【资源·共享】

FAIR原则背景下基于机构知识库的高校科学数据管理平台研究*

●孙清玉　梁美宏　张友华　（河海大学　南京　210098）

[摘　要]科学数据不仅是科学研究的成果，也是科学研究的对象与工具，是科研机构、科研人员研究交流的重要资源。科学数据管理是对科学数据进行搜集、整理、保存、共享，促进科学数据的价值再提升。随着FAIR原则研究的不断深入，越来越多的机构接受并支持该原则成为科学数据管理的国际准则。文章以河海大学为例，从系统架构、质量控制、服务体系等方面探究高校机构知识库在科学数据管理方面的应用以及实践科学数据的可发现、可访问、可互操作、可重用等原则，旨在为机构知识库在FAIR原则背景下实现科学数据的有效管理提供对策和参考。

[关键词]科学数据　机构知识库　FAIR原则　管理平台

[中图法分类号]G253　　[文献标识码]A　　[文章编号]1003－7845(2022)01－0037－04

[引用本文格式]孙清玉，梁美宏，张友华. FAIR原则背景下基于机构知识库的高校科学数据管理平台研究[J]. 高校图书馆工作，2022(1)：37－40

科学数据是科研工作开展的重要基础资源，也是科研人员的重要学术产物。有效管理科学数据，可以在一定程度上降低科学研究的重复性成本，提高科研产出效益，推动科学进步，收获更多生产力。目前，提供科学数据管理已成为高校图书馆的重要发展方向和服务趋势之一。2017年《地平线报告：图书馆版》指出，随着开放出版和数据收集的增加，图书馆在科学数据管理中的作用进一步巩固[1]。科学数据管理需要以数据存储库为平台，实现数据的有效管理、公开共享、规范引用和出版传播。机构知识库作为重要的数据管理平台，在数字资源存储与管理方面有着重要作用[2]。许多高校、研究机构已经建设机构知识库用于本单位研究出版物的存储，因此技术基础设施可以在无须开发或购买全新软件平台的情况下进行功能扩展。在国家出台一系列有关数据管理、数据共享的纲要和办法的大环境下[3－4]，机构知识库除了承担成果存储中心的功能外，也应成为科学数据管理中心。

2016年FORCE11组织(The future of research communication and e－scholarship)正式提出在科学数据管理领域引入FAIR原则。同年，FAIR指导性原则正式稿发表在《科学数据》第3期上[5]，FAIR原则包含可发现(Findable)、可访问(Accessible)、可互操作(Interoperable)及可重用(Reusable)四项内容，明确了科学数据管理的目标。河海大学图书馆基于机构知识库构建科学数据管理平台，在平台架构设计中实践FAIR原则，具备采集、标引、分类、保存、检索科学数据的组织功能，具备实现原始数据、中间数据、结果数据的利用和数据挖掘服务功能。

1　基于机构知识库的科学数据管理平台构建

科学数据管理的具体内容包括数据创建、数据存储、数据检索、数据安全、数据保存、数据共享和数据再利用等方面。

1.1　平台目标与功能

河海大学机构知识库是以本校学者公开产出的各种文献数据为主构建的集元数据采集、存储、清洗、展示和自主维护于一体的数据管理平台。目前存储的资源主要包括已公开发表或已申请的学术论文、图书著作、学位论文、会议文献、专利等。此外，为突出机构知识库科研成果全面、数据类型丰富的特色，图书馆在机构知识库建设之初，就为平台设置了广泛的成果类型，除了上述已有的文献类型外，还包括报纸、标准、研究报告、科学数据集、课件、教学视频、系统软件、演讲稿、实验报告、设计图纸、工作

*　本文系江苏省图书馆学会课题“区块链技术在机构知识库数据共享中的应用研究”(20YB03)、江苏省社科应用研究精品工程课题“基于机构知识库的高校科学数据管理模式研究”(20SYC－228)和河海大学图书馆科研项目(TSG2021A02)研究成果之一。

文稿等,这些都为科学数据管理提供了基础保障。

基于机构知识库构建的科学数据管理平台的目标是实现科学数据的开放、共享和引用,推动科学数据的长期保存与数据资产管理[6]。学者可以随时向平台添加相关数据及文献,平台可以为用户提供检索和浏览功能,并通过权限设置实现数据共享。

