科学通报 2024 年 第 69 卷 第 9 期: 1183 ~ 1191
《中国科学》杂志社 SCIENCE CHINA PRESS
香山科学会议专栏 观 点
第S69次学术讨论会·科学数据治理与利用的前沿和热点

FAIR原则全球影响分析及其实施策略建议

王卷乐[1,5*], 李扬[2], 王淑强[3], 石蕾[4], 段博文[1]

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. 中国科学院文献情报中心, 北京 100190;
3. 中国科学院地理科学与资源研究所图书馆, 北京 100101;
4. 国家科技基础条件平台中心, 北京 100862;
5. 江苏省地理信息资源开发与利用协同创新中心, 南京 210023
* 联系人, E-mail: wangjl@igsnrr.ac.cn

数据是大数据时代重要的科技基础条件和战略科技资源. 虽然我们生活在数据丰富的世界中, 但是数据形态和大小各异, 可共享使用的数据十分有限. 在大数据和人工智能时代, 数据共享变得越来越重要, 当前国内外高度重视科学数据共享和治理. 国际科学理事会(International Science Council, ISC)下属的世界数据系统(World Data System, WDS)在全球拥有128个数据中心(https://www.worlddatasystem.org/community/membership). 美国国家航空航天局(National Aeronautics and Space Administration, NASA)自20世纪90年代开始已建立了12个分布式活动存档中心(Distributed Active Archive Centers, DAACs)[1]. 国际数据仓储系统(re3data.org)截至2023年6月已收录了3125个数据库(https://www.re3data.org/). 中国于2019年组建了国家高能物理科学数据中心、国家基因组科学数据中心、国家地球系统科学数据中心等20个国家科学数据中心和31个资源库馆(https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/qtwj/qtwj2019/201906/t20190610_147031.html). 围绕科学数据的共享和治理, 一系列新的数据期刊也蓬勃发展, *Earth System Science Data* (ESSD)和PANGAEA数据仓储的结合实现了高质量论文和高质量数据同步发表(https://www.pangaea.de/). 国家自然科学基金委员会于2009年启动共享航次计划, 促进基金项目海洋调查数据开放共享实践[2]. 以数据共享和知识发现为目标的深时数字地球(Deep-time Digital Earth, DDE)国际大科学计划于2019年由国际地质科学联合会(International Union of Geological Sciences, IUGS)发布, 其愿景是通过数据和知识驱动推动地球科学研究范式的变革[3]. 随着数据密集型科学研究范式和开放科学[4]的发展, 对开放数据的获取、访问、互操作等治理需求也更为紧迫.

在这个背景下, Wilkinson等人[5]于2016年在*Scientific Data*以“The FAIR Guiding Principles for scientific data management and stewardship”为题, 发表了FAIR(可发现、可访问、可互操作和可重用, Findable, Accessible, Interoperable and Reusable)原则. FAIR原则旨在定义一套使机器和人类都能够查找、访问、互操作和重用数据与元数据的规则条款. FAIR原则共涉及4个部分, 15个条款, 是科学数据治理领域的重要成果. FAIR原则提出后, 得到许多国际组织和学界的响应. 2018年11月, 世界数据系统联合90多个国际组织共同签署了《赋能FAIR数据承诺声明》(*Enabling FAIR Data Commitment Statement*), 确定了共同支持开放和FAIR化数据的目标(https://www.copernicus.org/news_and_press/2018-11-05_enabling-fair-data-commitment-statement.html). 国际研究数据联盟(Research Data Alliance, RDA)成立了GO FAIR利益组织并出台了RDA Interest Group Draft Charter GO FAIR Interest Group(https://rd-alliance.org/group/go-fair-ig/case-statement/go-fair-ig-charter-statement)章程, 进一步将FAIR原则从理念推向实践. 欧洲联盟(简称欧盟)迅速跟进这一原则, 在欧盟委员会的报告《将FAIR变为现实》里提供了27项FAIR ACTION计动建议[6].



王卷乐 中国科学院地理科学与资源研究所研究员、博士生导师, 任国际科学联合会世界数据系统科学委员会委员、中国自然资源学会资源生态专业委员会主任. 长期参与国家科学数据中心建设和科学数据治理研究.

FAIR原则的提出迅速得到了多个领域数据治理的支持和响应. 这个学术传播影响过程辐射多少全球学科领域和区域? 影响传播过程如何? 其对科学数据治理实践有何直接作用? 针对这些问题, 本研究结合文献计量方法和科学数据共享实践开展分析, 试图量化揭示FAIR原则的影响力, 并结合

© 2024《中国科学》杂志社

科学数据治理的行动和实践提出相关发展建议.

1 研究方法

(ⅰ) 数据源. 本研究的数据源主要包括Web of Sciences™ (WoS)核心合集Sciences Citation Index(SCI)、Social Sciences Citation Index(SSCI)索引数据、Incite数据库、Essential Science Indicator™(ESI)数据库、Journal Citation RecordsTM (JCR)数据库. 统计时间窗: 2016年1月1日~2022年12月31日. 为加强数据分析的可比性, 本文以WoS学科分类体系为依据进行学科领域分析.

