PARIS 原则：开放协作环境下科学数据的可用性*

沈志宏[1]，张晓林[2]，郑晓欢[3,4]
[1](中国科学院计算机网络信息中心 北京 100083)
[2](中国科学院文献情报中心 北京 100190)
[3](中国科学院大学 北京 100049)
[4](中国科学院科学传播局 北京 100864)

摘要:
科学数据利用的需求日益迫切，且在“第四范式”、“融合科学”等新型科研范式带来的开放协作环境下呈现出跨边界、端到端、动态性和协作化的特征。作为“数据仓储时代”的产物，FAIR、TRUST 原则已无法为开放协作环境下科学数据的高效利用提供深入的指导。本文详细分析了科学数据利用的典型场景，提出开放协作环境下促进科学数据利用的 PARIS 原则：可处理（Processable）、可问答（Askable）、可信赖（Reliable）、可联合（Incorporable）与可供给（Suppliable），并重点分析了 PARIS 原则对科学数据可用性的促进作用。最后，本文探讨了实现 PARIS 原则可参考的技术路径。作为 FAIR、TRUST 原则的有益扩展，期望 PARIS 原则能有效提升科学数据的可用性。

关键词: FAIR 原则；TRUST 原则；PARIS 原则；数据利用；数据可用性
中图分类号: TP393
doi:10.11959/j.issn.2096-0271.2023013

From FAIR to PARIS: Improving the Usability of Scientific Data in the Open Collaborative Environment

Abstract:
The demand for scientific data utilization is increasingly urgent, and in the open environment brought by the new scientific research paradigms such as "Fourth Paradigm" and "Convergence Science", the data utilization shows the characteristics of cross-the-boundary, end-to-end, dynamic and collaborative. As products of the "era of data repository", the FAIR and TRUST principles can no longer provide in-depth guidance for the efficient use of scientific data in the open environment. This paper analyzes the typical scenarios of scientific data utilization in detail. Then, it presents the PARIS principles to promote scientific data utilization: Processable, Askable, Reliable, Incorporable, and Suppliable. Finally, this paper gives a technical practice path that the PARIS principles can refer to. As beneficial extensions of the FAIR and TRUST principles, it is expected that the PARIS principles can effectively improve the usability of scientific data.

Key words:
FAIR principles;TRUST principles;PARIS principles;data utilization;data usability

* 收稿日期：2022-09-29
通信作者：bluejoe@cnic.cn
基金项目：国家重点研发计划项目“面向国家科学数据中心的基础软件栈及系统”（No. 2021YFF0704200），中国科学院“十四五”网信专项工程建设项目“科学大数据工程（三期）”（No. CAS-WX2022GC-02）

1. 科学数据利用的需求与挑战

1.1 科学数据与科学数据中心

科学数据（Scientific Data，又作科研数据，Research Data）主要包括在自然科学、工程技术科学等领域，通过基础研究、应用研究、试验开发等产生的数据，以及通过观测监测、考察调查、检验检测等方式取得并用于科学研究活动的原始数据及其衍生数据[1]。

国际上，欧美等发达国家已经将科学数据的持续积累和开放利用提高到了国家战略的高度进行部署，并将国家科学数据中心建设作为科学数据管理的重要手段[2,3]。为促进科学数据的汇交整合和开放共享，我国加强了科学数据中心的建设。2019 年 6 月，科技部、财政部在原有科学数据类国家平台的基础上进一步优化调整形成了 20 个国家科学数据中心，涉及地球系统、人口健康、农业、林业、气象、海洋等多个领域[4,5]。同年，中国科学院启动了科学数据中心体系建设，初步建成了由总中心、18 个学科中心和 13 个所级中心组成的院科学数据中心体系[6]。

多元化科学数据中心生态正在迅速形成。大中型科学数据中心如：国家科学数据中心、学科数据中心、省部级科学数据中心[7,8,9]等，小型科学数据中心如：研究所数据中心、高校数据中心[10,11,12]、企业数据中心、实验室数据中心[13]等，微型科学数据中心如团队科学数据中心[14]、个人科学数据中心[15,16]等。此外，大型科学数据中心往往还按照学科或者区域下设分中心（分部）[17,18]。这些科学数据中心势必形成复杂的生态，共同推动科学数据的共享，为科技创新发挥更大作用。

1.2 FAIR、TRUST 原则与科学数据共享

随着数据开放运动的不断深入，科学数据的共享得到了较大的发展。2016 年，FAIR 原则被正式确定为科学数据管理的指导方针[19]。FAIR 原则规定了数据的开放共享需要满足可发现(Findable)、可访问(Accessible)、可互操作(Interoperable)、可重用(Reusable)等四个方面的要求。类似的，TRUST 原则从透明性（Transparency）、负责任（Responsibility）、用户导向（User Focus）、可持续性（Sustainability）、技术（Technology）等五个方面定义了数据仓储（Data Repository）的可信任能力[20]。

欧盟、荷兰、澳大利亚等重视 FAIR 原则在数据密集型科学数据管理中的重要作用，在人文社科、环境科学、生命科学等领域开展了应用实践[21,22]。越来越多的科学数据中心遵循 FAIR 原则将数据提供开放共享，并在 TRUST 原则的指导下构建可信的数据仓储。作为例子，国家青藏高原科学数据中心收集并发布青藏高原及周边地区的科学数据集 4600 多个，不断研发新技术实践 FAIR 原则，采用国际标准提供数据引用方式和数据关联文献引用方式，支持数据出版，开发在线大数据分析、模型应用等功能[23]。截止 2022 年 3 月，该中心累计页面访问量超过 1.5 亿，月均下载量达 50TB。同时，国家青藏高原科学数据中心也成为国内首个通过 Nature 数据期刊 Scientific Data 认证的数据仓储中心，大大提高了数据中心的影响力和权威性。

1.3 科学数据从共享到利用

数据只有动起来、用起来才能产生价值。全球著名咨询公司 Frost&Sullivan 在 2019 年发布的《2025 年世界顶级全球大趋势及其对商业、社会和文化的影响》一文中提到，数据支撑着未来，90%的变革性转变严重依赖数据的流通和使用[24]。

在大数据时代，科学发现越来越依赖于对海量数据的集成和分析，科学研究水平不仅仅取决于科研人员的水平，也越来越多地取决于对数据的积累以及将数据转换为信息和知识的能力[2]。如，北京正负电子对撞机北京谱仪国际合作组利用国家高能物理科学数据中心存

储的北京谱仪 III（BESIII）数据完成世界上最精确的正反科西超子衰变不对称性测量。该结果证实了一种新方法，它为研究物质和反物质之间的差异提供了极其灵敏的探针[25]。再如，国家生态科学数据中心兰州大学分中心赵长明等利用长期定位土壤呼吸观测数据，在土壤呼吸与土壤温度间滞后性的研究方面取得重要成果，强调了光合产物在植物和生物土壤结皮中的转运在土壤呼吸和土壤温度间滞后模式调控中的重要作用[26]。

