学术文献和科学数据的融合机制研究*

白海燕
（中国科学技术信息研究所，北京 100038）

摘要：针对科研资产主体的学术文献和科学数据之间的互联互通问题，研究融合学术文献与科学数据以提升科研效率的具体模式，探讨开放、可靠、可扩展科研信息基础设施的数据与文献融合机制，为面向数据密集型科研的传统文献信息服务转型和创新提供新的思路和参考。通过研究不同利益相关者在学术文献与科学数据融合方面的最佳实践和典型案例，基于驱动因素、技术实施、行业规范、实际效果等方面的分析比较和归纳总结，提炼出4类科研共享生态环境下的实现机制和3类融合学术文献与科学数据的服务模式，并评价其科研效率。

关键词：开放科学；科学数据；开放数据；信息服务
中图分类号：G251　DOI：10.3772/j.issn.1673-2286.2023.06.005
引文格式：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 (6)：40-47.

随着语义网、本体、云计算、关联数据等技术的发展，科学研究环境也从电子科研（e-Science）逐渐向关联科学（Linked Science）[1]转变，其特征为科学资产（Scientific Assets），包括科研资源、科研成果和科研条件等，以互联互通的方式支撑和推动透明的、可重复和跨学科的科学研究。学术文献和科学数据无疑是科学资产的主体[2]。曾提出第四范式——数据密集型科研(Data-Intensive Science)[3]的Jim Gray将科学研究资源划分为金字塔型的3个层面，从上到下包括处于塔尖的文献，中间层的派生数据、重组数据和底层的海量原始数据[4]。Jim Gray提出关联所有的科学数据与文献，形成一个互操作的世界，读者在阅读文献时，可以访问论文的原始数据甚至重复研究过程，或者从数据中找到与之相关的所有文献。这种数据与文献的互操作可以提升“信息速率”，进而提高科学生产力[5]。

文献与数据融合已成为科学交流的现实，并不断深化。以高能物理领域的科学研究资源金字塔为例：INSPIRE文献服务平台位于最顶层，在原有的文献发现基础上，开始提供论文图表数据检索、发现功能和引用链接；HEP OpenData服务位于下一层，提供论文支撑数据存储与获取功能；再下一层，CERN Analysis Preservation这类服务可实现对科学数据（经过仿真和模拟的部分科学数据）的分析和挖掘；最底层则是研究所产生的海量原始数据。

本研究主要聚焦科学研究资源金字塔的顶层，即面向学术文献和相关科学数据（文内数据[6]和出版数据[7]），通过文献调研和系统分析，重点发现有利于提升科研效率的数据与文献融合的主要场景和服务模式，探讨开放、可靠、可扩展的文献与数据融合机制，试图为面向数据密集型科研的文献信息服务系统转型和创新提出新的思路。

1 研究现状

国外相关研究集中于学术文献和科学数据关联的起因和驱动机制。Smit[8]分析证明数据和文献之间

收稿日期：2023-03-14
*本研究受到国家科技图书文献中心专项任务“下一代国家科技创新知识服务开放系统先期研发”子项目“面向关联科学（Linked Science）的科技文献与科学数据整合研究”（编号：XQYF0203）资助。

的联系提升了数据和文献的可见性、发现能力和检索效率。Ball等[9]、Cook等[10]、Lagerstrom[11]发现，文献和科学数据关联的最大驱动力来自科学成果的信用归因机制，良好的信用归属被认为是激励研究者发表数据的重要因素，而通过测度一个数据集在文献中或其他数据集中的被引频次，可以评价其影响力。此外，数据重用和研究再现的危机也驱动文献与科学数据关联[12]。有调查[13]显示，数据缺乏通常会导致研究成果不可复制，文献为正确的数据重用提供了有价值的上下文信息。在应用研究方面，Borgman[14]认为，文献提供了正确的数据解释和重用所需的上下文信息，而一种链接数据和文献的标准化通用方法将提升重现科学研究的可能性。Martone[15]、Starr等[16]提出一种学术论文系统，该系统可实现数据和文献之间的双向链接。