(ⅱ) 方法. 首先, 以FAIR原则为核心概念构建了检索式, 并在WoS核心合集中进行了检索, 获得研究FAIR原则的论文. 其次, 以首篇提出FAIR原则的论文(The FAIR Guiding Principles for scientific data management and stewardship)的施引文献作为主线(以下简称种子论文), 检索到引用该文的所有论文. 针对这两类检索结果进行文献计量分析, 分析项目包括引用时间、引文分布学科、引文发文国家、关键词词频、关键词主题聚类等. 然后, 从网络和文献中跟踪FAIR原则影响下的各国际组织和学术团体活动, 进一步形成总体认识, 并提出FAIR数据治理发展建议.

2 FAIR原则的文献计量分析

2.1 FAIR原则研究的发文态势分析

截至2022年底, FAIR原则的相关研究共计发文333篇. 首篇论文为发表在*Scientific Data*上的"Comment: The FAIR Guiding Principles for scientific data management and stewardship", 作者的第一单位是莱顿大学. 如图1所示, 依据发文数量可将研究阶段分为两个时期: 阶段一(2016~2018年), 年发文数量较少, 但是年发文增长率大于45%; 阶段二(2019年至今), 年发文数量明显增加, 但发文增长率降低, 2021年年发文增长率达到42%, 其他年份发文增长率最高未超过25%.

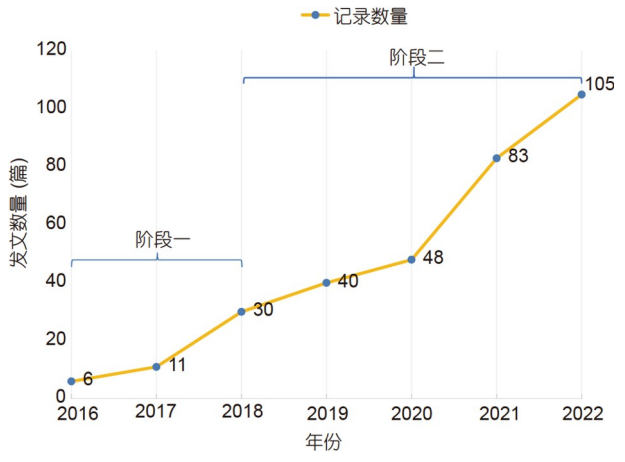


图 1 FAIR原则研究发文趋势
Figure 1 The trend of publications on the FAIR principles

截至2022年底, FAIR原则施引文献集合共计发文3393篇. 相关论文在种子论文发表当年即有FAIR主题论文出现在期刊上. 最早的一篇文献可以追溯至2016年12月, 这篇题为"I'll take that to go: Big Data bags and minimal identifiers for exchange of large, complex datasets"的会议论文发表在2016 IEEE *International Conference on Big Data*上, 作者第一单位是美国芝加哥大学. 该文讨论了大数据工作流中的数据组装和数据交换等问题, 并以生物医学数据集为例, 提出了大数据工作流中创建数据集的典型方法、工具、用途和可能出现的遗漏错误. 引文的年增长率达为78.5%, 2022年共有874篇论文引用了该文(图S1).

2.2 FAIR原则研究的期刊与学科分布

FAIR原则研究的发文分布在214个期刊, 期刊影响因子(impact factor, IF)最高的期刊为*Nature*. 影响因子小于1的期刊数量极少, 多数论文发表在影响因子在1~10区间内的期刊上, 约1/4的论文发表在影响因子在10~20区间的期刊. 如表1所示, 发文主要涉及生物、生命、材料、医药等多个基础和前沿领域, 以及数据管理和归档的多学科交叉领域. 依据WoS学科分类体系(WoS对期刊按照252个学科方向进行了分类标引, 一篇可以同时属于多个WoS学科分类), 本领域研究共涉及110个研究方向, 发文数量大于100篇论文的学科共计7个, 其中, 数量多的研究主要集中在计算机科学信息系统、计算机科学、生物学(特别是医学)、环境科学等多学科科学.

施引论文中共有81篇高被引论文, *Frontiers in Marine Science*和*Nucleic Acids Research*刊载了不止一篇高被引论文. 这反映出在海洋科学和生命科学领域对FAIR原则的高度关注. Top 10的高被引论文期刊主要发表在*Nucleic Acids Research*、*BMJ-British Medical Journal*、*New England Journal of Medicine*、*Nature*和*Journal of Neural Engineering*, 引用最高的一篇达2664次. Top10作者第一单位依次为剑桥大

表 1 FAIR原则发文期刊
Table 1 The FAIR principles issuing journals

期刊名称	载文数量(篇)
Scientific Data	16
Database-The Journal of Biological Databases and Curation	12
Nucleic Acids Research	8
Frontiers in Marine Science	5
GigaScience	5
Journal of Biomedical Informatics	5
PLoS Computational Biology	5
PLoS One	5

学、爱丁堡大学、南加利福尼亚大学、莫纳什大学、罗格斯州立大学(2篇)、麻省理工学院、美国国立卫生研究院、马里兰大学和蒙特利尔大学.