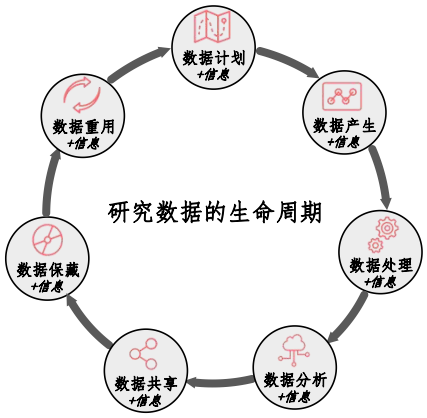

图 1 研究数据生命周期[27]

对科学数据进行处理、分析、挖掘与可视化等消费操作，将数据转换为信息和知识的过程，即数据利用（Data Utilization）。图 1 示出科学数据生命周期的核心阶段，包括：数据产生→数据处理→数据分析→数据共享→数据保藏→数据重用。其中，数据利用分别发生在数据产生之后、数据保藏之前，以及数据重用的阶段。

为深入了解科学数据共享与利用的现状，本文选取了 20 个国家科学数据中心的微信公众号以及科技部主办的微信公众号“锐共享”，对其中的文章内容进行分析，分别统计了 2021 年 7 月至 2022 年 12 月期间“数据发布”、“数据利用”两类文章的发表情况。图 2 示出 21 个公众号在两类文章的数量对比，图 3 示出“锐共享”公众号文章发表数量趋势。

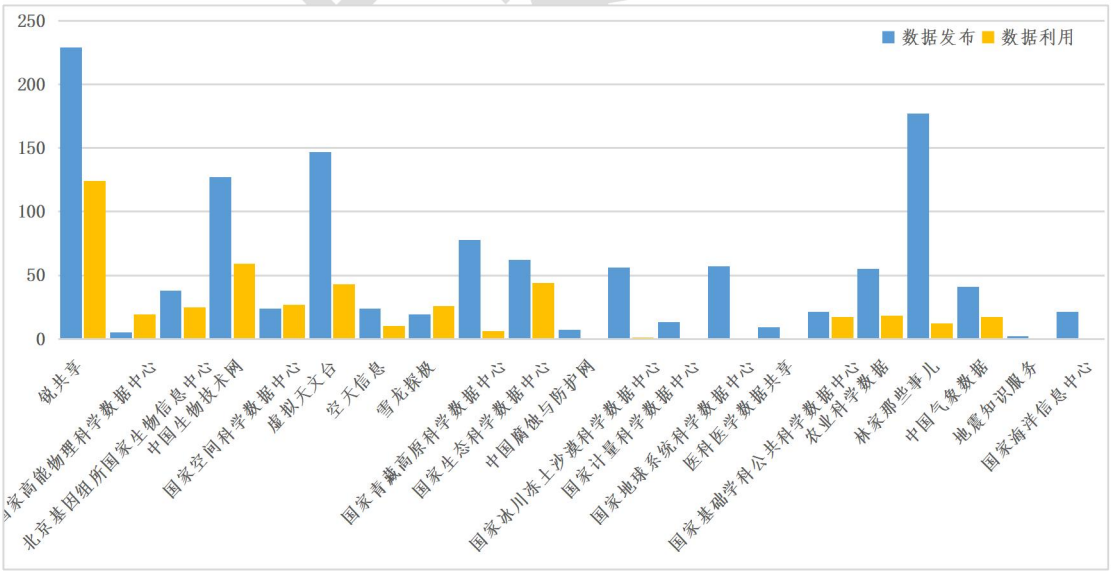

图 2 国家科学数据中心公众号文章统计

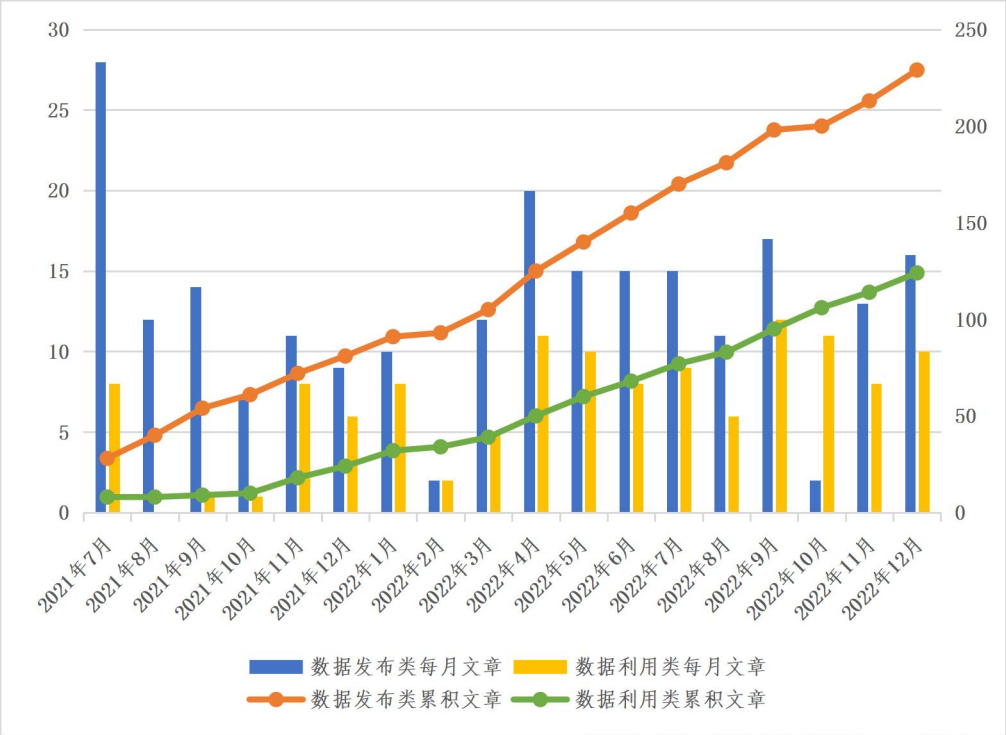

图 3 “锐共享”公众号文章发表数量趋势

统计发现，现有科学数据中心在科学数据的发布共享与数据利用方面都呈现出较好的发展势头，但相比而言，目前各科学数据中心在落实 FAIR 原则方面其工作重点更多的还是侧重于数据发布与共享（即 F、A 和 I 方面），利用方面（即 R 方面）稍显不足。科学数据中心的数据服务仍然以“数据仓储”服务为核心，以“数据汇聚/汇交”为主要的数据共享模式，发布出来的数据集往往还不能满足分析可用（Analysis Ready），甚至是在线分析可用（Online Analysis Ready），科学数据的利用环境与仓储环境还存在着较大的脱节，科学数据的利用能力和水平还存在着急需改进的地方。

1.4 开放协作环境下科学数据的利用

“第四范式(Fourth Paradigm，又称数据密集型科学发现范式)”、“融合科学（Convergence Science）”、“关联科学(Linked Science)”等新型科研范式对科学数据的共享和利用提出要求。“第四范式”强调从数据中心保存的海量的、由各种设备收集到的数据中查找所需数据进行分析研究[28]；“融合科学”强调所有学科之间的数据开放和共享、科研全流程的数据开放和共享，并强调对解决重大经济社会问题相关的全景式数据的开放和共享[29]；“关联科学”则强调科学数据之间的关联性，提出一种实现科学资产互联的方法，以支持透明的、可重复的和跨学科的研究[30]。

