高校开放科学数据平台建设探索*

——以浙江大学开放数据平台为例

□姚翔宇　黄晨　葛杭

摘要　开放科学数据平台是开放数据活动的信息基础设施，高校建设相关平台是加强科学数据服务的重要步骤。通过调查国内外实践现状，提出国内平台在建设逻辑、数据协作功能、数据融合标准等方面有待探索的问题。结合浙江大学开放数据平台实践阐释平台自主开发的总体框架、围绕协作过程和面向数据融合的建设重点，以及建设成效和思考。以期为国内高校开放数据平台建设提供参考，推进国内开放科学数据活动进一步走向落地实施。

关键词　数据开放　科学数据　研究数据　数据平台

分类号　G250.7

DOI　10.16603/j.issn1002－1027.2023.06.013

1　引言

开放科学数据是开放科学运动向纵深发展的重要领域，其重点在于对既有数据进行发布和共享，以发掘数据的新价值或验证既往研究。由数据公平港湾倡议组织（Data FAIRport Initiative）提出的FAIR原则已经成为开放科学数据活动的宏观目标和重要共识，该原则包括可发现、可访问、可互操作、可重用4个方面[1－2]。以专注于科学数据管理和传播标准研究的数据文档倡议组织（Data Documentation Initiative，DDI）提出的数据生命周期模型为参照[3]，开放数据活动主要涉及数据生命周期中的存储、发布、发现、重用环节。

围绕科学数据的开放共享，各国先后提出了相关倡议和政策要求。2012年欧洲研究型大学联盟（League of European Research Universities）发表《关于开放研究数据的声明》支持科学数据开放[4]。2013年美国白宫科技政策办公室（Office of Science and Technology Policy）发布《联邦资助科研成果开放获取政策》，要求促进对联邦资助研究中的科学数据提供开放获取[5]。2016年英国国家科研与创新署（UK Research and Innovation）发布《开放研究数据协议》，制定了10项利用开放研究数据的原则[6]。2018年中国国务院印发《科学数据管理办法》[7]，2019年国际科学理事会数据委员会（Committee on Data of the International Science Council）发表《科研数据北京宣言》[8]，显示出开放科学数据相关工作在国内关注度和重要性的提升。同时，近年来随着数据驱动的人工智能技术快速发展，各细分科学领域的研究范式也在快速演化，其中基于高质量的科学数据建立大数据模型、以人工智能辅助研究发现和科技创新已成为各国寻求科技突破的重点竞争领域，这一趋势对科学数据的快速积累、有效组织和开放获取提出了更迫切的要求。

开放科学数据活动涉及诸多要素，包括数据政策、数据平台、数据服务、数据素养等。其中数据平台是数据活动的主要载体，是开放科学数据活动实践不可或缺的信息基础设施，承担着将数据本体与数据生产者、使用者、管理者等利益相关方联结起来的作用。现有相关平台可归纳为存储型、索引型、导航型3类，它们自下而上共同构成了开放科学数据平台体系。存储型平台直接保存科学数据，能够涵盖数据的存储、发布、发现、重用等全过程，例如哈佛大学的Harvard Dataverse，密西根大学主导的ICPSR等。索引型平台收割、整合与揭示多个平台

* 浙江省高等学校图书情报工作指导委员会科研项目“高校开放科研数据平台的建设实践研究”（编号：2022TKT005）的研究成果之一。
通讯作者：姚翔宇，ORCID：0000-0002-8476-1450，邮箱：yaoxiangyu@zju.edu.cn。

中科学数据的元数据，一般不直接存储科学数据，例如作为澳大利亚国家科研基础设施的 Research Data Australia 等。导航型平台通过自主提交或主动采集的方式登记其他平台的信息，并进行标引、揭示甚至评价，以协助用户发现潜在的数据来源或存储设施，例如德国研究基金资助的数据存储库注册平台 Re3data 等。