国内相关研究主要侧重于实践应用，经历了从局部链接技术到全局框架，从农业、医学领域示范到全学科应用的过程。涂勇等[17]介绍了基于DOI的链接实现方式，邱春艳[18-19]、卫军朝等[20-21]、黄筱瑾[22]对文献与数据的关联方式进行了比较全面的研究。在特定领域，国内的学者进行了内容、语义和引用层面的研究。韩涛[23]在生物信息学领域提出交叉引用、内容与知识层面聚合、知识关联等5种关联方式。丁培[24]结合基因组测序项目数据，验证了基于本体、语义标注、语义推理等技术来实现数据和文献关联的可行性。黄筱瑾[25]提出向量空间模型的内容特征相似性计算方式。贺姝祎等[26]分析比较了天文领域科学数据与文献在元数据方面的相似性及差异。孙巍[27]建立了农业科学数据和科技文献实体间的语义关联，实现了两种类型实体间的关联与集成发现。郭学武[28]从引文角度分析科技数据与科技文献相互关联的3种主要模式。丁楠等[29]基于引文分析的方法，选取数据发布量、被引量、被引频次、h指数等指标，构建了基于引用的数据评价体系并进行了实证研究。黄永文等[30]提出科学数据检索与关联服务系统架构，并实现学术资源元数据采集与融合、科学数据元数据丰富与增强以及科学数据检索与关联发现服务。

可见，国内外研究都更多关注技术方法的突破，而在实践过程中，关于传统文献信息服务如何集成和融合数据服务、不同利益相关者如何有效驱动和协同、如何消除科研资产的异构性、如何从机制和全局业务流层面制定融合的规范和标准等，还缺失相应的系统研究，这直接影响技术方法最终的实现和工程化应用。本研究通过调研知名出版商、搜索引擎、数据中心、基金协会等利益相关者在文献与科学数据融合方面的最佳实践和典型案例，提炼出实现学术文献与科学数据融合的4类基本机制，在此基础上就学术文献与科学数据融合之后的应用和服务模式进行归纳和总结，形成完整的学术文献与科学数据融合实现机制与应用策略，为面向数据密集型科研的传统文献信息服务转型和创新提供新的思路和参考。

2 学术文献与科学数据融合的实现机制

对于如何实现学术文献和科学数据的融合，目前还没有统一和最佳的方案，研究和实践呈现出多样化趋势，因此需要从机制层面研究构建融合学术文献与科学数据的框架性和全局性方案。本文通过对相关项目、应用和系统的分析，以驱动和协同方式、全局业务流规范和标准等为基础，提炼出融合学术文献与科学数据的4类基本机制。

2.1 双边协议机制

双边协议机制是指文献出版商与数据仓储或数据中心通过协议建立学术文献与科学数据的双向链接，协议基础是学术期刊的数据公开政策。学术期刊要求作者将与论文相关的科学数据提交至公开的数据仓储，通过与数据仓储的双边协议支持数据关联。例如，Elsevier与众多领域的数据仓储或数据中心基于双边协议建立数据双向链接，同时也指定了如Mendeley等公共数据仓储；Wiley、Springer Nature等出版商要求作者将科学数据存储在PANGAEA等数据存储库中；*PLoS ONE*要求所有论文涉及的数据都必须无限制开放，要求作者在提交论文时同时提供一份“数据可用性说明”以描述论文涉及的科学数据的使用和访问方法；BioMed Central鼓励所有作者将数据以一种可机读的方式存储在公共数据库中，并且在“数据可用性说明”中提交数据DOI或Accession Numbers。学术出版的数据政策涉及提交数据、保存数据、开放数据、共享数据等措施，对数据的提交方式，数据的内容、共享范围、有效期、版权、保存方式等进行了具体规定[31]。

