[12];而科学数据中心则可以包含一个或多个数据库。为便于理解并与国际表述一致,本文采用SDR泛指数据仓储、数据平台、知识库、数据库、科学数据中心这一类信息基础设施。此外,本文中的科学数据集指科学数据文件的集合。re3data是由德国研究基金会提供资助,于2012年启用的涵盖全球范围内不同研究学科SDR的注册平台。本文基于re3data采集了2 803个SDR样本数据,数据采集截至2022年2月。FAIRsharing由英国牛津大学创办于2011年,是一个社区驱动的资源服务机构,包含科学数据标准、SDR以及科学数据政策3类相互关联的资源目录。本文基于FAIRsharing共采集了1 851个SDR样本数据,数据采集截至2022年2月。DCI数据库提供了对来自跨学科全球知识库的高质量研究数据的单点访问功能。本文基于DCI数据库采集了398个SDR样本数据,数据采集截至2022年2月,数据时间段为1990—2022年。此外,基于DCI数据库采集了10 038 370个科学数据集样本数据,数据采集截至2022年2月,数据时间段为2010—2021年。我国的20个国家科学数据中心是在原有国家平台的基础上于2019年优化调整形成

的,旨在完善我国科技资源共享服务体系,推动科技资源的开放共享。对国家科学数据中心的数据采集截至2023年1月。

2 中国科学数据资源现状与国外的比较

SDR是管理和传播科学数据资源的重要载体与平台。基于re3data、FAIRsharing、DCI国际主要SDR注册和发现平台,从SDR层面将中国的科学数据资源现状与国外进行对比分析。美国的SDR数量处在绝对领先地位,在re3data平台登记了1 141个,占该平台SDR总量的40.7%。德国、英国、欧盟等国家或国际组织的SDR数量也较多,与美国类似主要在re3data平台上登记。相比之下,中国的SDR数量较少,主要在FAIRsharing平台上登记(102个),但仅占到该平台SDR总量的5.5%;在re3data和DCI平台中则分别占各平台SDR总量的1.7%和2.5%。中国在上述3个国际主要SDR注册和发现平台登记的SDR有119个,按规则在世界范围内进行SDR元数据等信息的共享。

基于DCI从科学数据集层面对中国与其他国家的科学数据资源体量及存储现状进行分析可以发现,中国科学数据集的数量以较为明显的优势超过除美国以外的其他国家而位居第二,达到158 243个。这与我国在国际主要SDR注册和发现平台登记的SDR数量所处的排名形成鲜明对比。对这些科学数据集的主要分布SDR进行分析发现,科学数据集数量排名前20的SDR大部分来自美国,占总量的55%,其他主要来自英国、德国等欧洲国家以及欧盟等国际组织。结果表明,我国丰富的科学数据集与有国际影响力的SDR数量不匹配,包括我国在内的各国科学数据资源主要依靠欧美国家的SDR实现全球的共享传播。

3 基于FAIR原则的中国典型科学数据仓储分析

本文从FAIR原则的可发现、可访问、可交互、可重用四方面(见图1)出发,对我国科学数据资源的管理与共享现状进行分析。保障科学数据的可发现性是落实FAIR原则其他方面的重要前提。而科学数据一经潜在的数据使用者和计算机发现,就会通过可信的SDR提供的服务被访问及获取。科学数据通常需要与其他数据集成整合,通过使用标准定义和通用的语言

与应用程序或工作流进行交互操作,以开展数据的分析、存储和处理等工作。实现科学数据的广泛重用是实践FAIR原则的目标。通过对数据(元数据)进行准确、充分的描述和说明,以便潜在数据使用者正确使用数据。