期刊引文指标(journal citation indicator, JCI)是衡量期刊在最近3年内发表的可引用项目(文献和审阅)的平均类别归一化引文影响力(category normalized citation impact, CNCI), 可用于比较不同学术领域期刊的引文影响力. 其中11篇高被引论文的发文机构包含中国机构, 如表2所示.

2.3 FAIR原则的国家竞争力与国家间合作

InCites数据库以WoS核心合集数据库为基础, 集成了其特有的"分析集成指标系统", 结合ESI和JCR关键指标, 可以充分揭示各学科领域内国家、机构、学者间的学术竞争力. 本节采用InCites数据库开展国家间发文分析与国家间合作分析.

2.3.1 FAIR原则施引文献的国家发文分析

检索结果显示, 参与本领域研究的国家共计137个, 其中发文数量最高的国家包括美国、德国和英国, 发文数量均超过了800篇. 如图2所示, 发文数量大于300篇论文的国家共计10个, 欧美国家构成了施引文献发文的主体(中国发文159篇, 排第15位). 其中, 美国、德国、英国和荷兰的发文数量均超过了500篇. 英国和澳大利亚的篇均论文被引频次最高, 反映了这两个国家对FAIR原则关注程度相对较高.

2.3.2 FAIR原则施引文献的国家间合作

我们选取了发文数量最高的10个国家作为研究对象, 绘制了各国间的合作力导图, 如图3(a)所示. 它展示了各国间合作的密切程度, 图中节点代表发文国家, 两个国家间有连线说明两国间合作发文数量大于等于100篇论文; 节点的大小代表发文数量, 节点间连线的粗细代表了各国合作发文的数量多少, 连线越粗表明两国间的合作发文量越大. 美国、德国和英国位于本领域研究的中心, 与多数国家都建立了密切合作, 且美国、德国、英国三国之间也有着大量合作.

高发文机构间的合作关系见图3(b). 机构间的合作关系

表 2 中国机构FAIR原则施引文献的高被引论文

Table 2 Highly cited papers in the FAIR principles citation literature in Chinese institutions

标题	期刊	期刊JCI指数	中国作者单位	发文年	被引频次	文献编号
Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence	Nature Medicine	1.26	广州医科大学[a)]	2019	221	[7]
Global observing needs in the deep ocean	Frontiers in Marine Science	1.05	中国科学院海洋研究所	2019	104	[8]
Toward a coordinated global observing system for seagrasses and marine macroalgae	Frontiers in Marine Science	1.05	中国科学院南海海洋研究所	2019	80	[9]
Citizen science and the United Nations Sustainable Development Goals	Nature Sustainability	3.88	香港"CitizenSci Asia"期刊	2019	195	[10]
The ProteomeXchange consortium in 2020: Enabling "Big Data" approaches in proteomics	Nucleic Acids Research	3.12	中国科学院北京生命科学	2020	318	[11]
BioModels-15 years of sharing computational models in life science	Nucleic Acids Research	3.12	中国科学院北京生命科学	2020	94	[12]
Crop phenomics and high-throughput phenotyping: Past decades, current challenges, and future perspectives	Molecular Plant	1.43	华中农业大学	2020	200	[13]
The future of sensitivity analysis: An essential discipline for systems modeling and policy support	Environmental Modelling & Software	1.04	河海大学	2021	92	[14]
An agenda for open science in communication	Journal of Communication	0.89	香港浸会大学	2021	64	[15]
Fungal taxonomy and sequence-based nomenclature	Nature Microbiology	5.52	南京农业大学、吉林农业大学、中国科学院云南植物园	2021	47	[16]
The International Soil Moisture Network: serving Earth system science for over a decade	Hydrology and Earth System Sciences	1.49	中国科学院地质与地球物理研究所、中国科学院大学、清华大学、武汉大学	2021	50	[17]

a) 通讯作者机构

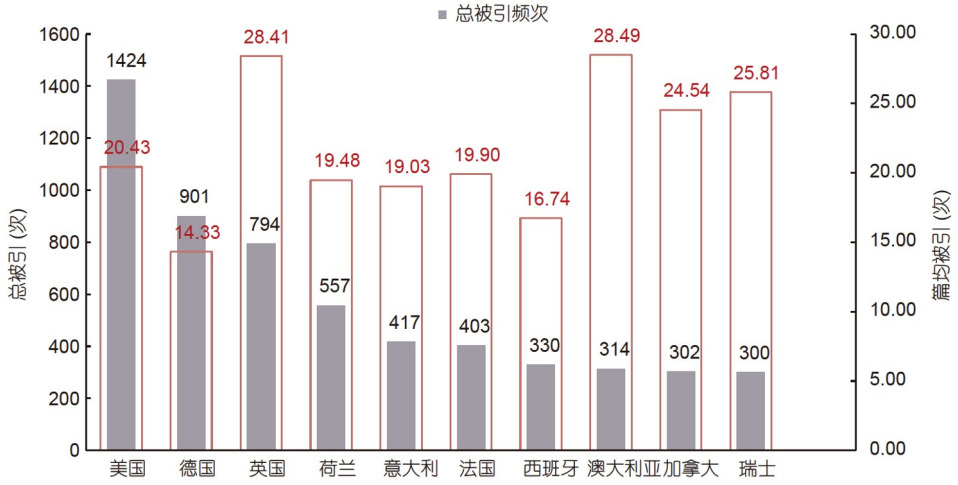


图 2 高发文国家论文引用统计
Figure 2 Citation statistics for papers from countries with high-volume publications