可以看出，随着各种新型科研范式的开展，科学数据的利用被置于一个“多主体、多要素、全景式”的开放协作环境。（1）多主体：多元化科学数据中心已形成日益繁荣的生态，彼此竞争和合作；（2）多要素：除了传统的科学数据资源要素，科学数据软件及服务、科学数据团队等要素也参与了科学数据的利用；（3）全景式：科学数据的利用不再局限于独立的视角，而以最终任务为目标，有效整合交叉学科机构各类数据资源与服务。

作为例子，国家天文科学数据中心牛晨辉等在处理 FAST 数据的过程中，发现 2019 年 5 月 20 日的数据存在重复的高色散脉冲[31]。基于这一发现，团队通过与美国甚大阵列望远镜合作，在 2020 年 7 月完成亚角秒量级的精确定位，并探测到了一颗与之对应的致密的持续射电源（PRS）。随后通过美国帕洛玛 200 英寸望远镜和凯克望远镜、加拿大-法国-夏威夷

望远镜和日本斯巴鲁近红外光学望远镜确定了 FRB20190520B 的宿主星系和红移，推导出其宿主星系贡献了总色散值的 80%，为目前已知所有快速射电暴源中最高。进一步结合散射特征，团队提出宿主星系的色散主要来自邻近 FRB 爆发源的区域，该区域高电子密度导致的高色散值也使得 FRB 20190520B 远远偏离经典的色散与红移关系。在这个例子中，我们可以看到综合利用到 FAST 数据、美国甚大阵列望远镜、美国帕洛玛 200 英寸望远镜和凯克望远镜，加拿大-法国-夏威夷望远镜和日本斯巴鲁近红外光学望远镜的数据。

为有力应对新型科研范式的应用场景，各科学数据中心纷纷形成相互联合的态势，如：2021 年 7 月，国家高能物理科学数据中心、国家空间科学数据中心、国家天文科学数据中心签订战略合作协议，并发布“高能物理-空间科学-天文学”首批联合主题数据目录[32]，从而满足多信使天文学使用探测电磁波、引力波、中微子、宇宙线等多种技术手段对天体进行观测的需求[33]。再如，2021 年 8 月，国家农业科学数据中心联合国家地球系统科学数据中心、国家林业和草原科学数据中心、国家气象科学数据中心，建立黄河流域生态保护与高质量发展专题库，为促进黄河流域生态保护与高质量发展国家战略实施、保障黄河长治久安提供全方位科技支撑[34]。不仅如此，科学数据中心内部也存在各分中心“联合”的态势，如：国家天文科学数据中心按照根据中国“虚拟天文台”的思路整合了国家天文台、紫金山天文台、上海天文台、云南天文台、新疆天文台等天文台的天文科学数据及其他类型的天文数据，形成了物理上分散、逻辑上统一的覆盖天文科学数据全生命周期的管理与开放共享平台[35,36]。

在“多主体、多要素、全景式”的开放协作环境下，科学数据的利用不再是传统手工的、单点作坊的方式，而是逐渐呈现出“跨边界、端到端、动态性和协作化”的特征：

（1）跨边界：科学数据的利用不再发生在单一数据中心、单一团队内，往往是跨系统、跨中心、跨领域的；

（2）端到端：科学数据的利用往往是“端到端”的无人工交互、弱人工交互的自动化、智能化的方式，科学数据需要技术协议层面的高度无缝衔接；

（3）动态性：科学数据的利用、流向不再是静态的、预设完好的，而是针对复杂利用场景，在线的、按需的、动态的执行的；

（4）协作化：科学数据的利用不再是针对某个单一的数据源、数据中心进行的，而是通过综合调度多个数据中心的资源，完成一个复杂场景的任务。

FAIR 原则的推出，在科学数据的共享和利用过程中起到了较大的指导作用。但应注意到，FAIR、TRUST 等原则更多关注于科学数据的开放与共享，实现了科学数据的“不可见→可见→可用”，但仍无法有效解决科学数据及其服务目前普遍存在着的分布式、孤岛化、差异化、权益约束等问题，无法实现科学数据的“可用→可协作”，无法有效满足“第四范式”、“融合科学”等新型科研范式提出的“跨边界、端到端、动态性和协作化”的科学数据利用需求，科学数据面临的主要矛盾已转化为分布式孤岛化异构科学数据资源与新型范式对科学数据高质量供给（按需、高效、可信供给）需求之间的矛盾。

2. 科学数据利用的场景分析

如上所述，跨边界、端到端、动态性和协作化的科学数据利用的需求广泛存在，本节梳理出不同环境下几类典型的科学数据利用场景：离线与在线消费模式、数据即服务、算法找数据、数据找算法、数据管道、数据协同分析。

2.1 离线与在线消费

传统的科学数据共享方案中，数据利用通常采用的是一种“下载—解释—使用（Download-Interpret-Load，简称 DIL）”的离线方式，即：

（1）用户通过浏览数据网站，获取到数据集地址，通过 HTTP 或者 FTP 等方式下载该数据集，获得一份拷贝；

（2）用户结合数据集的描述信息，对下载的数据集结构、内容进行理解和确认；

（3）用户启动一个消费程序，如 Excel，加载数据并进行进一步分析应用。

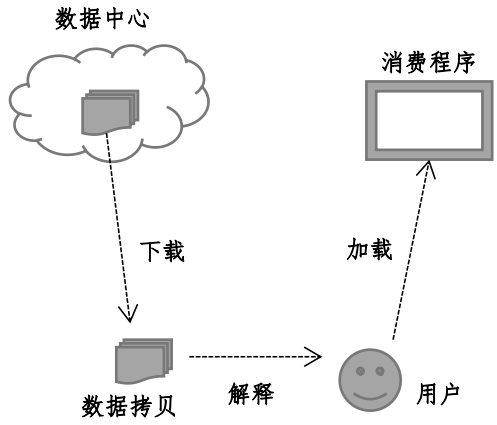

图 4 “下载—解释—加载”模式

图 4 示出“下载—解释—加载”模式，这种模式的缺点是需要人工干预，效率不足，在面对复杂动态任务的场景无法做到及时有效。另外一种模式为“在线消费（Online Consumption，简称 OC）”模式，如图 5 所示，这种模式下，消费程序按照固定的协议和格式直接接收数据并生成消费结果，消费过程中无需人工干预。

图 5 在线消费模式

近年来，数据出版成为一种新型的科学数据共享方式，数据出版可使数据达到可引用和追溯的状态, 核心内容是为数据引用提供标准的数据引用格式和永久访问地址[37]。数据仓储是一种常见的数据出版方式，它往往以数据文件包的方式提供数据下载服务。这种情况下的数据利用，即 DIL 模式。

2.2 数据即服务

图 4 展现的是一种传统的“移动数据”的消费模式，即消费程序不动，将数据从发布端迁移到消费端，其特点可概括为“给程序喂数据”、“程序不动数据动”。对于海量规模的科学数据集，由于数据迁移和传输的成本较高，往往需要采用一种“数据即服务（Data As A Service, DAAS）”的形式提供数据服务。在这种形式下，“科学数据集（Dataset）”和“数据服务”统一，可称作在线数据集（Online Dataset），在线数据集需要配套程序执行引擎或者容器，接受用户设定或提交消费程序或逻辑，并输出消费结果，这种方式为“移动计算”，特点是“给数据喂程序”、“数据不动程序动”，消费模式如图 6 所示。