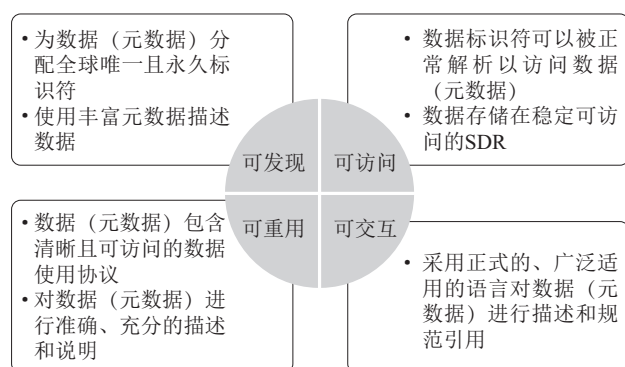


图1 面向科学数据使用和传播的FAIR原则及主要内容

3.1 可发现性

保障数据可发现性有两个重要技术手段。其中一

个是为数据(元数据)分配全球唯一且永久的标识符,这被认为是FAIR原则中最重要的方面^[13]。表1列出了我国20个国家科学数据中心在数据使用和传播方面的现状及面临的问题。在20个国家科学数据中心中,有10个为数据分配了DOI,但其中有6个只为部分数据分配了DOI,占比60%;有4个国家科学数据中心的SDR为全部数据分配了DOI,其中在国际主要SDR注册和发现平台上登记的只有国家天文科学数据中心和国家青藏高原科学数据中心两个。此外,基于我国自主制定的科技资源标识体系,有14个国家科学数据中心为数据分配了中国科技资源标识符(China Science and Technology Resource, CSTR),占比达70%;但其中有3个只为部分数据分配了CSTR。相比之下,有5个国家科学数据中心没有为数据分配DOI和CSTR这两类国内广泛使用的标识符,其中包括在国际主要SDR注册和发现平台上登记的国家基因组科学数据中心和国家气象科学数据中心。有4个国家科学数据中心为所有数据分配了DOI和CSTR两类标识符,分别是国家天文科学数据中心、国家青藏高原科学数据中心、国家生态科学数据中心、国家冰川冻土沙漠科学数据中心。

表1 国家科学数据中心数据使用和传播现状

序 号	名 称	使用文档	引用方式	DOI	DOI是否可解析	中国科技资源标识(CSTR)	CSTR是否可解析	是否在国际主要SDR注册和发现平台上登记
1	国家高能物理科学数据中心	有	有	无		有	否	否
2	国家基因组科学数据中心	无	有	无		无		是
3	国家微生物科学数据中心	有	有	有(部分数据)	是	有	是	否
4	国家空间科学数据中心	有	有	有(部分数据)	是	有(部分数据)	是	是
5	国家天文科学数据中心	有	有	有	是	有	是	是
6	国家对地观测科学数据中心	有	无	无		无		否
7	国家极地科学数据中心	有	有	有(部分数据)	否	无		是
8	国家青藏高原科学数据中心	有	有	有	是	有	是	是
9	国家生态科学数据中心	有	有	有	是	有	是	否
10	国家材料腐蚀与防护科学数据中心	无	无	无		无		否
11	国家冰川冻土沙漠科学数据中心	有	有	有	是	有	是	否
12	国家计量科学数据中心	无	无	无		无		否
13	国家地球系统科学数据中心	有	有	有(部分数据)	是	有	是	是
14	国家人口健康科学数据中心	有	有	有(部分数据)	是	有	否	是
15	国家基础学科公共科学数据中心	无	有	无		有	否	否
16	国家农业科学数据中心	无	有	无		有	否	否
17	国家林业和草原科学数据中心	有	有	无		有	是	是
18	国家气象科学数据中心	有	无	无		无		是
19	国家地震科学数据中心	有	有	有(部分数据)	是	有(部分数据)	否	是
20	国家海洋科学数据中心	有	有	无		有(部分数据)	否	否

注:“空格”表示无此项。

保障数据可发现性的另一个技术手段是为数据提供丰富的元数据。以国家青藏高原科学数据中心为例, 该SDR提供了描述性、技术性、管理性、权限管理等元数据, 从关键词、时空范围、引用方式、项目信息、数据贡献者等多个方面对数据进行了详细描述, 进一步提高了数据的可发现性。类似的, 国家生态科学数据中心提供了从DOI、CSTR、数据格式等基本信息, 到数据描述信息、生产者信息、共享服务信息、关联出版论文信息等丰富的元数据, 保障了数据的可发现性。