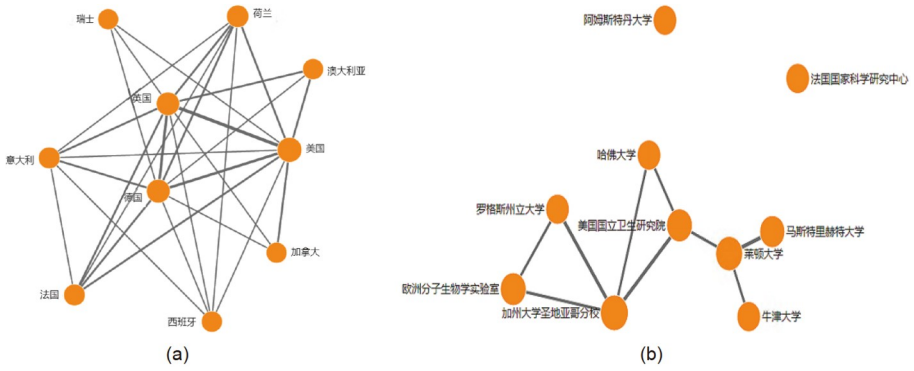


图 3 FAIR原则施引文献高发文国家间(a)和高发文机构间(b)的合作关系图
Figure 3 Collaborative relationship chart of the FAIR principles for citing literature between highly publishing countries (a) and between highly publishing institutions (b)

分析表明，加州大学圣地亚哥分校是本领域的领头羊，与美国国立卫生研究院、哈佛大学、罗格斯州立大学和欧洲分子生物学实验室有着密切合作. 其次是莱顿大学，其与美国国立卫生研究院、牛津大学、马斯特里赫特大学联系紧密. 相对于美国和荷兰，法国的研究结构相对独立，与其他单位联系较少.

2.3.3 FAIR原则研究的资助机构分析

我们遴选了资助发文大于10篇的基金开展分析. 主要资助基金包括美国国立卫生研究院、欧盟理事会、德国研究基金会、欧盟2020地平线研究与创新计划、美国国家自然科学基金、美国能源部、德国联邦教育和研究部、英国生物技术和生物科学研究理事会、德国DEAL计划、英国自然环境研究委员会(表3).

2.4 FAIR原则研究主题分析

基于关键词共现的方法，将本领域所有数据视为一个数据集，采用Thomson Data Analyzer软件将论文的关键词字段经过机器与人工清洗，选取出现频次大于5次的42个关键词作为分析对象. 之后利用VOSviewer软件，对论文核心主题词代表此主题中出现的高频主题词数据进行聚类，共得到4个簇(表S1). 结合专家判读，分别对每个聚类进行命名和解读，对本领域发文主题进行识别和分析.

分析结果显示，核心主题词平均被引频次代表包含此核心主题词的论文发文以来的平均被引频次；平均关联强度代表这个簇包含的核心主题词间联系的紧密程度，某个簇的平均关联强度越大，代表核心主题词间共现强度越大、研究越集中；反之，则代表共线强度相对较低、研究越分散. 核心主题词的总关联强度代表此主题词在共现网络中的重要程度，越高则说明此主题词对于构建网络越重要.

如图4所示，“基于FAIR标准的科学数据管理、监护和框架体系研究”主题中的关键词的平均关联强度最高，是研究内容最集中的主题，与其他研究内容交叉相对较少；“信息化

表 3　FAIR原则基金资助发文统计

Table 3　The FAIR principles funded publication statistics

资助机构	论文数	占全部发文百分比(%)
美国国立卫生研究院	47	14.11
欧盟理事会	32	9.61
德国研究基金会	23	6.91
欧盟2020地平线研究与创新计划	22	6.61
美国国家自然科学基金	19	5.71
美国能源部	16	4.80
德国联邦教育和研究部	15	4.50
英国生物技术和生物科学研究理事会	13	3.90
德国DEAL计划	11	3.30
英国自然环境研究委员会	8	2.40