1.2　平台架构设计

基于机构知识库构建的科学数据管理平台从逻辑框架结构角度分为数据底层、管理层和服务层,如图 1 所示。其中数据底层为科学数据资源池;管理层用于管理各种元数据及各级机构、学者、科研成果的对照关系;服务层展示各种成果数据,并提供数据分析及共享服务。

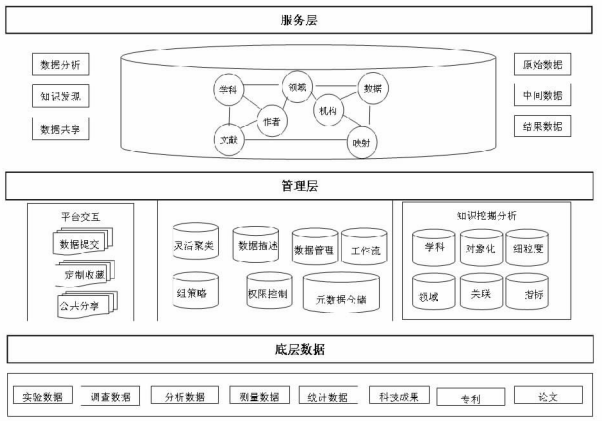

图 1　基于机构知识库的科学数据管理平台

(1)科学数据提交及存储模块。科学数据提交及存储模块基于底层数据而设计。FAIR 原则中数据可发现(Findable)子原则意在指出科学数据共享的前提是数据能够被用户及时发现,可发现原则为后续数据访问、操作和重用提供了条件[7]。可发现原则要求数据拥有一个唯一并且能永久存在的标识符(DOI),但是科学数据具有类型繁杂、量大且变化快、分布式异构多源等特征,较难进行准确描述。元数据作为数据的数据,可以有效地解决这一问题,能够对数据资源的内容、质量、条件、特性等相关描述性信息进行充分描述。

科学数据管理平台可依据元数据规则设置精准、完整、丰富的描述字段及字段内容要求,包括数据产生背景、样本描述、采集流程、质量评估方法、使用规则等,从而规范数据生产者、发布者上传数据资源的流程,避免科学数据因来源、格式等不同造成无序管理,无法利用。由于元数据的描述完整丰富且具有可扩展性,即使数据缺少唯一标识符,也可以在平台中被著录或标引,用户可以通过浏览、检索等方式方便地获得平台公开的数据资源。底层数据可存储本校科研人员的实验数据、调查数据、分析数据、测量数据、统计数据等数据资源,并可对科技成果、专利、论文等文献的科学数据进行抽取和存储。具体提交、存储的流程为:科研人员通过平台设定的元数据要求,建立新的数据项目,上传数据,或由平台对科学数据进行抽取和存储。数据上传之后进入到数据池中;相关管理员访问数据池,对上传数据进行校核与审查。图书馆数据馆员根据可读性、完整性等原则检查数据、元数据和文档,最终实现科学数据的顺利提交与有效存储。

(2)科学数据访问及检索模块。FAIR 原则中数据的可访问(Accessible)原则是指用户可以通过检索功能找到科学数据。对于科学数据管理服务平台,数据的访问和使用是最重要的目标之一,可确保数据无论位于何处,都能被查找和获取。科学数据访问及检索模块提供简单检索、高级检索、二次检索等多种检索方式,从而实现科学数据的可访问(Accessible)。该模块的检索界面、结果展示等设置均基于文献检索系统,检索字段包括题名、作者、机构、来源、出版日期、关键词、摘要等,检索结果以文本、表格、XML 等形式展示。

科学数据类型多种多样,而不同类型数据的格式不同,对应的检索字段也不同。为满足不同类型数据的存储、检索、展示与利用,基于机构知识库的科学数据管理平台将每个元数据项作为独立一行,并且不展示空元数据项。该方法能够灵活便捷地展示复杂多样的数据[8]。

基于机构知识库构建的高校科学数据管理平台制定了访问数据资源的协议规则,包括访问入口、身份验证、访问权限等方面。根据科学数据来源及内容,可将访问级别分为:管理员级别,这一级别的使用者为图书馆数据馆员,对发布的数据有认领、审核、修订、编辑等权限;数据拥有者或数据发布者级别,可以撤回、修订、编辑数据,可免费下载使用相关专业数据;数据使用者级别,这一级别的用户可以检索、浏览数据,但在下载数据时,需向管理员提交下载申请,经审核通过后,用户可下载获得数据[9]。