3.2 可访问性

科学数据的可访问性体现在数据(元数据)可以通过DOI等标识符的解析机制, 或使用某种链接规范实现从数据(元数据)发现到数据获取的能力。这样即使科学数据不再可用, 元数据仍可被人和计算机无障碍访问。这就需要一方面确保数据标识符可以被任何能够连接到互联网的潜在数据使用者正常解析, 进而访问科学数据; 另一方面保障数据存储在可信的、稳定可访问的SDR中。截至2023年1月, 20个国家科学数据中心官网均可以正常访问。在数据标识符的可解析性方面, 获取的数据中除国家极地科学数据中心外, 其他SDR的DOI均可被正常解析; 为数据分配CSTR标识符的14个国家科学数据中, 存在无法被正常解析情况的有6个, 约占43%, 其中包括4个没有DOI的SDR。CSTR标识符无法正常解析的SDR包括在国际主要SDR注册和发现平台上登记的国家人口健康科学数据中心和国家地震科学数据中心。虽然这两个国家科学数据中心通过在re3data、FAIRsharing等国际平台上登记提高了其全球显示度, 但数据标识符的解析问题阻碍了潜在使用者对数据(元数据)的正常访问。

在为所有数据分配了DOI和CSTR两类标识符的4个国家科学数据中, 两类标识符均可正常解析, 同时还明确定义了数据访问的条件和权限。以国家天文科学数据中心为例, 其数据简介中明确列出了数据的共享途径、共享范围以及获取数据的申请流程。类似的, 国家青藏高原科学数据中心、国家生态科学数据中心、国家冰川冻土沙漠科学数据中心在元数据中明确描述了数据的存在状态和获取方式, 促进了数据的共享与访问。

3.3 可交互性

数据的可交互性原则旨在保障潜在的数据使用者之间能够方便、可靠地交换和整合数据资源。这就需要采用正式的、广泛适用的语言和详细的数据资源描述框架对数据(元数据)进行描述。数据引用是国内外数据共享界提出的新概念, 旨在建立数据与数据之间以及数据与文献之间的关联, 进而促进数据的广泛交互。我国于2017年发布了《信息技术 科学数据引用》(GB/T 35294—2017)国家标准, 旨在规范科学数据引用元素的描述方法以及引用格式。分析国家科学数据中心的数据可交互性, 结果表明有4个国家科学数据中心没有提供明确的数据引用方式, 其中包括在国际主要SDR注册和发现平台上登记的国家气象科学数据中心。相比之下, 国家高能物理科学数据中心、国家基因组科学数据中心、国家空间科学数据中心、国家青藏高原科学数据中心、国家基础学科公共科学数据中心等大多数SDR提供了较为完整的数据引用信息, 包含数据贡献者、名称、版本信息、创建和传播机构以及数据唯一标识符等丰富的元数据信息。

国家科学数据中心提供的数据引用方式形式多样。以国家青藏高原科学数据中心为例, 除了对数据或数据出版文献的直接引用外, 还包括对与数据研究背景、产生过程、处理方法和质量评价等相关关联文献的引用, 同时提供数据使用者基于数据所发表文献的引用信息。类似的, 国家基因组科学数据中心主要通过提供数据关联文献的方式规范数据引用。国家农业科学数据中心在提供数据引用信息的同时, 通过制定《农业科学数据交换格式规范》来提高数据的可交互性。相比之下, 虽然国家微生物科学数据中心的数据具有DOI和CSTR标识符, 但在数据引用信息中并未提供上述两类标识符, 而提供了数据的创建和传播机构以及URL信息; 国家地球系统科学数据中心和国家海洋科学数据中心则在数据引用信息中提供了数据的创建和传播机构信息。

3.4 可重用性

保障科学数据可重用性的重要手段是通过采用标准化的数据组织方式对数据(元数据)进行准确、充分的描述和说明。在数据使用说明中提供法律层面的数据使用许可协议能够确保数据重用过程中知识产

权的清晰明确^[13]。同时,数据的来源、产生过程及涉及的相关贡献者等信息的提供也有利于潜在数据使用者对数据的重用,降低数据复用的复杂度。分析国家科学数据中心的数据可重用性,结果表明有5个没有提供清晰的数据使用文档或指南,其中包括在国际主要SDR注册和发现平台上登记的国家基因组科学数据中心,这阻碍了潜在使用者对数据的再利用。