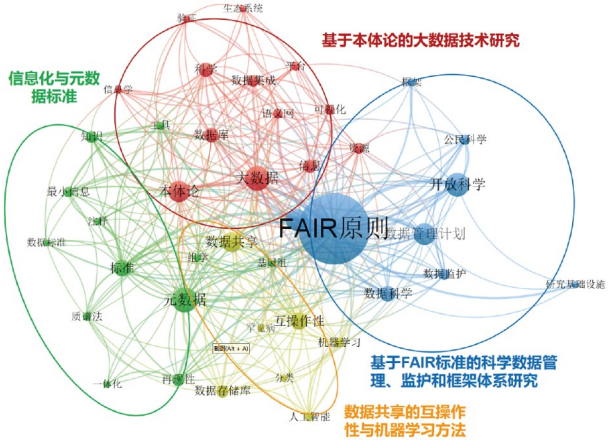


图 4　FAIR原则研究的主题分布

Figure 4　Distribution of topics studied by the FAIR principles

与元数据标准”的平均关联强度最低，是研究内容最发散的主题，与其他研究内容交叉相对较多；与“数据共享的互操作性与机器学习方法”主题相关的论文的平均被引频次最高，是4个主题中最受关注的主题. 可见，机器可读和互操作性是FAIR原则中的难点和关键.

3　FAIR原则活动影响分析

3.1　FAIR原则活动的总体情况

FAIR原则的讨论和推进带动着多个国际科学数据团体和相关机构的相关国际数据治理活动，如图5所示.

FAIR原则的影响可以从时间和空间两个角度进行观察. 在时间角度，FAIR原则早在2004年就已经被学界涉及，2014年形成共识，2016年FAIR原则正式论文发表后，迅速发展. 在空间尺度上，FAIR原则是跨学科的数据治理框架，因此在全球范围影响深远，多个活动都是由全球性的数据组织为主发起的. 相比而言，在国家和区域尺度的活动较少.

3.2　FAIR原则影响活动典型案例分析

进一步对图5所反映出来的国际组织和科技界对FAIR原则的响应，可以看出其在全球区域的影响过程. GO FAIR (https://www.go-fair.org/)是Global Open FAIR的缩写，旨在将利益相关者聚集在一起，为实施FAIR原则确定切实可行的具体方法，并形成3个活动支柱：GO CHANGE、GO TRAIN和GO BUILD. FAIR Metrics Group(http://fairmetrics.org)制定了评估FAIRness成熟度的14条通用指标，支持对FAIR原则的量化评估. RDA发布《FAIR数据成熟度模型：规范和指南》，旨在制定一套通用的FAIR核心评估标准. FAIRsFAIR(https://www.fairsfair.eu/)目标是在欧洲促进FAIR数据实践，并将提供一个平台，用于为欧洲研究数据提供存储和服务. World-FAIR：FAIR数据政策和实践的全球合作项目(https://codata.org/initiatives/decadal-programme2/worldfair/)是CODATA于2022年6月宣布启动的一个重大项目，其主要任务是推进FAIR原则的实施，特别是互操作性原则的落实，并制定一套跨学科领域的FAIR评估框架. 在区域应用层面，最值得一提的就是《将FAIR数据转化为现实：欧盟委员会公平数据专家组的最终报告和行动计划》[6]. 该报告提出了27条具体实施行动建议，见表S2.

值得一提的是，FAIR原则不能解决数据治理中的所有问题，尤其是数据保护方面的问题. CARE原则[18]的适时提出是一个从数据保护角度的重要治理理念. CARE原则具体指集体利益(collective benefit)、控制权(authority to control)、责任(responsibility)和道德(ethics)，是由全球土著数据联盟(Global Indigenous Data Alliance, GIDA)于2019年制定的. 它的提出主要是对FAIR原则和其他开放数据活动在伦理方面进行补充，强调对土著数据和用户要“公平且关怀”. FAIR原则和CARE原则的结合对更科学合理地践行数据治理具有重要的现实意义. 许多组织和机构正在通过合作、政策和机制采用CARE原则，包括研究数据联盟(RDA, www.rd-alliance.org)、史密森尼学会(Smithsonian, www.si.edu)、开放数据宪章(Open Data Charter, ODC, https://opendatacharter.net/)等[18].

4　我国应用FAIR原则的启示与建议

结合以上分析数据治理实践，可以发现，FAIR原则以其简洁、透明的治理规则迅速影响到全球和多个学科领域. 然而数据的治理更多依赖于理论与实践的结合，FAIR原则未在具体实践环节给出指导意见. 因此，有必要进一步思考如何通过实践来落实这些原则. 结合一些实践和专家研讨，提出以下在实施层面的启示和建议.