(3)科学数据共享及分析模块。FAIR 原则中数据的可互操作(Interoperable)原则的总要求是通过使用标准定义、通用数据元素等描述数据,实现数据间的互操作。可重用(Reusable)原则指的是数据与数据集应有明确的使用许可,同时包含准确的数据源信息。科学数据在具备互操作性的基础上,可

以实现不同类型数据的组织加工、分析处理、共享利用等功能。科学数据的开放共享可以实现研究人员引用或重现实验,有助于避免不必要的重复实验操作,缩短研究周期,加快整个领域的研究进程。所以科学数据唯有通过充分而广泛的共享,才能最大程度地发挥价值,实现整体增值,减少重复投入。

在科学数据共享管理中,数据发布者或数据管理员可以选择项目组内分享、二级单位内分享、二级单位间分享、校内分享及校外分享等,并可对分享的资源进行互操作,例如,评分、点赞、推荐等,系统可以根据全部历史用户的评价结果,对共享学术资源进行质量评定并进行排名等操作[10]。

2　科学数据管理平台服务体系研究

2.1　科学数据发现与数字化保存

河海大学图书馆基于机构知识库构建科学数据管理平台,通过设置管理元数据、规范描述数据等过程,实现了科学数据的结构化、流程化、数字化保存,能够有效避免数据的丢失、无序等问题,确保数据准确、完整、可复用;基于数据的结构化保存,设置相应的检索字段与检索浏览功能,实现了科学数据如同图书、期刊、专利等结构化数据一样被检索与发现,从而得以有效利用,提高数据价值。同时,基于机构知识库的科学数据管理平台通过集成 DataCite,进一步促进数据被发现与被引用。

2.2　数据参考咨询

数据参考咨询服务是针对用户在遇到特定的数据管理相关问题时,图书馆所提供的决策支持、定制解决方案等人工服务,其目的是为用户提供个性化的数据管理服务。例如,当用户在科学数据管理平台中提交上传数据时,相关数据馆员会及时收到该项操作的提示信息,此时数据馆员可直接与用户联系,帮助其解决在上传、发布、管理数据中遇到的问题,同时了解用户及其研究团队的相关研究,发掘与其开展进一步合作的机会,例如,可以合作开展数据密集型研究[11]。用户也可通过平台联系到相关馆员或专家,协助其解决在数据管理过程中所遇到的各种问题,还可以根据自身需求,提出定制化、个性化的数据管理支持服务。

河海大学图书馆将基于机构知识库构建的科学数据管理平台纳入到参考咨询服务框架中。基于该平台,图书馆可为用户提供馆员咨询、技术专家咨询等多种咨询渠道,以满足其不同层面的科学数据需求。其中数据馆员可以利用自身专业知识和业务技能帮助用户有效检索、发现、利用科学数据平台中已有的数据资源及相关服务;技术专家可以为用户提供科学数据平台中关于数据访问、元数据创建等方面的技术知识和相关技能。

2.3　数据素养培训

基于已构建的科学数据管理平台,河海大学图书馆通过线上、线下的不同方式,为不同层面的对象定期开展数据素养培训服务。培训服务的具体课程有数据素养课程、数据管理课程、实践操作课程等。

针对馆员的数据素养培训内容主要包括以下两个方面:一是提高馆员的数据管理服务意识。馆员作为科学数据管理平台的管理者、服务者,需注重自身对于数据管理的内在意识,积极主动了解用户需求,并提供相关服务。二是强化馆员的数据管理能力。馆员应结合本学科专业知识,将本学科相关科学数据纳入馆藏、教学和咨询工作,了解学科专业知识,不断学习开展数据管理的先进技术、手段,创新服务内容。

针对用户的数据素养培训内容主要包括以下三个方面:第一,培养用户,尤其是科研人员的数据管理意识,使其了解国家、相关机构对于科学数据管理的政策要求,充分认识到科学数据对于学科发展、科研工作、履行义务等方面的重要性,同时提高其对于数据所有权、隐私权、知识产权的保护意识,以及数据开放获取的共享意识。第二,提高用户数据管理操作技能,包括科学数据的元数据描述方法、上传要求、发现方法、检索策略等方面的知识与技能,帮助用户了解科学数据提交、获取、使用、评价等方面的数据素养能力。第三,提高用户数据管理能力,例如,通过一小时讲座的形式开展数据分析、数据管理、数据可视化等各类型数据分析软件的培训课程,面向不同需求和不同层次的用户开展针对性、个性化、持续性的数据素养讲座、培训,促进用户更好地进行科研数据管理工作。