各个国家科学数据中心提供的数据使用说明没有一个相对统一的形式。国家青藏高原科学数据中心和国家冰川冻土沙漠科学数据中心提供了明确的知识共享许可协议(Creative Commons license, CC协议)。Kindling等^[11]的研究显示,CC协议是全球范围内的SDR最常用的数据使用许可协议。虽然国家生态科学数据中心没有提供CC协议这类较为清晰的数据使用协议,但以数据出版的形式提供了数据使用指南,从数据产生背景、数据采集和处理方法、数据质量控制、数据使用方法等方面着手,帮助潜在数据使用者了解和正确使用数据。国家地震科学数据中心制定了《地震科学数据共享管理办法》,并针对特定类型数据提供了包括数据来源、数据产生或加工办法、数据质量说明等内容在内的元数据信息。类似的,国家地球系统科学数据中心、国家微生物科学数据中心、国家空间科学数据中心等大多数SDR提供了从数据使用声明到数据来源、数据产生或加工方法、数据质量说明等丰富且清晰的元数据,引导潜在数据使用者正确使用数据。受学科特点和数据敏感性等因素影响,虽然国家高能物理科学数据中心、国家对地观测科学数据中心、国家人口健康科学数据中心在数据使用方面有一定的限制和约束条件,但仍通过提供数据来源、数据校准方法、数据分析软件等元数据尽可能地让潜在使用者了解和正确使用数据。相比之下,国家天文科学数据中心、国家海洋科学数据中心除了提供数据作者或所有者信息外,并未提供清晰的有关数据来源、数据加工处理方法、数据质量说明等方面的元数据。

4 国内外科学数据管理实践现状及对我国的启示

4.1 国内外SDR建设实践

我国注重加强和规范科学数据管理,但相比于欧美国家起步较晚,在国际主要SDR注册和发现平台

上登记的SDR数量有限,与我国科学数据集的体量不匹配,科学数据资源基本依靠欧美国家中具有国际影响力的SDR进行传播共享。实际上,美国、英国等欧美国家长期支持SDR的发展。英国于2000年启动了e-Science计划^[14],推进在信息化基础设施支持下的科学研究活动,并为全球的学术合作提供支撑。美国国家科学基金委(NSF)于2003年提出通过网络基础设施(cyberinfrastructure)促进科学和工程学科的革新^[15]。2008年,欧洲开放获取基础设施研究项目(OpenAIRE)在欧盟第七框架计划(FP7)资助下展开^[16],并于2012年开始支持科学数据的开放共享。此后的欧盟第九框架计划(FP9),即“欧洲地平线”项目要求2021年及之后由欧洲研究理事会(ERC)资助的项目在形成科学数据前必须提交数据管理计划,然后将数据存入可信的SDR,并按照“尽可能开放,必要时封闭”的原则提供数据访问服务^[17]。类似的,我国在2018年发布的《科学数据管理办法》中,对政府预算资金资助形成的科学数据提出了“开放为常态、不开放为例外”的共享原则。2021年出台的国家标准《科技计划形成的科学数据汇交 技术与管理规范》(GB/T 39912—2021),推动政府预算资金资助的各级科技计划(专项、基金等)项目将科学数据汇交至我国的20个国家科学数据中心,规范数据汇交管理,促进科学数据的共享和重用。

4.2 国内外FAIR生态建设现状

我国已有部分国家科学数据中心在国际上亮相,但在科学数据资源的可发现性、可访问性、可交互性、可重用性方面仍有较大的发展空间。实现科学数据可持续传播与共享的重要前提是为数据分配全球唯一且永久的数据标识符。我国部分国家科学数据中心为数据分配了DOI和CSTR两种标识符,提高了科学数据的显示度,同时也保障了数据贡献者的知识产权^[18]。此外,国家科学数据中心也通过提供丰富的元数据保障了数据的可发现性。在科学数据的可访问性方面,部分国家科学数据中心曾出现过不能访问或数据标识符无法正常解析的情况。在这种情况下,即使明确了数据访问的条件和权限,仍会阻碍科学数据被潜在使用者访问和获取。实际上,王辉等^[5]、Kindling等^[11]对全球SDR的研究均显示有超过90%的SDR可正常访问,仅有不足0.5%的SDR处在无法访问的状态。在科学数据的