(1) 丰富FAIR数据治理在实践方面的领域特色. FAIR原则的提出引发多个学科领域的响应，例如涉及生命科学、材

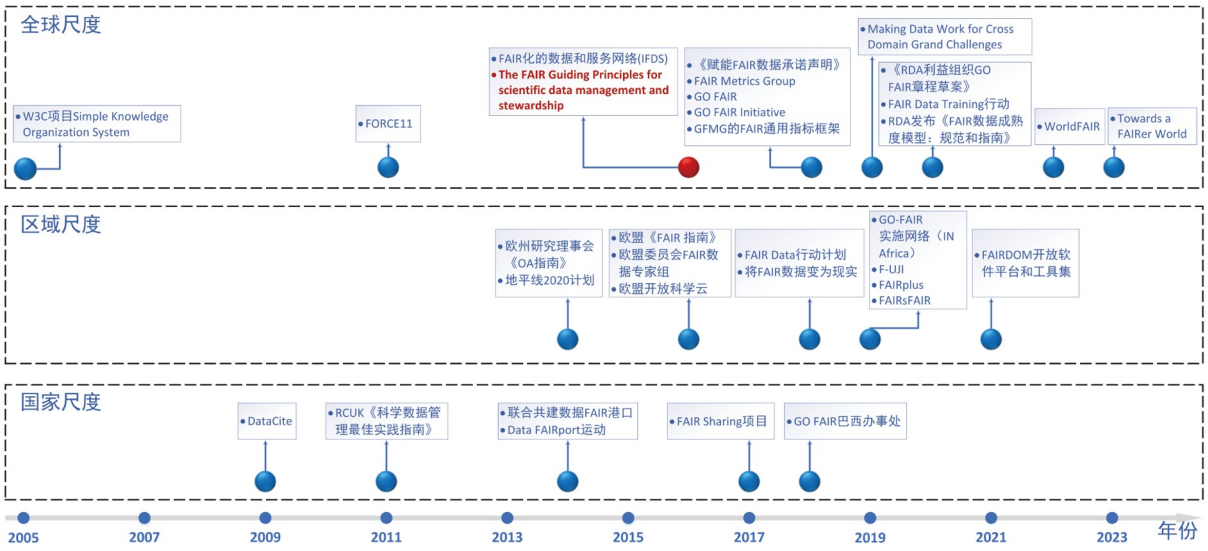


图 5 FAIR原则对相关国际数据治理的活动影响
Figure 5 Impact of the FAIR principles on relevant international data governance activities

料科学、地球科学、空间科学等领域，反映出科学数据治理在相关数据密集型领域的紧迫性. 充分利用好FAIR这一共性原则，并结合本领域和本基础设施特点，开展全面的数据治理实施设计将是各领域普遍面临的任务. 然而，学科领域在数据科学的共性认识和“学科信息学”的差异性认识上是共存的[19]. 在不同学科领域数据、信息、知识加工演化过程中，必然会不断丰富FAIR原则在实施中的内涵，同时形成X-FAIR的不同学科领域实践特色.

(2) FAIR原则是支撑科学范式变革的重要环节. FAIR是数据治理层面的典范，这是支撑以数据密集型研究为核心的科学范式变革的重要组成部分. 在整个数据驱动生命周期中，前期的数据处理工作，有大量需要新的人工智能技术支持下的数据提取. 在数据治理阶段，借助FAIR原则可最大化地为机器可读的人工智能应用创造条件，这也正是有些学者所说的“FAIR, For AI Ready”. 在科学发现阶段，通过FAIR原则治理的数据将直接为类似大模型的人工智能软件所访问和使用，从而产生更多从无到有的新知识.

(3) 机器可读是FAIR原则的关键和难点. 随着数据量的增加和各个学科领域大量长尾数据的出现(如海量的文献和报告数据)，迫切需要机器可读的方式获取、加工、访问和利用数据. 对于结构化的元数据信息，实现机器可读是相对可行的，然而针对复杂的数据实体，实现机器可读是非常困难的. 这需要更多基于本体或知识图谱的语义互通，而这正是前文发现的FAIR在领域的研究热点之一.

(4) 形成可互联互通的全球数据网络. Findable、Accessible、Interoperable、Reusable是全球各独立数据存储互联互通的基础. 在相近或交叉的学科领域中，都存在着数据库、数据仓储、数据中心等共存的现象，充分建立这些基础设施之间的互联互通，是对FAIR原则的深化，也是提高数据治理效益的体现. 因此，建议全面促成局部或整体性的数据网络，充分形成更开放、透明的全球公共数据基础设施.

(5) 元数据仍然是FAIR原则治理的重点. 元数据的结构化处理尽管在技术上可行，然而纷繁的元数据标准、繁琐的元数据著录、多维的元数据表达需求等均反映出这一领域尚未形成完全成熟的FAIR治理应用. 在很长一段时间内，元数据的FAIR治理仍然是各领域的重要任务，也是形成全球多学科领域数据网络互联互通的基础.

(6) 增强FAIR原则的整体软件架构. 当前FAIR原则的研究热点仍然多在计算机科学，这也反映出FAIR在信息技术开发方面的问题较多. 这包括多个层次的挑战. 首先是计算机架构上的，早期的分布式模式与当前业界推崇的存算一体架构并存，如何建立既最大程度共享利用基础设施硬件资源，又最大程度自治管理，且满足数据权益和隐私保护等要求，是一个重大挑战. 即，如何在一个公共平台下既共享数据，又可以借助区域链、隐私计算、联邦计算等解决数据的保护问题.

(7) 加强FAIR原则应用工具研发. 无论是FAIR数据治理，还是FAIR治理效果的评价，均需要多种多样的应用工具支持. 当前，国际和国内均在推进许多FAIR治理软件的开发，如中国科学院的FAIR软栈[20]. 然而总体上还是缺少可用的工具，尤其是在开放标准协议、语义层互通、异构系统访问、多源数据集成汇聚等方面存在挑战.