通过文献出版商和数据存储方的双边协议，双方可获得准确、双向的数据链接，双边协议机制是大多数文献服务平台最早采用的关联模式。但是，双边协议

机制存在很多问题：①对于双边对象之外的大多数发布者、数据中心、存储库和基础设施提供者而言，关联无效；②不同体系的内在异质性阻碍了全球互操作的实现，规范和行业标准缺乏。

2.2 事件数据机制

“事件”是指文献或数据已注册特征项与特定活动之间产生关系，例如文献和数据之间的引用、提及、重用等。事件数据（Event Data）机制是一种不通过协议的自主构建模式，打破了机构壁垒，具有广泛性和行业标准化的特征。

CrossRef和DataCite从2014年开始采用DOI进行数据发表和引用，并联合开发了生成文献和数据关联的服务——事件数据服务[32]。通过事件数据服务，提供DOI和其他数据源之间的通用链接，使出版物与数据引用、软件重用等之间的联系变得清晰和规范[33-34]。数据引用是事件数据服务捕获的事件子集：通过DataCite元数据的relatedIdentifier和nameIdentifier属性发现，事件的例子包括对相关数据的引用、期刊文章中的数据引用等。事件数据关联不局限于DOI，还包括其他永久性标识符（Persistent Identifier，PID），如研究者ID，即研究者的活动成为事件的子集。例如，RMap项目[35]、National Data Service、bioCADDIE、Open Science Framework和THOR（Technical and Human Infrastructure for Open Research）等都使用研究者ID、DOI等事件数据构建文献和数据乃至研究者的关联[36]。

2.3 多边模式下的第三方机制

为了打破双边协议机制的封闭性，“one-to-all”的多边解决方式出现。国际研究数据联盟（Research Data Alliance，RDA）的多个工作组和兴趣组将文献和数据关联列为重要的研究主题，探索以第三方枢纽方式建立一个通用的、具有共同标准的服务体系结构，使科学数据领域的各利益相关者均受益[37]。

2.3.1 数据聚集模式DLI

RDA数据出版工作小组发起DLI（Data Literature Interlinking Service）项目，通过利益相关者联盟，就共同的标准达成一致意见来突破双边协议机制的局限。DLI项目成员包括RDA数据出版工作小组、OpenAIRE（Open Access Infrastructure Research for Europe）、国际科联世界数据系统（ICSU-WDS）、国际科学技术与医学出版商（STM）协会、CrossRef、DataCite、ORCID、National Data Service和RMap项目[38]。

DLI提供的数据-文献互连服务是一个基于Web的开放服务，用于识别数据集与给定的相关文章的关联。DLI项目从各种数据中心、出版商和研究组织收集和聚合“权威”数据和文献，构建两者的链接图并开放访问权限。数据和文献的链接图可服务以下用户：通过网络门户搜索和浏览链接的终端用户、通过应用程序编程接口（API）访问链接图中出版物和数据集的第三方服务、愿意提供高质量权威链接的内容提供者。基本服务包括全文搜索字段或出版物和数据集匹配查询（限制10 000个结果）、通过开放存档计划元数据收割（OAI-PMH）协议批量访问、PID解析功能、返回含有给定PID的链接。

DLI项目初期面对大量的问题：首先，缺少明确的内容获取政策和规范标准，因此缺少对出版物、数据集的最低质量要求；其次，在技术方面存在数据访问效率低、数据可伸缩性差等问题。目前，DLI已成为OpenAIRE基础设施的一部分，通过在OpenAIRE的基础设施上部署服务来解决上述问题。

2.3.2 软件解决方案RD-Switchboard

RDA的DDRI（Data Description Registry Interoperability）工作组开发了RD-Switchboard作为开放协作的软件解决方案，以解决跨平台发现科学数据的问题。RD-Switchboard的主要功能是将数据集与跨注册中心的研究基金、出版物和研究人员相关联，实现规模化的数据收割、可扩展的图模式创建、机读接口的定制，读取和收割来自Dryad、ANDS（Australian National Data Service）、INSPIRE、ORCID、Figshare等数据库的信息，并且使用Research Graph模式将结果捕获为数据库[39]。