可交互性方面, 部分国家科学数据中心缺乏数据引用信息, 而提供引用信息的SDR也没有较为统一的引用规则或标准, 阻碍了数据的交互操作, 甚至数据的追踪计量^[18]。在科学数据的可重用性方面, 部分国家科学数据中心没有对数据(元数据)进行准确、充分的描述和说明, 存在使用说明缺失、数据使用协议不清晰的问题, 限制了潜在使用者对科学数据的了解和正确使用。这些在一定程度上反映出我国在FAIR原则的政策支持和规范实施中存在的不足。

欧美国家的组织、科研机构已在政策制定和实施过程中积极落实FAIR原则。欧盟是最早实践FAIR原则的国际组织^[19-20], 其将FAIR原则贯穿数据管理的政策和法规之中^[21]; 而欧洲开放科学云(EOSC)则为落实FAIR原则提供了重要的基础设施保障^[9]。欧盟委员会在“欧洲地平线2020”中启动了一项“开放研究数据试点”项目, 通过制定《FAIR数据管理指南(2020)》(Guidelines on FAIR Data Management in Horizon 2020)帮助研究人员更好地管理其数据^[22]; 同时, 欧盟委员会还在2016年成立了FAIR数据专家组^[23], 从技术和政策角度更好地推进数据的FAIR管理。欧洲研究图书馆协会(LIBER)也在积极推动由“欧洲地平线2020”委员会制定的FAIR数据行动计划^[24]。此外, 国际数据委员会、研究数据联盟(RDA)等国际组织也在积极倡导科学数据的FAIR生态建设^[25-26], 以提高数据的可交互性和可重用性。2016年, 澳大利亚在FAIR指导工作组的推动下提出了关于《FAIR获取澳大利亚研究成果》的声明^[27]。2017年, 德国、荷兰和法国联合成立GO FAIR国际支持与协调办公室^[28], 推动FAIR倡议的落地。国外高校也积极实践FAIR原则。2020年, 包括美国大学协会、欧洲研究型大学联盟、加拿大研究型大学U15集团在内的9个大学组织共同制定了《索邦大学研究数据权益宣言》, 其中做出了大学支持数据FAIR的承诺^[29]。

4.3 科学数据管理实践现状对我国的启示

本文从SDR层面和科学数据集层面将我国的科学数据资源现状与国外进行比较, 并从数据(元数据)的标识符及其可解析性、数据(元数据)的规范引用、数据(元数据)的使用协议及使用说明等方面分析我国典型SDR在落实FAIR原则中存在的挑战。当前, 我国面临的挑战主要有: 不重视SDR的国际传播工作, 即

使建设了国家层面的SDR, 也未在国际上亮相; 在SDR的可访问性、数据标识符的分配及可解析性方面存在不足, 一定程度上使得我国部分科学数据资源只能通过国外SDR进行传播和共享; 科学数据的引用和使用说明存在不完整、不清晰的问题, 阻碍了科学数据的共享和重用。

本文虽然没有涉及FAIR原则中的所有技术细则, 但上述问题的存在是建立科学数据传播与共享生态的主要障碍, 不利于对数据贡献者知识产权的保护以及数据使用者对科学数据的正确使用。我国在科学数据管理中需要针对FAIR原则提供政策支持和规范标准。为此, 我国可以借鉴国际组织和机构在SDR建设、数据标识符应用、数据引用原则和数据使用协议规范等方面的举措和经验, 推动和激励数据生产者和使用者参与到数据的开放共享生态中, 加强我国科学数据资源的管理水平和国际传播能力。