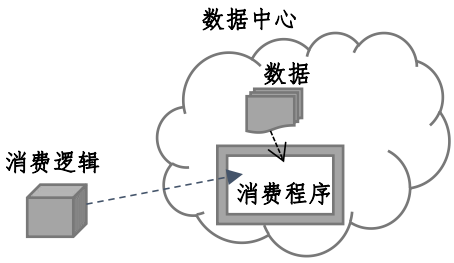

图 6 在线数据集模式

“数据即服务”一个典型的例子即 Google BigQuery[38]，BigQuery 为用户提供了一个在线分析的环境，允许用户选择数据集，提交一个 SQL 查询语句，从而获取到关心的查询结果。在这里，程序逻辑是 SQL 结构化查询，另外一种常见的程序逻辑是类似于 Map-Reduce 的大数据操作，如提交一段脚本，要求返回符合条件的数据的某列的总和，这种情况下，“移动计算”要比“移动数据”要经济的多。

2.3 算法找数据、数据找算法

在数据丰富、算法贫乏的时代，数据消费过程中主动权在于消费程序，因此需要根据消费程序（算法）的输入输出格式要求来准备数据，即“算法找数据”，这个过程中数据的预处理往往是一项重要的准备工作。

随着机器学习、深度学习、神经网络，以及容器技术、微服务技术的发展，根据数据以及任务来找算法已经成为可能。如：针对一幅在植物园拍摄的植物照片，格式为 JPEG，需要识别其中的植物物种，甚至识别其中的生态场景。这种模式即“数据找算法”，该模式如图 7 所示。这种场景需要数据提供详细的信息（如：JPEG 格式、植物照片），任务的定义（如：识别植物物种），同时需要一个类似于“算法市场”，其中每个算法具有更为详细的描述（如：用途、输入数据的约束、编程语言、执行环境等）。

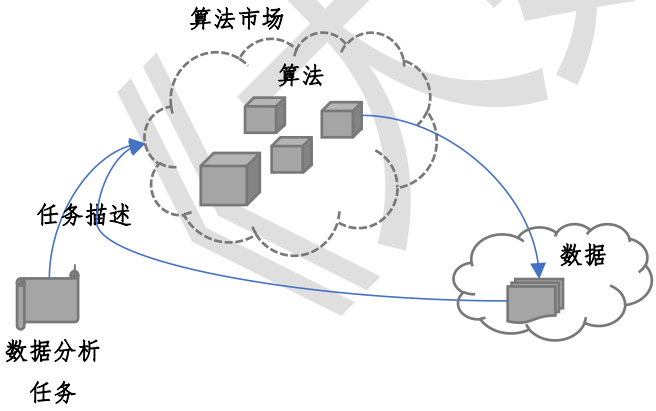

图 7 “数据找算法”模式

2.4 数据管道

数据的消费过程经常是复杂的、多轮迭代的，这种情况下需要多个消费程序形成数据管道（Data Pipe），这样消费程序就可以按需串接起来，即：消费程序 A 的输出可以作为消费程序 B 的输入，从而完成复杂的、个性化数据使用需求。这种模式即数据管道，该模式如图 8 所示。

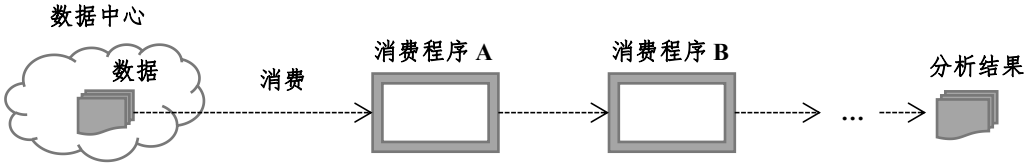

图 8 数据管道

数据管道广泛存在于科学数据的分析利用中。以 GWAC(The Ground-based Wide-Angle Camera array)为例，GWAC 是中法合作伽玛暴探测天文卫星 SVOM 的关键地面设备，一个 GWAC 相机每 15 秒钟产生一个大小为 32MB 的天区图，图像的点源提取和接下来的光变曲线处理流程应该在一帧的 15 秒内快速处理完[39]。这个实时处理过程实际上构建了一个典型的数据管道：天区图采集→图像处理→点源提取→交叉证认→光变曲线处理。

2.5 数据协同分析

在面向复杂任务时，由于数据天然的分布性，往往需要多个数据资源、数据服务协同完成。例如对一个区域的机构与企业数据、人口数据、市场数据、治安数据等的融汇处理。这种模式即数据协同分析，该模式如图 9 所示。

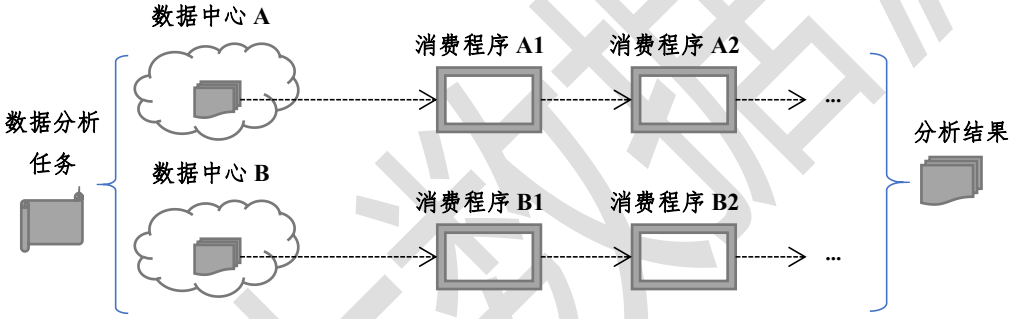

图 9 数据协同分析

前文提到的利用 FAST 数据发现快速射电暴事件[31]就是一个典型的数据协同分析的例子。其中的数据服务涉及到 FAST、美国甚大阵列望远镜、美国帕洛玛 200 英寸望远镜和凯克望远镜，加拿大-法国-夏威夷望远镜和日本斯巴鲁近红外光学望远镜。

3. PARIS 原则

为了有效提升开放协作环境下科学数据的可用性，促进科学数据的“可用→可协作”，本文提出一套针对科学数据的 PARIS 原则，即：可处理（Processable）、可问答（Askable）、可信赖（Reliable）、可联合（Incorporable）与可供给（Suppliable）。

3.1 可处理（Processable）

可处理原则（或者称为可计算原则）指的是科学数据其内容可以被计算机进行处理和分析。建议如下：

数值可计算。如：针对数值、日期等类型的数据值，建议采用数值类型（整数、浮点数、日期类型等），方便算术运算。数值类型的属性，宜在元数据中标注其单位（如：米、千克等）；