澳大利亚国立计算基础设施（National Computational Infrastructure，NCI）应用RD-Switchboard的软件解决方案，采用元数据构建研究人员、出版物和数据集之间的关联，并用于跟踪和分析链接，能够解答用户的如下问题：“在NCI发表的数据集，被研究期刊文章引用的有多少？被哪些文章引用？”“与给定的数据集有

关联的是哪些研究人员和研究机构？”RD-Switchboard可提供数据链接的数量，通过分析可以确定高价值的数据集，并测度数据集对已出版文献的影响[40]。

德国最大的社会科学基础设施机构GESIS（Leibniz Institut für Sozialwissenschaften）应用RD-Switchboard对110 949种出版物、6 259个数据集、53 914项基金，利用ORCID、CrossRef、DataCite、Dryad、INSPIRE、ANDS等进行关联、扩展和增强，生成研究图，并发布到关联数据云上[41]。

2.4 顶层框架机制

双边协议机制内在差异很大，各种解决方案之间的互操作性非常有限。虽然第三方机制分散地建立了大量的文献与数据关联，但是所有关联互相独立，无法形成完整的链接图，如何在技术和社会层面上克服这一分裂问题是目前的一个挑战。

Scholix（Scholarly Link Exchange）是由RDA提出的一个顶层的互操作性框架。该框架设想了一种通用互联服务，并提出了“多hub”互操作框架的技术指南，用于交换学术文献和数据以及数据集之间的链接信息[41]。

Scholix可以看成是“批发商到批发商”的交换框架，由如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性，且尊重现有的特定社区的实践。Scholix确立的原则具体包括：①信息模型（Scholix学术链接的概念定义）[42]；②链接元数据模式（表示一个Scholix链接的元数据字段集）；③XML和JSON交换格式[42-43]。

Scholexplorer是第一个由Scholix提供的聚合和查询服务，目前包含文献2 000余万篇、数据集5 326万个、双向链接2 623万个，资源来自1.3万个出版商、10个数据中心以及CrossRef、DataCite和 OpenAIRE等。期刊出版商、数据仓储库和数据中心都可以向其提供数据-文献链接信息，还可开发第三方服务，在自己的服务中使用来自Scholix的数据-文献链接信息。Scholix Swagger API允许用户在Scholexplorer索引上运行REST查询命令，以获取匹配给定条件的链接[41,44]。

对上述4类机制进行横向对比，有以下发现。①基于驱动机制，4类机制可分为“基于协议”和“非基于协议”两种模式。双边协议机制、多边模式下的第三方机制和顶层框架机制是协议驱动的协同方式，而事件数据机制是科研活动自主驱动的。②从业务流的规范和标准性来看，自主的事件数据机制通过科研活动，打破行业壁垒，更容易形成行业的规范。③从资产异构性来看，多边模式下的第三方机制和顶层框架机制在社会组织和技术层面试图形成全局业务流的规范和标准。

在实践中，上述各种数据与文献融合机制并不是单一和僵化存在的，而是具有多元和不断演化的特征。以OpenAIRE为例，作为国际性开放获取基础设施和欧洲开放科学云（European Open Science Cloud，EOSC）的可靠支柱，OpenAIRE经历了5个项目周期：第二项目周期的OpenAIREplus通过自动推断数据集和出版物之间的关联关系，生成丰富的数据图表[45]，采用多边模式下的第三方机制，兼具数据聚集和软件解决方案的特点；第三项目周期的OpenAIRE2020演化为顶层框架机制，即遵循Scholix的建议，充当出版物和数据集之间的链接代理[46-48]。