5 结语

科学数据的开放共享是推动科技创新和经济社会发展的重要前提。FAIR原则的提出旨在进一步指导和促进科学数据资源的可发现、可访问、可交互和可重用, 已被国际科研机构和组织广泛采纳。虽然我国在科学数据管理的FAIR生态建设方面较欧美国家起步晚且存在诸多挑战, 但借鉴国际上科学数据FAIR文化和生态建设中的经验做法, 开展对FAIR原则的政策支持, 并形成实施FAIR原则的标准规范, 增加科学数据发现、访问、交互和重用的可能性, 推进科学数据资源的可持续开放共享, 能为我国成熟的开放科学生态的形成和国家科技创新和经济社会发展提供重要保障。

参考文献

- [1] TENNANT J P, CRANE H, CRICK T, et al. Ten hot topics around scholarly publishing [J]. Publications, 2019, 7 (2): 34.
- [2] 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.
- [3] 科技部 财政部发布国家科技资源共享服务平台优化调整名单 [EB/OL]. [2022-05-24]. http://www.gov.cn/xinwen/2019-06/11/content_5399105.htm.
- [4] 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: 160018.
- [5] 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.
- [6] CHO J. Study of Asian RDR based on re3data [J]. The Electronic Library, 2019, 37 (2): 302-313.
- [7] ZHANG Y, HUA W, YUAN S. Mapping the scientific research on open data: a bibliometric review [J]. Learned Publishing, 2018, 31 (2): 95-106.
- [8] 李春秋, 杜博雅, 耿寒, 等. 医学科学数据开放平台FAIR原则的应用评估与调查分析 [J]. 图书情报工作, 2022, 66 (3): 72-82.
- [9] 翟军, 梁佳佳, 吕梦雪, 等. 欧盟开放科学数据的FAIR原则及启示 [J]. 图书与情报, 2020 (6): 103-111.
- [10] 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.
- [11] KINDLING M, PAMPEL H, VAN DE SANDT S, et al. The landscape of research data repositories in 2015: a re3data analysis [J/OL]. D-Lib Magazine, 2017, 23 (3/4) [2023-01-01]. https://www.researchgate.net/publication/314966856_The_Landscape_of_Research_Data_Repositories_in_2015_A_re3data_Analysis. DOI:10.1045/march2017-kindling.
- [12] SANSONE S A, MCQUILTON P, ROCCA-SERRA P, et al. FAIRsharing as a community approach to standards, repositories and policies [J]. Nature Biotechnology, 2019, 37 (4): 358-367.
- [13] 邢文明, 郭安琪, 秦顺, 等. 科学数据管理与共享的FAIR原则——背景、内容与实施 [J]. 信息资源管理学报, 2021, 11 (2): 60-68, 84.
- [14] HEY T, TREFETHEN A E. The UK e-science core programme and the grid [J]. Future Generation Computer Systems, 2002, 18 (8): 1017-1031.
- [15] ATKINS D E, DROEGEMEIER K K, FELDMAN S, et al. Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure [EB/OL]. [2022-11-24]. <https://www.nsf.gov/cise/sci/reports/CH1.pdf>.
- [16] RETTBERG N, SCHMIDT B. OpenAIRE - Building a collaborative open access infrastructure for European researchers [J]. LIBER Quarterly: The Journal of the Association of European Research Libraries, 2012, 22 (3): 160-175.
- [17] FAIRsharing.ORG. Open Research Data and Data Management Plans - Information for ERC grantees by the ERC Scientific Council [EB/OL]. [2022-11-24]. <https://doi.org/10.25504/FAIRsharing.iKa3Xm>.
- [18] 潘小多, 李新, 冉有华, 等. 开放科学背景下的科学数据开放共享: 国家青藏高原科学数据中心的实践 [J]. 大数据, 2022, 8 (1): 113-120.
- [19] HODSON S, JONES S, COLLINS S, et al. Turning FAIR Data into Reality: Interim Report from the European Commission Expert Group on FAIR Data [EB/OL]. [2022-12-21]. <https://doi.org/10.5281/zenodo.1285272>.
- [20] MONS B, NEYLON C, VELTEROP J, et al. Cloudy, increasingly FAIR: revisiting the FAIR Data guiding principles for the European Open Science Cloud [J]. Information Services & Use, 2017, 37 (1): 49-56.
- [21] Sustainability of (Open) Data Portal Infrastructures-Dataset Reus: A Method for Transforming Principles into Practice [EB/OL]. [2022-12-21]. https://data.europa.eu/sites/default/files/sustainability-data-portal-infrastructure_3_dataset-reuse.pdf.
- [22] H2020 Programme-Guidelines on FAIR Data Management in Horizon 2020 [EB/OL]. [2022-12-21]. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.
- [23] EC Expert Group on Turning FAIR Data into Reality [EB/OL]. [2022-12-20]. <https://codata.org/initiatives/previous-codata-working-groups/fair-data-expert-group/>.
- [24] LIBER. Open Consultation on FAIR Data Action Plan [EB/OL]. [2022-12-21]. <https://libereurope.eu/article/fairdataconsultation/>.
- [25] 陈书贤, 刘桂锋, 刘琼. 国内外科学数据管理FAIR原则研究进展及应用综述 [J]. 农业图书情报学报, 2022, 34 (8): 30-41.
- [26] CODATA. Making Data Work for Cross-Domain Grand Challenges: the CODATA Decadal Programme [EB/OL]. [2022-12-21]. <https://codata.org/initiatives/decadal-programme2/>.
- [27] THE FAIR STEERING GROUP. Policy Statement on F.A.I.R. Access to Australia's Research Outputs [EB/OL]. [2022-12-21]. <https://www.fair-access.net.au/fair-statement>.
- [28] GO FAIR. GO FAIR International Support and Coordination Office [EB/OL]. [2022-12-21]. <https://www.go-fair.org/go-fair-initiative/go-fair-offices/>.
- [29] Sorbonne Declaration on Research Data Rights [EB/OL]. [2022-12-21]. <https://sorbonnedatadeclaration.eu/>.

