

开放科学环境下科学数据与科技文献 关联模式研究^{*}

支凤稳 赵梦凡 彭兆祺
(河北大学管理学院, 保定 071002)

摘要: 数据密集型科研范式下, 科学数据成为与科技文献同等重要的科研资源。开放科学的迅速发展为科研资源管理带来机遇和挑战, 探究开放科学环境下科学数据与科技文献关联模式对提高资源的利用效率具有重要意义, 同时也可对相关理论研究提供有益参考。分析开放科学环境下科学数据与科技文献关联面临的新形势, 构建科学数据与科技文献的关联模式, 其基本流程包括数据汇交、质量审核、关联建立、关联应用、成果反馈。提出促进科学数据与科技文献关联的对策建议: 完善相关政策体系、丰富平台功能、增加关联文献类型、加强相关主体合作。

关键词: 开放科学; 科学数据; 科技文献; 关联模式

中图分类号: G250 **DOI:** 10.3772/j.issn.1673-2286.2023.10.006

引文格式: 支凤稳, 赵梦凡, 彭兆祺. 开放科学环境下科学数据与科技文献关联模式研究[J]. 数字图书馆论坛, 2023 (10): 52-61.

大数据时代的来临掀起了开放科学运动的浪潮。2021年11月, 联合国教科文组织 (UNESCO) 审议通过《开放科学建议书》, 标志着开放科学迈入全球共识新阶段^[1]。如今, 科学研究正在向自由平等和成果共享的开放科学模式前进, 开放科学的迅速发展也为科研资源的管理与建设提供了新思路、新方法、新路径。作为同等重要的科技资源, 科学数据与科技文献共同推动科研创新和知识发现。在过去一段时间内, 科技文献因具有数量多、增速快、可获取性强、易于传播等特点, 在科研活动中占据主导地位, 成为科研人员参考的首要资源。在数据密集型科研范式下, 科研成果的论证往往需要大量数据的支持^[2], 且随着学科的不断交叉与融合, 科研人员时常需要跨学科数据的支持以完成复杂的科研任务。然而, 目前大量的科学数据和科技文献孤

立存在, 具有分散、异构、多源等特征, 无法体现学科知识的内在联系, 从而为科研人员跨学科获取和利用数据带来了困难。建立科学数据与科技文献之间的关联是解决这一问题的有效途径, 不仅可以促进科研资源的精准发现, 提高资源的利用效率, 还有助于推动科研合作和学术创新, 为科研活动的顺利开展提供有力支撑。

科学数据与科技文献的关联不仅指两者之间客观存在的连接关系, 即论文需要支撑数据, 还包括数据引用、数据复用、数据出版等活动引发的间接关系。因此, 关联的数据对象既包括作者提交至数据仓储进行存储的自身数据, 也包括引用的他人数据^[3]。当前, 国内外数据平台、学术期刊、出版商、科研机构等相关主体逐渐意识到科学数据与科技文献关联的重要性, 陆

收稿日期: 2023-08-14

^{*}本研究得到河北省高等学校人文社会科学研究项目“元宇宙时代科学数据共享模式及其应用研究”(编号: BJS2022027)、河北大学校长基金资助项目“基于科学数据关联的科研合作网络研究”(编号: 2022HXZ009)、河北大学研究生创新能力培养资助项目“政策工具视角下京津冀智库人才政策演化研究”(编号: HBU2024SS023)资助。

续参与相关的学术研究和服务实践,并出台了相应的关联管理政策。国内外学者也针对该主题展开了系列探索,主要从元数据关联、引用关联和语义关联等方法层面分析其可行性,或从实践与应用层面探究关联的实现方案与策略。例如:在元数据关联方面,黄筱瑾^[4]基于元数据分析科学数据与科技文献关联的可行性与优势,又进一步从元数据描述中提取出内容特征^[5],并由此建立两者的关联;也有学者对全球变化科学研究数据出版系统中的数据论文及相应的期刊论文元数据关联的特征进行分析^[6]。在引用关联方面,张鑫等^[7]设计了基于引文探针的关联推断算法,从而提升文献与数据关联的覆盖率;Riedel等^[8]利用文本挖掘算法,提取出生物医学领域文献引用的数据集。在语义关联方面,丁培^[9]通过Uni Prot案例验证利用语义标注技术实现科学数据与文献间关联的可行性;Attwood等^[10]则开发了一款语义分析工具Utopia Documents以解决文献和数据的孤立问题;Aalbersberg等^[11]识别并挖掘论文中隐藏的语义实体,并与科学数据仓储中对应的数据建立链接。此外,卫军朝等^[12]以学术期刊与科学数据仓储的关联为研究对象,提出图书馆推进文献与数据关联的途径与策略;Venkatesan等^[13]开发了SciLite平台,整合不同来源的文本挖掘注释,并将这些注释和基础数据与原始文献链接。可见,科学数据与科技文献关联的研究不断取得进展,但尚处于起步阶段,相关制度仍有待完善,统一的关联流程和框架尚未形成。在开放科学的环境下,如何更好地发挥科学数据与科技文献的价值,有效建立两者之间的关联是当前亟需解决的问题。因此,本研究尝试构建开放科学环境下科学数据与科技文献的关联模式,充分挖掘科技资源的价值,以推动科学数据管理与共享的实践进程。