语义可计算。如：针对物种分类、地区机构、关系等属性的值，尽量采用公认的语义词汇。此外，建议采用公认的语义词汇作为数据（元数据）的属性名；

采用通用的数据格式。如：图片采用 JPEG/PNG 等格式；

具有适用的处理程序。科学数据可以交付给某个公认的处理程序（如：Excel、Matlab、Python 等）消费；

可处理程度越高越好。数据的格式和结构不同，其可处理的程度也具有差别。例如，针对一张图片进行人脸识别，将识别的结果输出到一个包含人脸坐标的 CSV 文件，比直接输出一张包含画框标注的图片，更容易被计算机处理。

3.2 可问答（Askable）

可问答原则要求科学数据不是一种仅供下载和在线浏览的静态资源，而是以一种“活的”服务的形式存在，可以接受用户提出的问题并给出答案。建议如下：

可在线问答。大规模的数据集建议包装成“在线数据集”形式，且采用某种共识的访问和调用协议，在线数据服务具备明确的调用协议、接口说明，以及其接受的消费逻辑，产出结果的格式描述；

支持静默问答方式。在线数据集的服务尽量采用静默方式，即无人工交互或弱人工交互方式；

可持续问答。支持会话和状态保持。

3.3 可信赖（Reliable）

可信赖原则要求科学数据在获取和消费过程中得到信赖保证。建议如下：

数据使用许可。发布明确的数据使用许可协议，帮助用户全面、快速了解数据的使用方式及其限制，规范用户使用行为，保障数据作者合法权益；

数据安全可靠。根据相关法律法规及规范性文件的要求，对数据进行分级分类管理，遵守法律法规和科研伦理，确保数据安全可控；

数据拥有可信赖。开放数据使用的同时，限制对数据的全量和大量获取请求，最大限度保障数据流失；

服务输入可信赖。在线数据集场景中，针对用户提交的消费逻辑，采用必要的手段（如：沙箱技术、代码审查等）进行限制，以保证数据和服务系统不被侵害；

数据访问可信赖。可针对用户身份、数据访问频次等信息，实现数据访问的限制。

3.4 可联合（Incorporable）

可联合原则指的是科学数据彼此可以合作，并非孤立。FAIR 所提出的可互操作原则重在强调机器对数据的理解，可联合原则则更强调机器与机器之间的联合、合作。建议如下：

数据可关联。发布数据的同时，发布数据的内部关联和外部链接，如：针对关系型数据，可描述其外键关联。针对图数据，可采用边（edge）描述数据之间的关联。宜采用 RDF 图数据模型，使用 RDF link 描述数据之间的关联；

数据可溯源。对科学数据的溯源信息进行描述，如：记录数据集的创建过程、创建者、数据生成设备、数据处理流程等信息。准确且丰富的机器可读溯源信息可为研究人员或代理计算机评估数据集提供凭证和支撑；

数据可互补。科学数据的描述可利于便捷的水平方向、垂直方向的数据联合。如：数据集的描述可提供互补属性，如：不同的学科分布、不同的时空分布、不同的属性等；

服务可联动。科学数据服务具备可联合的能力，如：科学数据服务提供机器可理解的服务调用协议、输入输出描述等。

3.5 可供给（Suppliable）

可供给原则指的是科学数据可以作为一种资源对外供给，同时可以提供给后续数据服务进一步使用，满足“算法找数据、数据找算法”以及“数据管道”的场景。建议如下：

提供数据目录；

提供面向供给的元数据。包括数据的覆盖面、产生频率、数据精度、数据加工级别等信息；

面向数据消费的供给。提供开放的数据消费协议、接口说明；

提供版本化供给。针对数据资源的不同版本产品，应描述其批次和版本信息；

提供流式供给。数据资源的传输尽量满足供给链的要求，如：科学数据与服务的输出的格式应尽量适合流式（Streaming）传输，这与传统的基于“完整文件包”的利用形式不同；

提供可靠供给。提供合理的镜像、副本，从而满足动态供给的需求；

针对供给的计费体系。针对科学数据的使用进行统计和计费。

3.6 PARIS 原则与科学数据可用性

PARIS 原则可实现科学数据的“可用→可协作”，从协作方面丰富了科学数据的可用性（Data Usability）的内涵。Peter Bloland 等针对免疫数据将数据可用性定义为：关联性、高效性、全面性、及时性、完整性与一致性[40]；H. PRINS 结合医疗数据，从数据源的收集、用于记录的管理规程、选择数据记录的初衷、以及与论文数据的比较几个方面来衡量可用性[41]；空间数据可用性研讨会认为，数据可用性涉及五个要素：推广、质量、软件与工具、人类理解与认知，以及应用[42]；李建中等从数据的一致性、完整性、精确性、时效性、实体统一性等五个方面定义大数据的可用性[43]。本文按照数据利用过程中从用户到科学数据的不同纵深，将数据可用性分解成四个层次：数据可获取、政策可容许、来源可信赖、技术可处理，如图 10 所示。

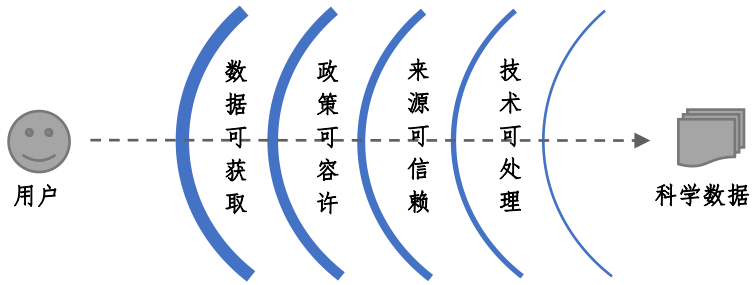

图 10 科学数据利用过程的不同纵深

（1）数据可获取。包括数据源的可发现、可访问、可理解，主要通过访问协议、以及关于数据文件的元数据描述来判断。

（2）政策可容许。这不仅指数据文件本身是否授权使用（数据文件许可），还包括数据采集、记录、处理、传播等是否符合相关法律，尤其是个人信息以及“敏感数据”的保护。使用者需要检验数据是否符合法律要求，还要考虑后续对这些数据的使用本身是否符合法律要求或者需要什么法律程序。

（3）来源可信赖。包括数据仓储的可信赖与数据内容的可靠性。前者即一个机构或领域的数据中心所提供的数据服务能否有效地支持人们的使用；后者涉及数据内容本身针对具体研究问题的有效性和可靠性的判断，包括但不限于数据处理是否符合标准、准确、精确等。多数情况下这需要通过对数据采集、处理、计算等的方法、参数、工具等的描述来判断，可能需要数据格式、数据溯源信息等来确定。

（4）技术可处理。即使用者能否有效调用、配置合适的工具和方法来进行所需要的数据处理，往往需要研究协议、方法的配置等有明确的具体的甚至是计算机可读的描述，以及工具和方法使用授权。

FAIR、TRUST 及 PARIS 原则从不同的方面促进了科学数据的可用性，满足矩阵如表 1 所示。其中，FAIR 原则为数据可获取方面提供了有力保障、同时为政策可容许、来源可信赖、技术可处理方面都提供了部分保障；TRUST 原则为来源可信赖方面提供了针对数据仓储的可信赖保障；PARIS 原则是对 FAIR、TRUST 原则的有益补充，主要针对开放协作环境提出，实现科学数据的“可用→可协作”。