3 学术文献与科学数据融合的服务模式

学术文献与科学数据融合机制只有应用于真实的科研信息环境和服务，才能真正支撑全生命周期的科研活动，发挥提升科研效率和促进科研成果转化的作用。因此，综合目前可见的各类具体文献-数据服务场景、功能实现和应用方式等，提炼出3类基本的学术文献与科学数据融合服务模式。

3.1 数据链接与发现

文献与数据融合的基本服务模式为实现数据链接与发现，目标是提升科研资源的可发现性。链接与发现的前提是科学数据的开放与共享，具体表现为文献与数据在结构和语义层面的关联。

数据链接与发现服务应实现以下粒度的功能：①精确到具体数据，同时将数据链接与发现融入文献的检索和发现过程，如INSPIRE、Elsevier平台等都提供文献与数据的双向检索、发现与关联功能；②深入基于语义内容的文献内实体，例如英国皇家化学学会（Royal Society of Chemistry，RSC）的语义出版项目RSC Semantic Publishing利用化学领域本体抽取和识别文章中的化学物质名称，并将其与专业化学数据库的数据条目相关联，读者通过点击文章中的化合物词汇，

即可链接到ChemSpider网站，并基于化学领域关系，扩展发现其他相关化合物以及资源；③扩展相应的数据引用与分析评价服务，如Web of Science提供数据引用链接，并基于数据引用进行分析评价。

3.2 数据可视化与阅读

数据可视化与数据阅读是对学术论文在线阅读模式的补充和突破：在读者的阅读过程中提供在线可视化数据，并将数据放于文本中，显著提升数据的可用性和易用性。数据可视化和数据阅读的前提是数据可获取，制定一定的规范和质量标准并集成数据阅读工具，实现前后台交互。

3.2.1 图表数据阅读

ActiveCharts.org项目将经济合作与发展组织数据库（OECDiLibrary）中的各种统计数据进行混合、可视化和共享，并集成到期刊论文页面。在读者阅读过程中，阅读器能够重新定向创建数据图表的数据集源[31]。读者不仅可以在文本阅读过程中获得动态数据图表，还可以选择显示/隐藏数据和重新绘制。F1000Research也开发了类似的技术，发布了测试版数据绘图工具，使读者能够重新绘制电子数据表格，并根据需要改变x和y轴。

3.2.2 地理数据可视化

在Elsevier、ScienceDirect的在线阅读界面中，如果论文内容涉及PANGAEA数据库的原始数据，读者会看到一个交互式地图应用程序，该程序将PANGAEA数据集的地理位置可视化，并提供数据记录链接和多种交互功能，例如：可将数据集的地理位置元数据应用于第三方地图（如谷歌地图）；如果涉及DataCite数据，可使用DataCite存储的其他元数据属性（如标题、创建者等）吸收多来源的地理定位数据，允许来自不同数据库的地理空间信息显示在同一地图上，读者因此能够根据论文提到的数据集的地理位置或空间范围来搜索论文。

3.2.3 领域特定数据可视化

蛋白质查看器是ScienceDirect的一个基于Jmol查看器的应用程序，应用于含有蛋白质标识符的论文。它使读者能够浏览论文标记的所有蛋白质模型，并允许读者交互式地探索每一个模型，例如读者可缩放模型、改变视点和背景颜色或以3D模式查看蛋白质结构。支持上述功能的蛋白质结构交互可视化三维模型来自蛋白质结构数据库（Protein Data Bank，PDB）。

3.3 数据在线操作

数据在线操作是结合了文本、原始数据和代码的动态服务。读者通过人机交互，操作文献阅读环境中的数据和代码。其前提是数据可获取及计算流程完整，且数据、工具、环境、代码集成及前后台可交互。数据在线操作能够提高科研的透明度、开放性和可重复性[49]。

在Elsevier的“可执行论文”（Executable Paper）项目中，论文作者可以将可执行的代码嵌入论文，读者则在阅读论文时执行这些代码。可执行论文允许读者在预先指定或交互提供的数据集上运行代码块，产生可验证的结果[50]。