1 开放科学环境下科学数据与科技文献关联面临的新形势

1.1 数据质量的高要求

开放科学环境下,用户对数据质量的要求越来越高,包括数据的完整性、准确性、及时性、相关性等。数据质量成为数据资源建设的关键影响因素,同时也决定着数据资源之间相互关联的效果^[14]。只有在确定数据准确的基础上提高内容的真实性和可靠性,保证科学数据的质量,才能提升科学数据与科技文献的关

联效果,使更多的优质数据被发现与利用,从而实现科学数据价值的最大化。为此,相关机构和平台应在数据收集、存储、出版等环节进行严格的质量控制,检查数据本身、元数据以及数据文档的完整性、一致性和有效性,并定期维护相关的元数据链接,避免出现无效链接等问题。

1.2 资源的可获取性

如果资源不可获取,或者只有通过付费的方式才能获取,那么数据资源的可用性将会大大降低。目前,公共科学数据仓储和机构数据库的科学数据获取方式多为协议共享或分级共享,完全共享占比比较少。因此,整体上看,科学数据的可用性还有待提升。开放科学强调数据资源的可访问性,消除科研人员获取数据和文献的障碍,免费为科研人员甚至是普通公众提供访问、下载服务,使用户能够在阅读论文或查看数据的同时,根据关联信息直接访问相关的科学数据和科技文献。这不仅有助于促进用户对知识的理解,为数据提供者和数据用户提供合作的机会,还能够提高科研资源的利用效率。政府机构、出版机构、科研机构、科研资助机构、图书情报机构以及科研人员等需要共同推动开放科学环境建设,通过建立数据共享和重用激励机制、数据平台安全监管机制以及开发强大的数据管理工具等方式,促进科研资源的开放共享。

1.3 数据的安全性

数据开放意味着其安全性也受到威胁。如今,出于对数据本身和数据平台安全的顾虑,科研人员较少以提供者身份参与数据共享。借助区块链、数据密钥等先进的技术维护科学数据、科技文献以及相关数据库、数据平台的安全,可以大大降低数据被泄露、剽窃的风险,有效保障数据提供者的权益不受侵犯,进而提升数据汇交和共享的意愿,最终扩大数据库资源的存储量。在数据量充足的情况下,通过建立科学数据与科技文献之间的关联,还能促进知识挖掘与融合。在开放科学环境下,需要根据不同类型、不同领域数据资源的特点,构建多元融合、组织有序、安全开放的资源保障体系,提高系统的兼容性和互操作性,从而对科学数据和科技文献进行有效管理。

2 开放科学环境下科学数据与科技文献关联模式构建

进一步加强科学数据与科技文献的关联,实现两者之间的深层次聚合与知识发现,对于促进科学数据的共享与复用、提高科学研究的效率意义重大。针对文献与数据的关联,研究数据联盟(Research Data Alliance, RDA)和世界数据系统(World Data System, WDS)提出顶层互操作框架Scholix(Scholarly Link Exchange),可以用于学术文献和数据以及数据集之间的链接信息交换^[15],进而实现了不同永久标识符体系、不同数据引用方式的学术成果间的互联互通。目前,该框架已得到多个数据库的采纳与使用。Scholix可以看成是“批发商到批发商”的交换框架,由DataCite、CrossRef、OpenAIRE等现有的“批发商”聚合实现。通过公共概念模型、信息模型和开放交换协议,Scholix可实现数量较少的大型链接中心之间的互操作,且尊重现有的特定社区的实践规则^[16]。在该框架中,用户提供的科学数据或科技文献为学术实体,数据存储

库、学术出版商等提供商则为学术实体的关联提供元数据,用户也将参与这一过程。此外,Scholix还将统一定义和标准化学术实体的指标和术语,以确保关联过程的一致性和可靠性。因此,本研究参考这一顶层互操作框架,基于开放性、系统性、标准化、动态性、互操作性等原则,构建开放科学环境下科学数据与科技文献关联模式的总体框架,反映数据与文献从散落分布到有序整合,再到增值利用的完整过程。

提出的科学数据与科技文献关联模式总体框架(见图1)参照计算机网络的结构模型,自下而上分为数据层、关联层和应用层,各部分紧密衔接、环环相扣。其中:数据层是各项工作顺利开展的基础,主要包括科学数据/科技文献收集、存储以及数据质量审核;关联层是该关联模式中最关键的部分,通过元数据关联、引用关联和语义关联等方法建立科学数据与科技文献的关联;应用层位于最上层,为用户提供知识聚合、知识挖掘、知识发现、关系预测、个性化推荐等多样化的知识服务,同时要求用户反馈应用成果。下面针对该关联模式的基本流程进行介绍。