(8) 促进更多FAIR原则的应用案例. 目前，FAIR原则已在美国、欧洲、澳大利亚等区域得到快速发展推广，尤其是在欧盟地区实施力度最大，但在其他区域的实践还不充分. 借

助我国国家科学数据中心和多个领域或部门科学数据中心体系建设的契机[21], 可进一步推动FAIR原则在不同领域数据中心治理应用的最佳实践. 例如, 推动FAIR在地球表层、深时地球、海洋、对地观测、气候变化、生命科学、医疗健康、核能、材料等领域的深化应用.

(9) 健全FAIR原则的标准生态. FAIR原则能够被科学共同体广泛接受和认可, 与其简明扼要的特点息息相关. 围绕这一原则的实施和应用, 还需要细化若干标准和规则, 包括建立围绕FAIR原则的标准体系, 形成FAIR原则各条款实施的规范, 形成可评价的标准等. 制定和发布适用不同场景的国家标准、行业标准、团体标准、企业标准等, 是共同建立FAIR标准生态的一个发展需求.

(10) 创新发展科学数据治理的新理念. FAIR原则是2016年提出的, 在此之后, 许多学者相继提出了TRUST原则[22]、CARE原则[18]、PARIS原则[23]等. 应对于大数据和人工智能时代的更多科学数据治理需求, 仍然需要继承和发展各类理论与方法, 并且建立适合于本地国情和学科特点的更多数据治理新理念.

致谢 感谢参与讨论的各位专家和参与调研的团队成员; 感谢深时数字地球(Deep-time Digital Earth, DDE)国际大科学计划的支持; 感谢国家重点研发计划(2022YFF0711600)资助.

推荐阅读文献

1 Wang J L, Wang M M, Shi L, et al. The situation of scientific data management and its enlightenment to Earth sciences of China (in Chinese). Adv Earth Sci, 2019, 34: 306–315 [王卷乐, 王明明, 石蕾, 等. 科学数据管理态势及其对我国地球科学领域的启示. 地球科学进展, 2019, 34: 306–315]
2 Leng S Y, Su T Y, Zhang L. Practice and reflections on open sharing of marine scientific data (in Chinese). Chin Sci Bull, 2023, 68: 2239–2249 [冷疏影, 苏天赟, 张亮. 海洋科学数据开放共享实践与思考. 科学通报, 2023, 68: 2239–2249]
3 Wang C, Hazen R M, Cheng Q, et al. The Deep-Time Digital Earth program: Data-driven discovery in geosciences. Natl Sci Rev, 2021, 8: nwab027
4 Qu J S, Huang K M. The development logic and future mission of open science (in Chinese). Chin Sci Bull, 2022, 67: 4312–4325 [曲建升, 黄珂敏. 开放科学的发展逻辑与未来使命. 科学通报, 2022, 67: 4312–4325]
5 Wilkinson M D, Dumontier M, Aalbersberg I J, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data, 2016, 3: 1–9
6 European Commission Expert Group on FAIR Data. Turning FAIR into Reality. Luxembourg: Publications Office of the European Union, 2018
7 Liang H, Tsui B Y, Ni H, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med, 2019, 25: 433–438
8 Levin L A, Bett B J, Gates A R, et al. Global observing needs in the deep ocean. Front Mar Sci, 2019, 6: 241
9 Duffy J E, Benedetti-Cecchi L, Trinanes J, et al. Toward a coordinated global observing system for seagrasses and marine macroalgae. Front Mar Sci, 2019, 6: 317
10 Fritz S, See L, Carlson T, et al. Citizen science and the United Nations Sustainable Development Goals. Nat Sustain, 2019, 2: 922–930
11 Deutsch E W, Bandeira N, Sharma V, et al. The ProteomeXchange consortium in 2020: Enabling “Big Data” approaches in proteomics. Nucleic Acids Res, 2020, 48: D1145–D1152
12 Malik-Sheriff R S, Glont M, Nguyen T V N, et al. BioModels—15 Years of sharing computational models in life science. Nucleic Acids Res, 2020, 48: D407–D415
13 Yang W, Feng H, Zhang X, et al. Crop phenomics and high-throughput phenotyping: Past decades, current challenges, and future perspectives. Mol Plant, 2020, 13: 187–214
14 Razavi S, Jakeman A, Saltelli A, et al. The future of sensitivity analysis: An essential discipline for systems modeling and policy support. Environ Mod Softw, 2021, 137: 104954
15 Dienlin T, Johannes N, Bowman N D, et al. An agenda for open science in communication. J Commun, 2021, 71: 1–26
16 Lücking R, Aime M C, Robbertse B, et al. Fungal taxonomy and sequence-based nomenclature. Nat Microbiol, 2021, 6: 540–548
17 Dorigo W, Himmelbauer I, Aberer D, et al. The International Soil Moisture Network: Serving Earth system science for over a decade. Hydrol Earth Syst Sci, 2021, 25: 5749–5804
18 Carroll S R, Garba I, Figueroa-Rodríguez O L, et al. The CARE Principles for indigenous data governance. Data Sci J, 2020, 19: 43
19 Zhang Z Q. Research on the Theory and Application of Subject Informatics for Domain Knowledge Discovery (in Chinese). Beijing: Science Press, 2023 [张自强. 面向领域知识发现的学科信息学理论与应用研究. 北京: 科学出版社, 2023]
20 Shen Z H, Zhao Z H, Wang H B. Big Data technology stack shifting: From SQL centric to graph centric (in Chinese). Data Anal Knowl Discov, 2020, 4: 50–65 [沈志宏, 赵子豪, 王海波. 以图为中心的新型大数据技术栈研究. 数据分析与知识发现, 2020, 4: 50–65]