作者简介

李骥安, 男, 1991年生, 博士, 馆员, 研究方向: 科学数据管理。

孟宪飞, 男, 1977年生, 博士, 副研究员, 研究方向: 科研管理与科技政策。

张书华, 男, 1982年生, 博士, 馆员, 研究方向: 数据治理。

张璐, 女, 1992年生, 硕士, 工程师, 研究方向: 数字资源标准。

张蓓, 女, 1979年生, 硕士, 副研究馆员, 研究方向: 图书馆信息化。

窦天芳, 女, 1973年生, 硕士, 副研究馆员, 通信作者, 研究方向: 电子资源管理、学术传播机制、信息系统, E-mail: ddoutf@mail.tsinghua.edu.cn。

Status Analysis and Implications of Scientific Data Resources in China Based on FAIR Principles

LI QiAn¹ MENG XianFei² ZHANG ShuHua¹ ZHANG Lu¹ ZHANG Bei¹ DOU TianFang¹

(1. Tsinghua University Library, Beijing 100084, P. R. China; 2. Research & Development Affairs Office, Tsinghua University, Beijing 100084, P. R. China)

Abstract: To promote open science in practice as well as scientific data sharing and reuse, three registries of scientific data repositories (SDRs), namely re3data, FAIRsharing and Data Citation Index, are employed as the main data sources to carry out a comparative analysis of the current status of scientific data resources in China and abroad. We also analyze the challenges faced by typical SDRs in China based on FAIR principles. The number of SDRs with international impacts in China is mismatched with scientific datasets, and the dissemination and sharing of scientific datasets from various countries, including China, mainly rely on SDRs in European and American countries. The findability and accessibility of scientific data and metadata in national scientific data centers in China are affected by the insufficient allocation of data identifiers and the inability to resolve these identifiers properly. In terms of the interoperability and reusability of scientific data and metadata, there are several challenges such as incomplete and inexplicit citation information and usage instructions, which impede the dissemination and sharing of scientific data. The representative experiences and initiatives of international organizations and institutions provide insights for scientific data management in China. By conducting the policy support and regulatory implementation of FAIR principles so as to improve the level of open sharing of scientific data resources in China.

Keywords: Open Science; Scientific Data; Scientific Data Repository; FAIR Principles; China

(责任编辑: 雷雪)