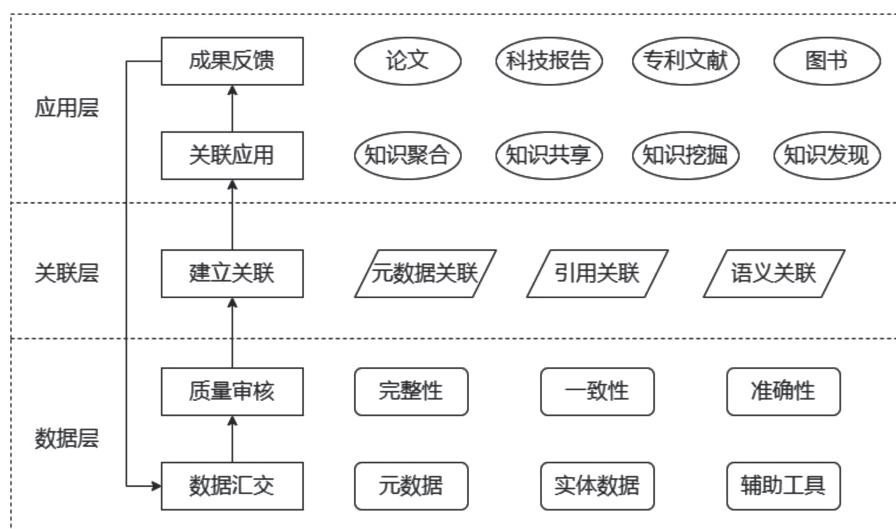


图1 科学数据与科技文献关联模式总体框架

2.1 数据汇交

在开放科学环境下,数据汇交与存储是科学数据与科技文献关联的前提和保障。其主要任务是对科学数据进行收集、存储与整理,相关主体主要涉及科学数据管理平台 and 期刊出版商。各级科学数据管理平台应按照《科学数据管理办法》和《国家科技资源共享服务平台管理办法》等相关政策的要求,开展科技计划项目

来源数据和科技论文支撑数据的数据汇交活动,在平台页面专门设置数据汇交入口,为用户提供数据汇交服务。部分期刊在意识到论文支撑数据的重要性后,要求作者在提交论文的同时将相关数据集上传至规定的数据仓储。通过汇交科技报告、科技论文或其他来源的数据,获取不同领域、不同主题的科学数据,为科学数据与科技文献的关联作充足的准备。

在数据汇交阶段,用户应按照科学数据管理平台

的要求, 上传正确格式的数据, 以便平台对科学数据进行进一步规范管理。科学数据汇交的基本流程为: 首先, 数据负责人注册相关信息后, 选择线上或线下提交汇交计划, 通过初步审核后, 填写汇交元数据、凭证信息并上传全部实体数据集; 然后, 平台管理员对其进行审核, 将审核结果通知数据负责人, 并对数据进行管理 与存储; 最后, 数据按照共享协议发布, 这也意味着项目数据汇交工作的结束。

科学数据的统一汇交不仅对科学数据的进一步规范管理有重要意义, 提高了数据的开放共享水平, 还为科学研究、公共服务及重大战略决策提供了丰富、可靠、高效、安全的资源保障, 同时也能更好地支撑国家科技创新和经济社会发展。此外, 科学数据管理平台也可以对科技计划科学数据增值成果进行回溯和再汇交, 比如科技论著、科技报告、软件工具等, 从而实现科技资源的汇聚和增值。

2.2 质量审核

在收集数据的过程中, 平台要严格进行质量审核, 包括对实体数据及元数据的审核。首先, 根据规定格式, 审核作者提交的实体数据内容是否符合规范, 如数据类型和格式是否符合要求, 数据内容是否完整、准确^[17]。其次, 元数据作为科学数据与科技文献建立关联的重要依据, 和实体数据同等重要。由于科学数据仓储通常要求作者提供元数据, 甚至允许用户自定义元数据信息, 元数据经常存在缺失、重复、错误、不一致、

不规范等问题。因此, 必须对元数据、实体数据加以控制, 只有经过严格的质量审核, 才能完成数据归档。

目前, 国内许多数据中心非常注重汇交数据的质量审核, 如国家空间科学数据中心、国家海洋科学数据中心、国家人口健康科学数据中心等。然而, 由于科学数据集本身规模大、结构复杂多样, 加上不同学科之间的差异性, 质量审核的难度大大增加。对此, 诸多数据中心均选择人机结合的方式来审核科学数据计划与实体数据。若上传内容规范、格式标准, 审核难度能在一定程度上减小。同时, 应以高效和安全为目标不断完善平台的基础设施, 建立规范的数据共享制度, 定时对平台和数据进行质量评估, 通过构建数据安全支持系统加强对平台的有效监管, 切实保障科学数据管理平台的安全。

2.3 建立关联

2.3.1 元数据关联

元数据作为描述数据的数据, 目前广泛用于各类资源的聚合与关联。本文所指的科学数据与科技文献元数据关联模式主要基于两者外部特征和内容特征的相似性建立关联。无论是科学数据还是科技文献, 通常都利用都柏林核心元素集 (Dublin Core Element Set, DC) 来描述其核心元数据, 并结合学科主题等方面的特点, 对相应元素进行增减。调查科学数据与科技文献的元数据, 结果见表1。