3　思考和建议

FAIR 原则对于数据管理平台规范数据管理流程具有重要意义。一方面,在围绕科学数据的产生、管理和发布的各个环节,明确各方的责任与义务,建立符合 FAIR 原则的规范、流程、评价标准,并不断建设支撑这些管理措施实施的技术环境;另一方面,通过 FAIR 原则的实施,建设可重用的科学数据,实现科学数据价值的最大化。

河海大学图书馆在 FAIR 原则指导下,基于机构知识库构建科学数据管理平台,基于都柏林元数据等标准收集、组织、存储数据,并在平台中集成

ORCID、DataCite 等数据共享工具,为河海大学科研人员提供了一个开展数据管理的专业平台,该平台不仅满足用户对科学数据的提交、发布、存储和检索等需求,同时还支持用户进行在线合作研究。在协助河海大学科研人员申请项目、协作科研、开展数据管理等方面发挥了重要作用,并且已在用户群体中产生了一定影响。

但是,基于机构知识库构建的科学数据管理平台在元数据支持、用户体验等方面尚存在一些问题,有待今后不断改进完善。例如,平台现有的元数据标准主要采用的是 DCMI 基础元数据,不足以支持影音、地图等特殊类型的数据描述,后续可结合本机构科学数据的具体特征、用户需求等综合情况进一步完善。

参 考 文 献

[1] Adams B S, Cummins M, Davis A, et al. NMC horizon report: 2017 library edition[R]. Austin, Texas: The New Media Consortium.

[2] 刘 莉. 基于机构知识库的科研数据管理分析[D]. 淄博:山东理工大学,2020:1-2.

[3] 国务院办公厅. 关于印发促进大数据发展行动纲要的通知[(EB/OL)]. [2021-04-12]. http://www.gov.cn/zhengce/content/2015-09/05/content_10137.htm.

[4] 国务院办公厅. 关于印发科学数据管理办法的通知[(EB/OL)]. [2021-04-12]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.

[5] 张玉娥,王永珍. 欧盟科研数据管理与开放获取政策及其启示——以"欧盟地平线2020线计划为例[J]. 图书情报工作,2017(13):70-76.

[6] 陈秀娟,吴 鸣. 面向学科领域科研数据发表的图书馆服务——以化学学科为例[J]. 图书馆论坛,2017(11):66-78.

[7] 宋 佳,温亮明,李 洋. 科学数据共享FAIR原则:背景、内容及实践[J]. 情报资料工作,2021(1):57-68.

[8] 洪正国,项 英. 基于Dspace构建高校科学数据管理平台——以蝎物种与毒素数据库为例[J]. 图书情报工作,2013(6):39-42.

[9] 王丹丹,任婧媛,吴思洁. 社会科学数据管理与服务平台研究——德国的经验[J]. 现代情报,2020(11):99-106.

[10] 孙清玉,梁美宏,胡晓辉. 区块链技术在机构知识库数据共享中的应用研究[J]. 新世纪图书馆,2020(7):49-52.

[11] 王 辉,Michael Witt,窦天芳. 普渡大学研究仓储及其支持的科学数据管理服务[J]. 现代图书情报技术,2015(1):9-16.

[作者简介]孙清玉,河海大学图书馆副研究馆员;梁美宏,河海大学图书馆馆员;张友华,河海大学图书馆馆员。

[收稿日期]2021-07-25 (宋小华 编发)

The Scientific Data Management Platform Based on the Institutional Repository under the FAIR Principle

Sun Qingyu Liang Meihong Zhang Youhua

(Hohai University, Nanjing, Jiangsu 210098, China)

Abstract Scientific data are not only the results of scientific research, but also the objects and tools. Scientific data are important resources for scientific research and communication among research institutions and researchers. Scientific data management refers to the collection, organization, preservation and sharing of scientific data, which can promote their value. With further study of the FAIR principle, more and more institutions accept it as the international norm for scientific data management. This paper, taking Hohai University as a case in point, examines the system construction, quality control and service system of the application of the institutional repository to scientific data management, and puts the principles of findability, accessibility, interoperability and reusability into practice. It aims to propose some countermeasures for the effective management of scientific research data in the institutional repository under the FAIR principle.

Keywords Scientific data. Institutional repository. FAIR principle. Management platform.