表 1 PARIS 原则与科学数据可用性

	FAIR 原则				TRUST 原则	PARIS 原则				
	F	A	I	R		P	A	R	I	S
数据可获取	√	√	√	√						√
政策可容许				√				√		
来源可信赖				√	√			√		
技术可处理			√	√		√	√		√	√

4. PARIS 实现技术探讨

本节针对 PARIS 的实现技术进行探讨，结合已有的相关实现技术分析，提出 PARIS 核心服务设想，并进一步结合该设想给出 PARIS 节点、PARIS 网络的技术实现路径。

4.1 相关实现技术参考

在基于 PARIS 原则实现数据利用技术框架时，有一些成熟的机制与技术可供参考和使用，如表 2 所示。

表 2 PARIS 原则相关实现技术

PARIS 原则	相关实现技术
可处理	数据格式协议、元数据等
可问答	函数式编程、在线 Notebook 等
可信赖	访问控制、安全多方计算、可信计算等
可联合	关联数据、联邦学习[44]、数据网格、数据经纬等
可供给	科学工作流[45]、大数据流水线等

基于联邦学习，可以实现科学数据的“可问答”和“可联合”。联邦学习自推出以来，已在工业、医疗、金融等场景中得到广泛使用。但这种方式目前仅适用于机器学习，其模型相对固定，各节点之间交换的是模型的参数而非分析结果，因此尚不支持大规模数据的跨中心联动。

数据网格的发展比较早，它实现了异构科学数据的一体化访问，可以有效提升科学数据中心的跨域数据“可联合”和“可供给”能力。数据网格(Data Grid)的概念来自网格(Grid)，网格技术的研究目标是实现网络虚拟环境下高性能资源的共享和协同工作，以解决一致使用各种分散资源的问题[46]。数据网格是以命名的透明性、位置的透明性、协议的透明性、时间的透明性为目标, 建立一个分布海量数据的一体化网格数据访问、存储、传输、管理与服务架构和环境。目前有 Globus 数据网格、欧洲数据网格、地球系统网格[47,48,49]等。

另外一个相关的技术是数据经纬（Data Fabric，又译作数据编织），数据经纬概念在 2000 年首先被 Forrester 提出[50]，2016 年 Forrester Wave 中增加了 Big Data Fabric 类别。从 2019 年

数据经纬开始入选 Gartner 各年度的技术趋势[51]。数据经纬在数据的发现、语义互操作、智能访问协同方面，可以较大程度的提升数据的利用水平。

4.2 PARIS 核心服务设想

PARIS 原则有利于促进科学数据中心彼此协作，从而实现一些面向应用的“协作式”创新性服务，如：数据关联网络、大文件系统、大数据库、大数据流、联邦分析等。

数据关联网络：通过采用关联数据规范，将分布的数据记录发布成统一的格式，同时提供统一的访问协议，数据之间彼此关联，该服务为应用提供一张完整的语义网络；

大文件系统：面向分布于多个科学数据中心的文件（或者对象），向应用提供一张逻辑上完整的文件系统视图。客户端可采用类似于连接 HDFS 分布式文件系统的方式进行数据消费，区别在于 HDFS 是局域的，而大文件系统是跨域的；

大数据库：面向分布于多个科学数据中心的结构化数据（关系数据库、图数据库、文档数据库、KV 数据库等）向应用提供一张逻辑上完整的数据库视图，可以采用统一的、智能的查询语句实现跨库数据检索和分析。客户端可采用类似于 MongoDB 分布式数据库的方式进行数据消费，区别在于 MongoDB 是局域的，而大数据库是跨域的。广义的大数据库可涵盖结构化数据以及半结构化、非结构化数据，采用一种统一的数据模型实现数据资源封装；

大数据流：数据从产生到处理、转换、以及分析，即数据流。大数据流则可能跨越多个科学数据中心节点，涵盖到传感器数据采集、数据校验、网络传输、分中心汇聚、大数据处理转换、模型计算分析等多个流程，面向应用形成一个虚拟的数据供应链路；

联邦分析：该服务为应用提供一张透明的、相对完整的分析能力调度网络，根据应用的需求，对输入的分析任务进行智能拆解、路径编排和执行。

4.3 PARIS 节点设计

可以将满足 PARIS 原则的数据中心节点理解成一个 PARIS 节点，它具备如下能力：

（1）维护一套本地的数据目录；

（2）维护一套本地的服务目录；

（3）开放标准化的访问接口，满足数据关联网络、大文件系统、大数据库的调用请求；

（4）开放标准化的服务接口，满足大数据流、联邦分析的调用请求。

针对以上设想，本文给出一个节点设计参考方案，如图 11 所示。

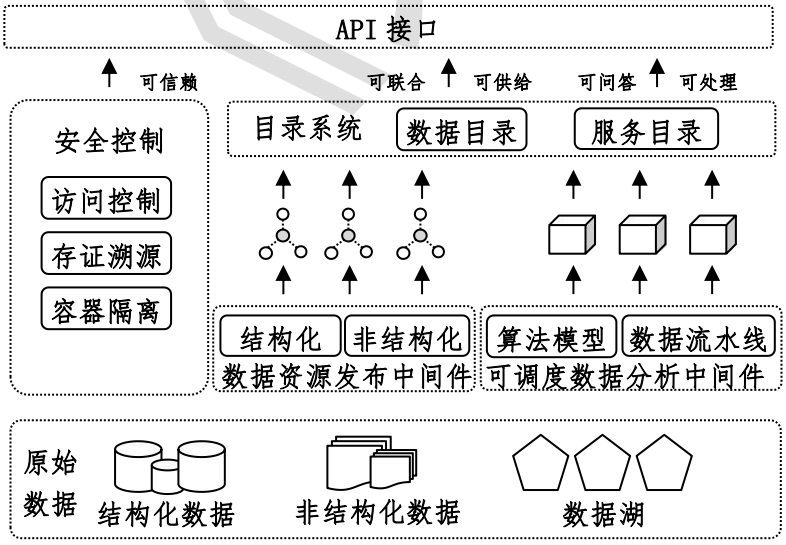

图 11 PARIS 节点设计参考方案

可以看出，在该设计方案中，PARIS 节点中包含几个关键组件：数据资源发布中间件、可调度数据分析中间件和安全控制系统。其中，作为可调度数据分析中间件的核心，数据流水线以算子和流水线的方式实现了对数据分析服务的抽象，实现了多元异构数据和计算的统一[52]。如图 12 所示，一条流水线由多个节点组成，每个节点被称为处理器（Processor），处理器之间具有数据的传输。其中，具有一个输入和一个输出的处理器被称为转换器（Transformer），具有多个输入的处理器形成了合流（Merge）的效果，具有多个输出的处理器形成了分流（Fork）的操作。