表1 科学数据与科技文献的元数据元素

元 素	标 签	组成部分			
		科学数据	科技论文	科技图书	专利文献
Title	题名	数据集名称	文献题名	书名	专利名称
Creator	创建者	数据集生产者	文献作者	著作者	发明人
Subject	主题	关键词	关键词	主题词	关键词
Description	描述	数据集描述	文献摘要	目次、摘要、附注	摘要
Publisher	出版者	发布机构	期刊单位	出版机构	专利权人名称和代码
Date	日期	发布时间	发表时间	出版日期、创建日期、获取日期	申请时间、公开时间
Type	资源类型	数据集	科技论文	电子图书	专利文献
Source	来源	科学数据管理平台	文献数据库	电子图书数据库	专利数据库
Identifier	标识符	CSTR标识、数据DOI	引用方式、论文DOI	ISBN、URI、图书DOI	专利号、专利分类号
Relation	关联	参考文献、关联论文、基金项目	参考文献、基金项目	其他版本、参照、被参照、包含、包含于、其他格式	引用专利、施引专利、参考文献
Coverage	覆盖范围	空间范围	时空范围	时空范围	空间范围 (将保护发明的国家/地区)

可以发现,虽然科学数据和科技文献存储于异构数据库,但是元数据元素相对一致。以作者为例,假如某个科研人员创建了一个数据集并据此撰写发表了相关论文,此时的科学数据与科技文献通过相同的作者产生直接关联。一般来说,科研人员的研究方向比较固定,其长期产出的科研成果通常仅涵盖少数研究方向,且它们之间存在一定的相关性,因此在没有上述直接关联的科学数据与科技文献之间也可能存在一定的主题相关性。综上,可以基于作者建立科学数据与科技文献的关联。此外,一项较为复杂的科学研究可能需要跨学科跨领域合作。根据科研人员之间的合作网络可以挖掘更多相关的研究成果,从而将科学数据与科技文献进行关联。以关键词为例,关键词通常能够反映科学数据和科技文献的主题内容,是重要的检索字段。具有相同或相似关键词的科学数据和科技文献在研究内容和主题方面也存在相关性,所以可以基于关键词建立两者的联系。同理,基于其他元数据元素的相似性均可以建立科学数据与科技文献之间的关联。

2.3.2 引用关联

科研人员按照相应的规范在文献中引用数据,形成基于引用的直接关联、基于同被引的关联、基于引用的扩展关联^[18]。

基于引用的直接关联模式是双向的关联关系,用户可以通过阅读科技文献找到相应的一个或多个支撑数据,也可以在查看科学数据时找到引用它的一篇或多篇科技文献。这不仅可以帮助用户深入理解数据的意义和

用途,还能够促进科学数据的发现与利用。当然,这是最直接的一种关联模式,仅突出存在引用关系的数据与文献,范围较窄,未明确体现同被引等引用关系。

基于同被引的关联模式与文献的同被引关系类似。如果一个科学数据D和一篇科技文献P同时被一篇或多篇科技文献引用,那么看似毫无关系的D和P建立起关联,它们有同被引关系。而且,同时引用D和P的文献数量越多,两者之间的关系越紧密,关联程度越高。

科学数据会被多篇科技文献引用,一篇科技文献也会引用多个科学数据,交叉引用所形成的引用网络就是基于引用的扩展关联模式的关键。在引用网络中,节点为科学数据或科技文献这两种实体,连线则是指两者之间的引用关系。只要有连线,两个实体之间就存在关联关系。科学数据会形成相关数据群,科技文献也具有相关文献群。那么,科学数据D不仅能与一篇直接引用它的科技文献P建立关联,还和与科技文献P相关的文献、与科学数据D相关的数据之间存在关联关系。在这个巨大的引用网络中,不同学科的知识交叉融合,科研人员可以利用基于引用的扩展关联模式找到跨学科的科学数据,从而进行多学科交流与合作。

2.3.3 语义关联

建立科学数据与科技文献之间的语义关联有助于知识的深层次发现。如图2所示,语义关联流程主要包括文本处理/数据类型判断与方法选择、概念/实体抽取、关系识别、语义标注结果、语义距离计算、语义发现6个环节^[9]。