可执行论文的一个变种是“可复制论文”（Reproducible Paper），即以模拟代码为核心的出版物。读者可以通过运行该代码生成数据并执行分析，从而重现整个研究过程和结果。其与可执行论文的主要区别在于，可执行论文依赖经验数据，而可复制论文的代码能够生成数据，因此可复制论文更强调代码的描述[51]。

数据在线操作服务面临两个挑战：首先，所有的资源（代码、图像和数据）需要能够访问；其次，可执行文件不是静态的，而是交互式的、动态的，它需要被托管以便用户能够与文档交互，而用户最好不需要安装额外的软件。

综上所述，这3类服务模式代表了目前学术文献与科学数据融合在具体服务和应用中的基本形态，呈现递进和深化的关系，即从增强泛在发现能力，到强调数据重现，再到强调数据重用。因此，现有的科技信息服务系统在进行服务升级和转型的过程中，应重点以上述服务模式为基础进行开发，发挥学术文献与数据融合的价值。

4 结语

在国内，国家科技图书文献中心（National Science and Technology Library，NSTL）近期推出了外

文科技文献与科学数据的双向整合检索与精准链接服务，在科学资产的互联互通方面迈出可喜的一步。但从服务规模来看，目前实现文献与数据链接的学术资源在NSTL中只占少数；从服务模式来看，NSTL仅采用双向检索发现与精准链接的单一模式，还需细化数据与文献关联的粒度和层次、丰富数据的可视化阅读与分析功能、实现数据的在线重用与复现等；从实现机制来看，NSTL主要通过与专业图书馆合作实现文献与科学数据融合，缺少多边协议和顶层机制，因此尚不具备具有顶层规范、自组织能力特征的自驱动机制，这是制约数据规模扩大、服务模式丰富和深化、科研效率提升的根本性因素。

建议数字图书馆深化与利益相关者的战略合作，设计多边与自驱动的多元融合机制。以NSTL为例：作为国家科技条件保障平台的一部分，应打通与科学数据平台的关联渠道，将不同科研生命周期所产生的数据资产与最终成果相关联；嵌入国家科技文献资源保障体系和网络服务体系，构建面向全社会的新型学术服务；基于现有的数据和论文发表模式与出版渠道合作，在科学信息传播的源头成为出版物和数据集之间的“链接代理”；通过构建专业社区，与论文作者互动，以用户认领的自驱动方式实现数据与文献的关联与共享等。

科研资产包含出版物、研究数据、软件代码、资助信息、机构信息等。只有实现了全域和深层的互联互通，即产生开放科学语义链接合集，科学研究才有可能完全情境化、实现开放和可遍历。这不仅能够催化科研发现，而且对于科研成果监控、研究重现以及信息资源统一评估等具有重要作用。

参考文献

[1] KAUPPINEN T, DE ESPINDOLA G M. Linked open science-communicating, sharing and evaluating data, methods and results for executable papers[J]. Procedia Computer Science, 2011, 4: 726-731.

[2] LinkedScience about[EB/OL]. [2022-11-11]. http://linkedscience.org/about/.

[3] The nature of science and the scientific method[EB/OL]. [2022-12-01]. http://www.geosociety.org/educate/NatureScience.pdf.

[4] GRAY J. A transformed scientific method[EB/OL]. [2022-12-01]. http://languagelog.ldc.upenn.edu/myl/JimGrayOnE-Science.pdf.

[5] Promoting access to public research data for scientific, economic and social development[EB/OL]. [2022-11-14]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.9016&rep=rep1&type=pdf.

[6] 丁培. 科学论文内的科学数据组织和发现研究[J]. 现代情报，2020，40（2）：34-43.

[7] 黄如花，邱春艳. 国外科学数据共享研究综述[J]. 情报资料工作，2013（4）：24-30.