21 Wang R D, Gao M X, Shi L, et al. Research and thoughts on the opening and sharing of scientific data under background of Big Data (in Chinese). China Sci Technol Resour Rev, 2020, 52: 1–5 [王瑞丹, 高孟绪, 石蕾, 等. 对大数据背景下科学数据开放共享的研究与思考. 中国科技资源导刊, 2020, 52: 1–5]
22 Lin D, Crabtree J, Dillo I, et al. The TRUST Principles for digital repositories. Sci Data, 2020, 7: 144
23 Shen Z H, Zhang X L, Zheng X H. From FAIR to PARIS: Improving the usability of scientific data in the open collaborative environment (in Chinese). Big Data, 2023, 9: 172–188 [沈志宏, 张晓林, 郑晓欢. PARIS原则: 开放协作环境下科学数据的可用性. 大数据, 2023, 9: 172–188]

补充材料

图S1 首篇FAIR原则论文施引文献发文趋势
表S1 FAIR研究主题分析
表S2 《将FAIR变为现实》提出的27条行动建议

本文以上补充材料见网络版csb.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

Summary for “FAIR原则全球影响分析及其实施策略建议”

Global impact analysis and implementation strategy recommendations of the FAIR principles

Juanle Wang[1,5*], Yang Li[2], Shuqiang Wang[3], Lei Shi[4] & Bowen Duan[1]

[1] *State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China;*
[2] *National Science Library, Chinese Academy of Sciences, Beijing 100190, China;*
[3] *Library of the Institute of Geographic Sciences and Human Resources, Chinese Academy of Sciences, Beijing 100101, China;*
[4] *National Science and Technology Infrastructure, Beijing 100862, China;*
[5] *Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China*
* Corresponding author, E-mail: wangjl@igsnrr.ac.cn

In the current era of Big Data and artificial intelligence, data sharing and its governance are increasing in importance. However, guiding scientific data governance is a major challenge in multidisciplinary situations. In this context, the FAIR (Findable, Accessible, Interoperable and Reusable) principles were proposed in a paper in 2016, “The FAIR Guiding Principles for scientific data management and stewardship”. Subsequently, the FAIR principles quickly spread across multiple domains. How many research areas and countries/regions have been impacted by this academic process? How did the dissemination process proceed? What are the direct effects of scientific data governance practices? To address these questions, we analyzed bibliometric methods and scientific data-sharing practices. The paper that introduced the FAIR principles (referred to as “cited paper”) was used to construct a retrieval type. Academic papers citing FAIR principles research, from January 1, 2016 to December 31, 2022, were also retrieved from the Web of Science Database. This study identified more than 300 articles on FAIR principles and 3300 papers citing the cited paper. Furthermore, we conducted a bibliometric analysis that included citation frequency, discipline-wise distribution of citations, country/region-wise distribution of citations, keyword frequency count, and keyword theme clustering. A total of 137 countries were involved in FAIR impact studies, and the countries with the highest number of publications were the United States, Germany, and the United Kingdom. They are core research countries, who have established close cooperation with most countries involved in this field. The impact mainly involves basic and cutting-edge fields, such as biology, life, materials, and medicine, with the largest number of disciplines reflected in multiple disciplines, such as computer science, biology, medicine, and environmental science. The relevant international data governance cooperation dynamics after the proposal of the FAIR principles were further analyzed from both temporal and spatial perspectives. The FAIR principles were found to have a greater impact on a global scale, mainly in related international organizations. An overall bibliometric and activity impact analysis showed that the FAIR principles had rapidly expanded globally and across multiple disciplines with its concise and transparent governance rules. Combining the above analysis with the actual needs of scientific data governance in China, we put forward ten insights and suggestions. (1) Enriching and developing the field characteristics of FAIR data governance in practice; (2) the FAIR principles are an important means to support science paradigm shifting; (3) machine-readability is the key point of the FAIR principles; (4) forming an interconnectable global data network; (5) metadata is still the main content for FAIR principles governance; (6) enhancing the overall software architecture of the FAIR principles implementation; (7) strengthening the research and development of the FAIR principles application tools; (8) promoting a wider application of the FAIR principles in different fields; (9) improving the standard ecology of the FAIR principles; (10) innovating and developing new concepts of scientific data governance beyond FAIR principles.

scientific data, FAIR principles, data governance, data sharing, data interoperability, suggestions

doi: 10.1360/TB-2023-0585