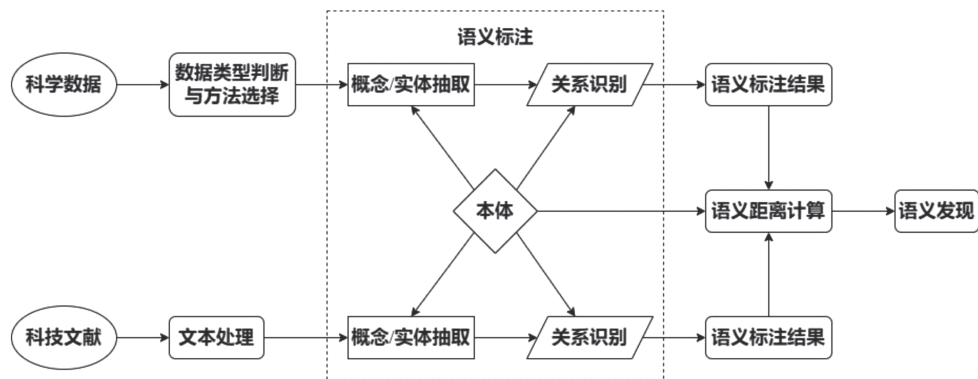


图2 语义关联流程

准确识别数据集和文献中的语义实体和关系对科学数据与科技文献的互联互通至关重要,语义实体主要包括具体的知识单元、抽象的科学概念或术语。借助

本体、语义标注、语义推理等技术对抽取出来的条目进行规范化处理和相互匹配,能够建立科学数据与科技文献之间的关联。科学数据的语义标注是指从科学数

据中提取概念内容并进行语义化表示的过程。由于不同学科领域的科学数据在类型、格式方面存在差异,难以使用相同的方法进行语义标注。因此,需要根据科学数据的特征进行相应的类型判断,再选择合适的标注方法,比如:对于数字、表格类型的科学数据,需要先借助本体将其转化为可理解的文本,再进行语义标注;对于图像、声音、视频等非结构化的科学数据,需要先利用已开发的语音识别工具、视频描述框架等对数据内容进行转换,再进行人工标注。科技文献的语义标注是指借助自然语言处理的相关技术、词典、工具等,将文献内容和结构转化为机器能够理解的文本,并进行下一步的实体识别和实体链接。为了更好地被计算机理解 and 处理,科学数据和科技文献的语义标注结果用RDF格式表示。基于本体语义关系,计算科学数据和科技文献标注结果之间的语义距离,达到异构系统间异构资源关联发现的目的。

2.4 关联应用

通过实现科学数据与科技文献之间的关联,能够有效拓展科学数据平台和文献数据库的服务范围,为开放数据服务奠定基础,从而促进科学数据和科技文献的共享与重用,最大限度发挥科技资源的价值。关联应用是指对科学数据和科技文献采集、描述、组织和整合之后,面向以科研人员为主的服务对象,开展知识聚合、知识挖掘、知识发现、关系预测、个性化推荐等多样化的知识服务。借助各种技术手段,引入Scholix索引,对多来源和多模态的科技资源进行物理、逻辑或规范层面的集中、关联、整合,方便用户查询、获取和利用资源,从而使科学数据的共享范围更加广泛,以便科学数据的发现、获取、重用。同时,完整、统一的资源系统结构为知识检索提供了便利,有利于进一步挖掘知识之间的潜在关系,促进深层次的知识发现,为科研活动的顺利开展提供有力支撑。

数据与文献的关联也能支持跨学科、跨系统的资源复用,从而提高资源的开发和利用效率,加快开放知识的速度。空间、天文、地球系统等领域数据中心的科学数据通常由文字、数值、表格、图片、视频等组成,数据内容涉及多个领域,隐藏着学科领域之间的相关关系。利用关联关系,不仅能够促进天文、地理、化学、生物、计算机等学科知识交叉融合,还可以对学科热点或发展趋势进行分析预测,拓展研究的深度和广度^[19]。

此外,可以根据知识之间的关联关系和预测结果进行个性化推荐,主动将用户感兴趣的内容定期推送至页面,同时让用户进行服务效果评价,平台根据反馈意见不断调整、改进相关功能,以便为用户提供更优质的知识服务。通过共享与重用科学数据,科研人员可以找到与自己研究兴趣相同或相近的研究者,针对某一主题进行深入的交流探讨,实现跨地区、跨机构、跨学科的沟通与合作,从而促进知识的传播和科学研究的发展^[20]。通过利用已有的科学数据进行更深入的研究,不仅可以避免重复劳动,还能够从中发掘创新潜力,为科学研究注入强大动力。

2.5 成果反馈

为保证关联模式能够持续不断地顺利运行,在数据共享与重用的过程中,科学数据管理相关部门、机构、平台应明确告知数据使用者使用权责。数据使用者需在获得使用资质后,将基于共享数据所取得的增值成果再汇交,具有反馈通报的责任。科学数据管理平台应明确相关约束制度,并签订框架协议作为约束条件,以此保障由共享和重用国家科技计划和论文关联的科学数据产生的增值成果如约回流和汇交。由于数据使用者通常在开放数据的基础上做进一步研究,科研成果的主题与科学数据本身的主题存在一定相关性。因此,平台管理人员可以基于各类增值成果的汇交情况进一步建立其与科学数据之间的关联,实现该模式的循环、反馈。通过建立科学数据与科技文献的关联,资源的共享程度不断提升、传播范围不断扩大,科研人员可以更加方便地获取和利用所需数据,同时也能够更加深入地理解他人的科研成果。这不仅有助于提高科研效率,加速科研成果产出,还为完善学术评价体系提供了有力支持。

3 促进科学数据与科技文献关联的对策建议

3.1 完善科学数据相关政策体系

完善的科学数据管理、关联、引用等相关政策是科学数据与科技文献关联工作的根本保障,可以规范和引导这两者的关联以及科研人员管理、引用数据。调研发现,大多数国外成熟的数据平台在其主页的醒目位置

设有相关政策栏目,且在這些政策中科学数据的价值和知识产权内容均处于首要位置。显然,无论是在科学数据与科技文献的关联上,还是在对科学数据的共享与重用方面,明确利益相关者的权益和责任都尤为重要。这不仅可以指导科学数据与科技文献的关联、减少数据共享和引用等方面的纠纷,还能够激励科研人员共享科学数据。