[8] SMIT E. Abelard and Héloise: why data and publications belong together[J]. D-Lib Magazine, 2011, 17（1/2）: 1045.

[9] BALL A, DUKE M. How to track the impact of research data with metrics[EB/OL].[2023-01-16]. https://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_To_Track_Data_Impact.pdf.

[10] COOK R B, VANNAN S K S, MCMURRY B F, et al. Implementation of data citations and persistent identifiers at the ORNL DAAC[J]. Ecological Informatics, 2016, 33: 10-16.

[11] LAGERSTROM J. Measuring the impact of the Hubble space telescope: open data as a catalyst for science[C]//World Library and Information Congress: 76th IFLA, 2010.

[12] SCHOOLER J W. Metascience could rescue the ‘replication crisis’ [J]. Nature, 2014, 515（7525）: 9.

[13] BAKER M. 1, 500 scientists lift the lid on reproducibility[J]. Nature, 2016, 533（7604）: 452-454.

[14] BORGMAN C L. Big data, little data, no data: scholarship in the networked world[M]. Cambridge: MIT Press, 2022.

[15] MARTONE M. San Diego.CA data citation synthesis group: joint declaration of data citation principles[EB/OL]. [2022-11-14]. https://doi.org/10.25490/a97f-egyk.

[16] STARR J, CASTRO E, CROSAS M, et al. Achieving human and machine accessibility of cited data in scholarly publications[J]. PeerJ Computer Science, 2015, 1: e1.

[17] 涂勇，彭洁. 基于DOI技术的科学数据与科技文献融合的研究[J]. 数字图书馆论坛，2007（10）：28-31.

[18] 邱春艳. 期刊文献与科学数据的关联服务研究[J]. 情报资料工作，2014（2）：63-66.

[19] 邱春艳. 科学数据与期刊文献的关联实现研究[J]. 图书馆杂志，2015，34（8）：29-33.

[20] 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.

[21] 卫军朝. 科学文献与科学数据关联实践研究：以Elsevier为例[J]. 国家图书馆学刊，2017，26（3）：93-101.

[22] 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.

[23] 韩涛. 科学数据与科学文献相关性研究：以生物信息学为例[J]. 图书情报知识，2008（3）：42-46.

[24] 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.

[25] 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.

[26] 贺姝祎，魏韧，吴茂春，等. 科技文献与观测数据的关联性在天文领域的应用研究[C]//2014年第五届全国知识组织与知识链接学术交流会，2014：182-191.

[27] 孙巍. 科学数据与科技文献关联发现系统研究与实现[C]//2011年全国知识组织与知识链接学术交流会，2011：186-197.

[28] 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.

[29] 丁楠，黎娇，李文雨泽，等. 基于引用的科学数据评价研究[J]. 图书与情报，2014（5）：95-99.

[30] 黄永文，孙坦，赵瑞雪，等. 科学数据与学术文献关联服务的研究与实现[J]. 图书情报工作，2021，65（23）：116-125.

[31] 李莉，王朝晖. 国外期刊科研数据发表政策分析与比较：以化学类期刊为例[J]. 情报探索，2019（9）：54-57.

[32] D-Lib Magazine[EB/OL]. [2022-09-16]. http://www.dlib.org/dlib/november14/11inbrief.html.

[33] DataCite[EB/OL]. [2022-09-16]. https://www.crossref.org/community/datacite/.

[34] BURTON A，ARYANI A，KOERS H，et al. The scholix framework for interoperability in data-literature information exchange[J]. D-Lib Magazine，2017，23（1/2）.

[35] The RMap project（white paper） [EB/OL]. [2022-09-16]. http://rmap-project.info/rmap/wpcontent/uploads/RMap_Project_Overview_Revised_Final.pdf.

[36] BURTON A，KOERS H，MANGHI P. On bridging data centers and publishers：the data-literature interlinking service[EB/OL]. [2022-09-16]. http://10.1007/978-3-319-24129-6_28.