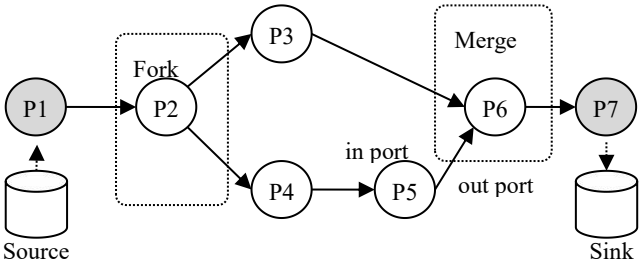

图 12 数据流水线的抽象模型

4.4 PARIS 网络设计

多个分布式的 PARIS 节点形成网络，该网络可有效连接数据孤岛、服务孤岛，发挥科学数据中心体系优势，形成完整视角的数据资源服务及协同分析能力。PARIS 网络具备如下特征：（1）开放性。任何新的节点按照约定的协议即可加入和离开网络；（2）去中心化。元数据、数据信息分布存在于各节点，网络中不存在一个中心节点，网络会自动选择一个节点作为 Leader 节点，该节点仅负责总体调度；（3）联邦式。即各节点遵循一套合作协议，提供相对完整的目录视图，实现联邦服务；（4）透明性。即对外屏蔽了数据的异构性、位置差异性、计算服务的差异性；（5）可靠性。

由于节点同时暴露数据目录和服务目录，因此 PARIS 网络具有两个平面：PARIS 数据平面和 PARIS 分析平面，如图 13 所示。

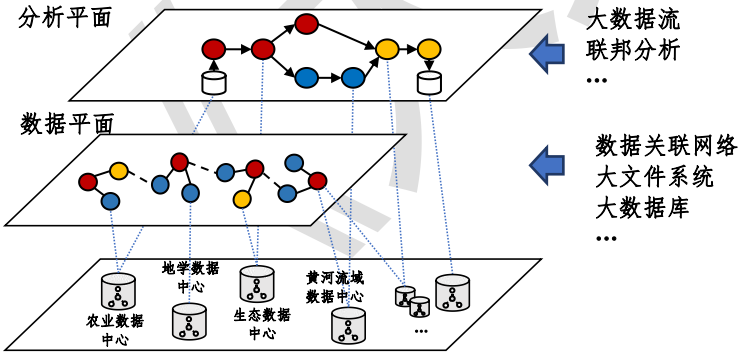

图 13 PARIS 网络架构

其中，PARIS 数据平面的核心功能是数据资源的融合，可以基于此平面构建数据关联网络、大文件系统、大数据库等服务；PARIS 分析平面的核心功能是分析能力的融合，可以基于此平面构建大数据流、联邦分析等服务。作为例子，图 14 示出基于跨域数据流水线调度的联邦分析流程。

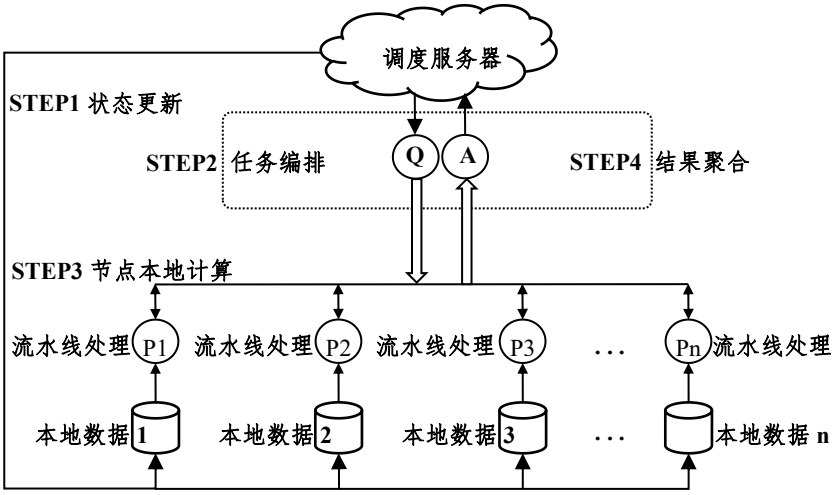

图 14 基于跨域数据流水线调度的联邦分析

5. 小结与展望

随着 FAIR、TRUST 原则的推出，科学数据的开放共享得到了较大的推动，作为科学数据的汇聚、存储、服务主体，科学数据中心日益发挥出重要的作用。然而同时，我们可以看到，目前科学数据的共享还是以科学数据的可发现、可访问为主，科学数据中心的数据服务仍然以“数据仓储服务”为核心，以“数据汇聚/汇交”为主要的数据共享模式，科学数据的利用与目前的仓储环境还存在较大的脱节，科学数据的利用能力和水平还存在着急需改进的地方。

随着“第四范式”、“融合科学”等数据驱动型的新型科研范式的发展，随着国家、部委、省市、机构等各级科学数据中心的成立和蓬勃发展，科学数据的利用在开放协作环境下呈现出跨边界、端到端、动态性和协作化特征，如何有效提高科学数据的协同服务能力，会成为下一步的研究焦点。

本文详细的分析了开放协作环境下科学数据利用的场景，提出促进科学数据利用的 PARIS 原则：可处理（Processable）、可问答（Askable）、可信赖（Reliable）、可联合（Incorporable）与可供给（Suppliable）。最后，本文给出可参考的技术实践路径。传统的 FAIR、TRUST 原则在开放协作环境下对科学数据利用工作的指导还存在着诸多限制，无法深入指导科学数据跨边界、端到端、动态性和协作化的利用，希望 PARIS 原则能在未来的科学数据共享工作中发挥更大的作用！

参考文献

[1] 国务院办公厅.关于印发科学数据管理办法的通知[EB/OL].(2018-03-17)[2022-07-25].http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm

[2]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[J]. 地球科学进展, 2019, 34(3): 306-315.

[3]徐波, 王瑞丹, 陈祖刚, 等. 科学数据中心综合运行评价体系赋权研究[J]. 中国科技资源导刊, 2021 (4): 96-103.

[4]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[J]. 中国科技资源导刊, 2020, 52(1): 1-5.

[5]科技部 财政部关于发布国家科技资源共享服务平台优化调整名单的通知[EB/OL].(2019-06-05)[2022-07-01]. http://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/qtwj/qtwj2019/201906/t20190610_147031.html.

[6]高雅丽. 在科技管理中，让科学数据“开放共享”[N]. 中国科学报,2022-06-08(001).DOI:10.28514/n.cnki.nkxsb.2022.001260.

[7]贵州省科技厅与贵阳市政府、贵安新区管委会联合召开贵州科学数据中心建设与运营专题会议 .[EB/OL].(2021-08-24)[2022-07-25].https://www.most.gov.cn/dfkj/gz/tpxw/20210 8/t20210824_176565.html

[8]甘肃省启动建设 3 个省级科学数据中心. [EB/OL]. (2020-12-20)[2022-07-25]. http://kjt.gansu.gov.cn/kjt/c111526/202012/1545072.shtml

[9]广东省科学数据中心组织召开广东省科学数据中心建设启动会. [EB/OL]. (2022-1-25)[2022-07-25]. https://www.gdcc.com.cn/news/trends/2081.html

[10]中国科学技术大学科学数据中心. [EB/OL]. [2022-07-25]. https://dc.ustc.edu.cn/science/

[11]北京大学管理科学数据中心. [EB/OL]. [2022-07-25]. https://dcms.pku.edu.cn/

[12]中国人民大学中国调查与数据中心. [EB/OL]. [2022-07-25]. http://nsrc.ruc.edu.cn/index.htm

[13]郑大一附院国家工程实验室数据中心召开救灾抢险感谢会. [EB/OL]. (2021-09-18)[2022-07-25]. https://www.zysbs.cn/html/caijing/2021_09/18/22156861.html

[14]北京大学开放研究数据平台-何涛水合物岩石物理研究课题组(北京天然气水合物国际研究中心).[EB/OL].(2021-06-15)[2022-07-25].https://opendata.pku.edu.cn/dataverse/gh_rp_res.