面向科学数据的提交与存储,国内相关政策虽要求汇交科学数据,但对于数据的类型、格式以及提交方式并没有统一的标准。科学数据引用政策多是鼓励型政策,只有一些科学数据平台在数据引用方面制定了较为详细的规定。虽存在面向不同学科领域的科学数据政策,但其未能充分考虑相应学科领域的的数据特点。整体上看,我国虽具有普适性的政策内容作为指导,但相应的政策内容不够丰富,学科、区域等细分领域政策的关联性较弱,尚未形成一套系统的且适应我国科学数据发展现状的科学数据开放共享政策体系。

政府部门、科研机构、出版商等还需结合我国整体的实践情况和各领域数据的具体特点,从宏观、中观、微观3个层面统筹协调,不断制定和完善数据开放共享政策,明确利益相关者的责任和不同类型数据的开放共享方式,充分挖掘科技资源的价值,避免造成数据的浪费和相关人员的权责纠纷。同时应借鉴国外的先进经验,在科学数据的存储、汇交、关联、重用等方面制定更为明确的政策,从引用标准的统一、引用内容的细化等方面不断进行完善。要关注不同学科领域的的数据特点,提供具有针对性的数据使用说明,从而增加科学数据与相关科研成果的关联强度,促进科研效率的提升。此外,在构建科学数据开放共享政策体系的基础上,细化与标准化关联政策,提升数据政策的适用性。

3.2 丰富科学数据平台服务功能

根据调查和分析发现,目前科学数据平台在关联服务功能建设方面仍处于起步阶段,数据准备尚不充分、数据检索方式单一、数据更新不及时、意见反馈机制缺少等问题都影响着平台的服务效果。为更好地满足用户需求,在检索方法方面,应增加高级检索功能,并设置检索指南,引导用户构建完整的检索表达式,支持按科学主题、细分学科、来源机构、出版日期等元数据进行组合检索,从而获取精确的检索结果。在检索界面方面,依据简洁、友好、易用的原则,对用户界面进

行优化,提高检索效率。同时还应提供用户意见反馈通道,通过人工或智能助手窗口及时回复用户疑问,并根据用户意见调整相应功能,提升平台服务质量。在支撑工具方面,平台要不断开发新的数据挖掘和分析工具,创新关联方法,融合人工智能、机器学习等技术深入挖掘科学数据与科技文献之间的关联关系,实现科学数据的精准定位与发现。

科学数据平台可以借鉴国外成熟的文献与数据关联案例,如ScienceDB通过引入Scholix索引进一步实现了数据与出版物的互联互通。国内,中国知网(CNKI)、维普、万方等文献数据库积累了宝贵的建设经验,并不断完善平台功能,满足用户的一站式检索和多样化检索需求。比如,出于知识管理的需要,CNKI搭建了知识网络结构框架,并基于相似性比较、自动聚类理论、方法和技术,实现了各种知识主体之间的相互链接,组成关联知识网络^[21]。在检索结果页面,用户点击任意文献、作者或出版物,即可进入相应的文献知网节、作者知网节或出版物知网节。随着共享平台上的资源变化,每一节点的链接信息都会动态实时更新。这种知识整合模式不仅可以直观展示知识的起源和不同阶段的发展进程,方便用户快速、高效地获取和利用知识,还充分揭示了各类信息之间的直接关联与间接关联,有利于知识的发现、交流、学习、共享、创新等。科学数据平台可以从元数据要素出发构建知网节,通过链接方式对数据进行关联,方便用户查找。但由于部分平台存在无效链接,今后还需对数据链接进行定期维护、及时更新。

除了基于元数据构建知网节以外,还可以基于数据与文献的引用关系构建引用网络,通过可视化的方式展现数据与文献的关联关系。与文献类似,科学数据也可以反复利用、不断增值。科学数据引用指的是科研人员通过参考文献、正文注、脚注、尾注或致谢等形式在论文中提供数据出处的做法^[22]。科学数据与科技文献的引用关系表现出两者之间交叉融合的关联关系,科学数据引用网络节点如表2所示。通过构建科学数据引用网络,将节点数据的参考文献、引证文献、共引数据、同被引数据整合到一个检索结果页面,方便用户轻松链接、获取引用文献的知网节及全文。同时,数据平台也应根据数据量和文献量的增加,对引用网络进行动态实时更新,帮助识别数据归属、追踪数据来源。这不仅有助于分析和追踪科学数据的发展历程,还可以为学科发展趋势的评价与预测提供依据^[23]。

表2 科学数据引用网络节点

节点类型	定义	功能
参考文献	节点数据所引用或参考的文献	反映节点数据的研究依据
引证文献	引用或参考节点数据的文献	反映节点数据相关研究工作的发展或评价
共引数据	与节点数据有共同参考文献的数据	动态提供与节点数据相同、相似、相关的研究背景、研究依据或研究内容的数据
同被引数据	与节点数据共同被引用的数据	动态提供与节点数据同时被引用的数据，可共同作为进一步研究的基础

3.3 增加关联文献的类型

目前,科学数据平台在建立科学数据与科技文献的关联时,关联对象集中于期刊论文,关联的文献类型较为单一。但事实上,科技文献种类非常丰富,学位论文、会议论文、专利文献、科技报告、科技图书等均属于科技文献的范畴。通过关联科学数据与不同类型的科技文献,可以拓展科学数据的发现途径,这对于追溯科学数据归属、促进数据的共享与重用、提高资源的利用效率等均发挥着重要作用。