[37] The DLI Service：an open one-for-all data-literature interlinking service[EB/OL]. [2022-09-16]. https://www.openaire.eu/dliservice.

[38] Publishing data services working group case statement[EB/OL]. [2022-09-16]. https://www.rdalliance.org/filedepot/folder/114?fid=239.

[39] WANG J B，ARYANI A. Graph connections made by RD-switchboard using NCI's metadata[EB/OL]. [2022-09-16]. https://nci.org.au/research/publications/research-articles/graph-connections-made-rd-switchboard-using-ncis-metadata.

[40] ARYANI A，POBLET M，UNSWORTH K，et al. A Research Graph dataset for connecting research data repositories using RD-Switchboard[J]. Scientific Data，2018，5：180099.

[41] Discovering research data links via GESIS LOD[EB/OL]. [2022-09-16]. https://scholexplorer.openaire.eu/#/about.

[42] BURTON A，KOERS H. The scholix framework for interoperability in data-literature information exchange[EB/OL]. [2022-09-16]. http://mirror.dlib.org/dlib/january17/burton/01burton.html.

[43] Enabling researchers to make their data count[EB/OL]. [2022-09-16]. https://globaljournals.org/GJCST_Volume19/1-Enabling-Researchers-to-Make.pdf.

[44] Scholix metadata schema for exchange of scholarly communication[EB/OL]. [2022-09-16]. https://www.rd-alliance.org/group/rdawds-scholarly-link-exchange-scholix-wg/outcomes/scholix-metadata-schema-exchange-scholarly.

[45] MANGHI P，BOLIKOWSKI L，MANOLD N，et al. OpenAIREplus：the European scholarly communication data infrastructure[J]. D-Lib Magazine，2012，18（9/10）.

[46] MANGHI P，MANOLA N，HORSTMANN W，et al. An infrastructure for managing EC funded research output：the OpenAIRE project[J]. International Journal on Grey Literature，2010（6）：31-40.

[47] ARTINI M，ATZORI C，BARDI A，et al. The OpenAIRE literature broker service for institutional repositories[J]. D-Lib Magazine，2015，21（11/12）.

[48] 赵展一，黄金霞. 开放科学基础设施的信息资源建设模式分析[J]. 图书馆建设，2021（3）：46-55.

[49] LASSER J. Creating an executable paper is a journey through Open Science[J]. Communications Physics，2020，3：143.

[50] Executable Papers-improving the article format in computer science[EB/OL]. [2022-09-16]. https://www.journals.elsevier.com/the-journal-of-logic-and-algebraic-programming/news/introducing-executable-papers.

[51] Reproducible paper template[EB/OL]. [2022-09-16]. https://savannah.nongnu.org/projects/reproduce/.

作者简介

白海燕，女，硕士，研究馆员，研究方向：数字图书馆、信息组织、关联数据、学术发现，E-mail：bhy@istic.ac.cn。

Fusion Mechanism of Academic Literature and Scientific Data

BAI HaiYan
(Institute of Scientific and Technical Information of China, Beijing 100038, P. R. China)

Abstract: Aiming at the interconnection between academic literature and scientific data of the subject of scientific research assets, this paper studies the specific model of the fusion of academic literature and scientific data to improve scientific research efficiency, and discusses the mechanism of data and literature fusion for open, reliable, and scalable scientific research information infrastructure. It provides new ideas and references for the transformation and innovation of traditional literature information service towards data-intensive scientific research. By studying the best practices and typical cases of different stakeholders in literature and scientific data fusion, based on the analysis, comparison, and summary of driving factors, technology implementations, industry norms, and actual effects, four implementation mechanisms under the ecological environment of scientific research sharing and three service models of literature and data fusion are extracted, and their scientific research efficiency is evaluated.

Keywords: Open Science; Scientific Data; Open Data; Information Service

（责任编辑：王玮）