[15] 山东科技大学.关于启用“教师个人数据中心”的通知.(2021-10-20)[2022-07-25].https://tech.sdust.edu.cn/info/1081/1888.htm.

[16] 北京化工大学.个人数据中心[EB/OL].(2020-04-02)[2022-07-25].https://cit.buct.edu.cn/grsjzx/list.htm.

[17] 中新社：中国首个国家计量科学数据分中心在浙江上线[EB/OL].(2021-05-31)[2022-07-25].http://zjamr.zj.gov.cn/art/2021/5/31/art_1229003093_58999964.html.

[18] 国家青藏高原科学数据中心西藏分中心正式揭牌建设.[EB/OL].(2021-12-23)[2022-07-25].https://www.most.gov.cn/dfkj/xz/zxdt/202112/t20211223_178701.html.

[19] Axton M, Baak A, Blomberg N, et al. The FAIR Guiding Principles for scientific data management and stewardship[J]. Scientific data, 2016, 3: 15.

[20]Lin D, Crabtree J, Dillo I, et al. The TRUST Principles for digital repositories[J]. Scientific Data, 2020, 7(1): 1-5.

[21]李春秋, 杜博雅, 耿骞, 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.

[22]Jones S. Open data, FAIR data and RDM: the ugly duckling[C]//Open Sciences Conference, Berlin. 2018: 13-14.

[23] 韩扬眉. 科学数据要像学术论文一样积极“共享”[N]. 中国科学报,2022-03-15(001).DOI:10.28514/n.cnki.nkxsb.2022.000676.

[24]梅宏.数据要素化仍是国际性难题[N]. 中国科学报,2022-09-01

[25] The BESIII Collaboration. Probing CP symmetry and weak phases with entangled double-strange baryons[J]. Nature 606, 64 – 69 (2022). https://doi.org/10.1038/s41586-022-04624-1

[26] Guan C, Chen N, Qiao L, et al. Photosynthesis regulates the diel hysteresis pattern between soil respiration and soil temperature in a steppe grassland[J]. Geoderma, 2022, 408: 115561.

[27]The Research Data Lifecycle[EB/OL].(2018-12-17)[2022-07-13].https://algonquincollege.libguides.com/rdm/research-data-lifecycle.

[28]Tansley S, Tolle K M. The fourth paradigm: data-intensive scientific discovery[M]. Redmond, WA: Microsoft research, 2009.

[29]肖小溪, 甘泉, 蒋芳, 等. “融合科学” 新范式及其对开放数据的要求[J]. 中国科学院院刊, 2020, 35(1): 3-10.

[30]Linked science. [EB/OL].[2022-07-21].http://linkedscience.org/

[31]Niu C H, Aggarwal K, Li D, et al. A repeating fast radio burst associated with a persistent radio source[J]. Nature, 2022: 1-5.

[32]国家天文科学数据中心.天文学、高能物理、空间科学三个国家科学数据中心签订战略合作协议[EB/OL].(2021-07-21)[2022-07-21].https://nadc.china-vo.org/article/20210721091122.

[33] Abbott B P, Abbott R, Abbott T D, et al.Multimessenger observations of a binary neutron star merger[J].The Astrophysical Journal Letters, 2017, 848:L12.

[34] “黄河流域生态保护和高质量发展”跨平台专题上线.[EB/OL].(2021-08-17)[2022-07-21].https://www.agridata.cn/detail.html?type=work&id=64#

[35]崔辰州, 薛艳杰, 李建, 等. 虚拟天文台——天文学研究的科研信息化环境[J]. 中国科学院院刊, 2013, 28(4): 511-518.

[36]卢逸航, 李国庆, 陈祖刚. 科学数据中心间互操作模式研究[J]. 数据与计算发展前沿, 2022, 4(1): 69-83.

[37]Klump J, Bertelmann R, Brase J, et al. Data publication in the open access initiative[J]. Data Science Journal, 2006, 5: 79-83.

[38]Ali M H, Hosain M S, Hossain M A. Big Data Analysis using BigQuery on Cloud Computing Platform[J]. Australian JofEng Inno Tech, 2021, 3(1): 1-9.

[39]Wan M, Wu C, Wang J, et al. Column store for GWAC: a high-cadence, high-density, large-scale astronomical light curve pipeline and distributed shared-nothing database[J]. Publications of the Astronomical Society of the Pacific, 2016, 128(969): 114501.

[40]Bloland P, MacNeil A. Defining & assessing the quality, usability, and utilization of immunization data[J]. BMC Public Health, 2019, 19(1): 1-8.

[41]Prins H, Kruisinga F H, Buller H A, et al. Availability and usability of data for medical practice assessment[J]. International Journal for Quality in Health Care, 2002, 14(2): 127-137.

[42]Wachowicz M, Riedermann C, Vullings W, et al. Workshop report on spatial data usability[C]//Proceedings of the 5th AGILE Conference on Geographical Information Science. 2002: 429-436.

[43]李建中,刘显敏.大数据的一个重要方面:数据可用性[J].计算机研究与发展,2013,50(06):1147-1162.

[44] McMahan B,Ramage D.Federated learning:collaborative machine learning without centralized training data[EB/OL].(2017-04)[2019-12-08].https://ai.googleblog.com/2017/04/federated-learning-collaborative.html

[45]Barker A, Hemert J. Scientific workflow: a survey and research directions[C]//International Conference on Parallel Processing and Applied Mathematics. Springer, Berlin, Heidelberg, 2007: 746-753.

[46]Foster I, Kesselman C, Tuecke S. The anatomy of the grid: Enabling scalable virtual organizations[J]. The International Journal of High Performance Computing Applications, 2001, 15(3): 200-222.

[47] Foster I, Kesselman C. The Globus project: A status report[C]//Proceedings Seventh Heterogeneous Computing Workshop (HCW'98). IEEE, 1998: 4-18.

[48] Shiers J. The worldwide LHC computing grid (worldwide LCG)[J]. Computer physics communications, 2007, 177(1-2): 219-223.

[49]Bernholdt D, Bharathi S, Brown D, et al. The earth system grid: Supporting the next generation of climate modeling research[J]. Proceedings of the IEEE, 2005, 93(3): 485-495.

[50] Alvord M M, Lu F, Du B, et al. Big Data Fabric Architecture: How Big Data and Data Management Frameworks Converge to Bring a New Generation of Competitive Advantage for Enterprises[J]. 2020.

[51]Gartner.Gartner Identifies Top 10 Data and Analytics Technology Trends for 2019[EB/OL]. (2019-02-18)[2022-07-25].

https://www.gartner.com/en/newsroom/press-releases/2019-02-18-gartner-identifies-top-10-data-and-analytics-technolo

[52]Koitzsch K. Data pipelines and how to construct them[M]//Pro Hadoop Data Analytics. Apress, Berkeley, CA, 2017: 77-90.