科学数据与科学文献、科技报告、专利、图书分别属于不同类型的数据资源,虽然存储于异构数据库,但是可以利用多种方法建立起关联。一方面,可以根据元数据描述内容分析作者、主题、名称等易于发现的外部特征,建立直接关联,链接同一作者发布的、相关主题的科技文献。除引用他人的数据外,大部分以原始数据为基础进行分析、推演、挖掘而形成的科技文献,其作者和其支撑数据的作者相同^[24]。另一方面,也可以基于相互之间的引用关系构建科学数据与科技文献的关联。科学数据是重要科研成果,根据引用关系识别其流动和扩散方向有助于实现知识的关联发现,从而推动科学创新和进步。比如,就专利文献而言,专利引用数据体现出科学数据对技术领域的贡献和支撑作用,反映了科学知识向技术领域的流动和扩散,而数据引用专利代表着科学领域对技术知识的吸收与融合。专利和科学数据的关联关系不仅能够反映数据的流动方向,还有助于研究者了解科学和技术的融合和相互作用情况^[25]。此外,还可以基于科学数据和科技文献之间的引用关系进行定量分析和知识发现,为科学数据的价值和影响力评价提供依据。

3.4 加强相关主体之间的合作

虽然我国科学数据平台、学术期刊、图书馆等相关主体从不同角度进行了关联实践探索,但是当前科学数

据与科技文献的关联关系大多体现在数据平台一方,即用户在科学数据平台可以点击查看关联论文或引用论文,而在文献数据库平台如CNKI上却无法直接查看相应的支撑数据。可见,科学数据与文献之间仅建立了单向联系,尚未建立双方或多方的相互联系。为实现科学数据与科技文献的双向关联,我国应充分借鉴国外的先进经验,加强相关主体之间的合作,通过元数据映射、接口链接、数据资源整合等多样化的方式实现两个或多个不同数据库之间的互联互通。例如,国外的Elsevier商业数据库与Dataverse、PANGAEA、Dryad、ICPSR等50余个科学数据仓储合作,通过链接模式建立了数据与文献的双向关联,还提供可验证数据DOI和语义实体链接、创建横幅标识、开发集成数据App等,在极大程度上发挥了科学数据与文献的价值,方便用户的数据获取行为。

无论是商业数据库还是开放获取数据库,都应该在相关政策的指导下,积极开展与科学数据仓储的交流合作,通过语义实体识别和链接来实现科技文献与科学数据仓储的语义关联。如此,不仅可以很好地支持科学研究与发现,还有利于扩大数据库中科技成果的影响力。对于图书馆数字资源平台而言,可以通过目录整合的方式实现科学数据与图书资源之间的关联,从而提供多样化的数据服务,满足用户个性化的数据需求。同时,在数据收集方面,科学数据平台也应与其他各级数据中心、科技文献出版方、图书馆等开展深入的合作,多渠道获取数据,注重数据收集的质量,构建科技资源的利益共同体,共同促进科学数据和科技文献的共享和重用。为获得足够的资金支持,相关平台或科研机构也可以面向企业或商业数据库开展数据服务,合作开发集成系统,满足用户一站式检索数据的需求。

4 结语

关联和融合科学数据与科技文献是数据密集环境

下科学研究的迫切需要,也是科学数据开放共享活动的关键环节。在开放科学环境下,科学数据与科技文献的关联显得尤其重要。建立两者之间的关联是实现科学数据获取和重用的可靠途径,是推进科学数据管理与共享工作的有效手段和完善科研成果评价体系的有力支撑,也是促进科研合作和学术创新的强大动力。

本研究虽然构建了开放科学环境下科学数据与科技文献关联模式,并提出两者关联的对策建议,但仍有一些不足之处,需要后续研究进一步思考与完善。①仅从理论层面探索两者的关联模式框架和基本流程,构建的框架结构尚不够完善,未来应结合关联数据、本体、语义实体等应用模型,实现资源间的多维度、深层次互联。②从宏观的角度对两者的关联进行整体分析和总结,未针对具体类型的科学数据进行分析,亦未充分考虑学科领域的差异。未来应针对具体学科领域科学数据的特点,结合科研人员在两者关联方面的具体需求,完善模型设计,提升关联框架的可行性。③仅立足于开放科学环境的影响来构建两者的关联模型,并未充分考虑开放科学与两者关联的内在关系,今后应结合开放科学的特点进行深入探讨。

科学数据经过前期的收集、中期的处理、后期的分析实现价值最大化,彰显学术意义,为社会解决现实问题。关联研究是挖掘不同数据之间的相互关系,并从中发现研究对象之间潜在联系的重要方式,对于促进知识的发现、融合、利用,推进科学研究的发展具有重要意义。目前,科学数据与科技文献的关联研究尚处于探索阶段,后续仍需各利益相关者共同努力,以推动科研资源整合和利用的实践进程。

参考文献

- [1] 温亮明,李洋,郭蕾,等.《开放科学建议书》制定背景、内容体系与科学价值[J].图书馆论坛,2022,42(4):18-26.
- [2] 周晓英.情报学进展系列论文之七 数据密集型科学研究范式的兴起与情报学的应对[J].情报资料工作,2012(2):5-11.
- [3] LAWRENCE B, JONES C, MATTHEWS B, et al. Citation and peer review of data: moving towards formal data publication[J]. International Journal of Digital Curation, 2011, 6(2): 4-37.
- [4] 黄筱瑾.基于元数据的科学数据与科技文献关联研究[J].情报理论与实践,2013,36(7):27-30.
- [5] 黄筱瑾.基于内容特征的科学数据与科技文献关联研究[J].现代情报,2018,38(1):56-59.
- [6] 陈帅印,刘桂锋,刘琼.基于元数据的数据论文与期刊论文的关联研究:以全球变化科学研究数据出版系统为例[J].数字图书馆论坛,2022(8):11-18.
- [7] 张鑫,文奕,杨宁,等.基于引文探针的文献与数据的关联算法与应用:以高能物理领域为例[J].情报理论与实践,2019,42(10):151-156.
- [8] RIEDEL N, KIP M, BOBROV E. ODDPub: a text-mining algorithm to detect data sharing in biomedical publications[J]. Data Science Journal, 2020, 19(1): 42.
- [9] 丁培.科学文献与科学数据细粒度语义关联研究[J].图书馆论坛,2016,36(7):24-33.
- [10] ATTWOOD T K, KELL D B, MCDERMOTT P, et al. Utopia documents: linking scholarly literature with research data[J]. Bioinformatics, 2010, 26(18): 568-574.
- [11] AALBERSBERG I J, KÄHLER O. Supporting science through the interoperability of data and articles[J]. D-Lib Magazine, 2011, 17(1/2): 1045.
- [12] 卫军朝,宋婧婷.学术期刊与科学数据仓储关联研究:兼论图书馆科学文献与科学数据关联的途径[J].图书与情报,2018(1): 126-133.
- [13] VENKATESAN A, KIM J H, TALO F, et al. SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data[J]. Wellcome Open Research, 2016, 1: 25.
- [14] 陶蕊,叶继元.开放科学环境下高校图书馆数据资源建设调查与分析[J].图书馆论坛,2023,43(7):75-83.
- [15] Discovering research data links via GESIS LOD[EB/OL]. [2023-09-16]. <https://scholexplorer.openaire.eu/#/about>.
- [16] 白海燕.学术文献和科学数据的融合机制研究[J].数字图书馆论坛,2023,19(6):40-47.
- [17] 田稷.开放科学环境下科学数据混合出版新模式研究[J].出版科学,2022,30(5):60-67.
- [18] 郭学武.基于引文的科学数据与科技文献关联研究[J].情报科学,2014,32(4):59-62,125.
- [19] 余文婷,吴丹.基于数据关联的开放科学数据长期保存利用研究[J].信息资源管理学报,2014,4(2):87-93.
- [20] 赵梦凡,支凤稳,彭兆祺,等.国内外中小企业竞争情报研究:基于ITGInsight的对比分析[J].情报科学,2022,40(10):123-130.
- [21] 张丽华.知网节与知识网络[J].现代图书情报技术,2006(9): 85-88.

- [22] 李丹丹, 吴振新. 研究数据引用研究[J]. 图书馆杂志, 2013, 32 (5): 65-71. 刊论文互联机制研究[J]. 情报理论与实践, 2023, 46 (5): 166-175.
- [23] 屈宝强, 王凯. 科学数据引用现状和研究进展[J]. 情报理论与实践, 2016, 39 (5): 134-138, 114. [25] 王诗炜, 陈春. 基于科学论文和技术专利关联关系识别潜在知识发现方法研究综述[J]. 数据分析与知识发现, 2023, 7 (7): 18-31.
- [24] 尚渡新, 刘桂锋, 王婧怡, 等. 基于关联数据的科学数据与期

作者简介

支凤稳, 女, 博士, 副教授, 硕士生导师, 研究方向: 竞争情报、科学数据。

赵梦凡, 女, 硕士研究生, 通信作者, 研究方向: 科学数据、竞争情报, E-mail: mfzhao98@126.com。

彭兆祺, 男, 硕士研究生, 研究方向: 科学数据、竞争情报。

Relevance Pattern of Scientific Data and Scientific Literature in Open Science Environment

ZHI FengWen ZHAO MengFan PENG ZhaoQi

(School of Management, Hebei University, Baoding 071002, P. R. China)

Abstract: Under the data-intensive scientific research paradigm, scientific data has become equally important scientific research resources as scientific literature. The rapid development of open science has brought opportunities and challenges to scientific research resource management. It is of great significance to explore the correlation mode of scientific data and scientific literature under the open science environment to improve the utilization efficiency of resources, and also to provide useful references for related theoretical research. This paper analyzes the new situation faced by the association between scientific data and scientific literature in the open science environment, and builds the relevance pattern of scientific data and scientific literature. The basic process includes data exchange, quality audit, association establishment, association application, and result feedback. Finally, some suggestions are put forward to promote the correlation between scientific data and scientific literature: perfecting relevant policy system, enriching platform functions, increasing types of related documents, and strengthening cooperation among relevant subjects.

Keywords: Open Science; Scientific Data; Scientific Literature; Relevance Pattern

(责任编辑: 王玮)