高校是从事科研活动的重要主体，具备科学数据产出的能力，面临数据保存、管理和发布的综合需求，高校开放科学数据平台（以下简称高校平台）的需求和能力特征主要对应存储型平台。从现有的情况看，存储型平台也主要由直接从事数据生产和发布活动的机构建设，包括科研院所、高等院校、科技成果出版机构等。但是从实践的总体情况看，一方面相对于国外，我国的平台建设与应用仍相对滞后，处于“平台数量少，尚未形成规模”的阶段[9]，Re3data 上登记的存储型平台有 3136 个，其中美国 1170 个、德国 501 个、英国 319 个，而中国大陆仅有 81 个[10]。另一方面在国内现有平台中，占主体的是科研院所建设的平台，例如国家高能物理科学数据中心等 20 个国家科学数据中心[11]；而高校方面仅有北京大学、复旦大学等数个高校进行了平台建设。结合我国高校不断提升的科研产出占比，这两个方面显示出在国际相关平台建设已经达到相当规模的情况下，国内高校平台建设仍处在十分初级的阶段，需要加大力度开展开放科学数据平台的实践探索、研究总结和应用推广，以汇聚散落在高校科研团队中的科学数据，更好支撑我国数据驱动的科学研究转型发展。

2 高校平台实践现状及问题

在高校主导的开放科学数据平台建设方面，国外特别是欧美地区起步较早、数量较多，平台建设路径、服务功能等方面都形成了较为丰富的实践。国内高校相关实践主要开始于 2011 年武汉大学高校科学数据共享平台的建设，此后复旦大学、北京大学、华东师范大学等先后建设了各自的开放科学数据平台。通过调查国外高校平台及国内前述 4 个高校平台，本文对国内外实践现状进行了比较和总结。

（1）建设路径和技术方案。国外高校体现出较强技术实力，许多平台采用自主研发形式建设，例如 ICPSR、UKDA、Harvard Dataverse 等均结合自身需求进行平台研发和迭代，其中哈佛大学从 2007 年开始搭建的 Dataverse 系统，作为开源社区项目已经开发至 5.13 版本，也被许多其他高校采用。欧洲部分高校采用了商业化的平台解决方案，如牛津大学、伦敦大学学院、荷兰四校联盟的 4TU.ResearchData 等均基于商业化云端数据系统 Figshare 开展服务。另外开源的通用机构文档存储系统 DSpace、Fedora 也有应用。国内平台主要是在国外开源系统基础上进行部署和二次开发，如武汉大学平台基于 DSpace 部署，复旦大学和北京大学平台基于 Dataverse 定制开发。华东师范大学则选择了自主建设路径搭建平台。

（2）主要功能。国内外高校平台在基本功能上均涵盖了开放数据活动周期的主要环节，主要以数据存储为基础，提供数据发布、发现、访问等功能。部分平台如 Dataverse 还集成了数据的在线预览和分析功能，但是相关功能一般比较简单。除 ICPSR 通过接入在线数据分析平台 SDA 进行高级数据分析外，较为复杂和专业的数据分析一般未在开放科学数据平台中提供。

（3）用户参与和数据组织。国外平台存在两种不同的取向，ICPSR、UKDA 等侧重于数据汇交和发布服务，它储存用户在线或离线提交的数据，在按照一定标准审查处理后以数据集为单元发布。用户对平台功能的参与度低，侧重在平台之外完成数据组织。新兴应用 Dataverse、Figshare 等系统的平台则侧重用户参与数据组织，例如 Dataverse 提出数据空间概念，用层层嵌套的空间将数据集和用户权限按照层次关系组织起来，提供了团队用户共同管理数据集的条件。Figshare 更强调用户空间，用户的账户与数据集建立关联，可以如社交平台一样运营自己的数据集。相应地 Figshare 更贴合用户的研究过程，数据集版本通常有较高的更新频次。国内平台主要偏向于对数据进行多层次组织，并允许用户自主管理权限范围内的数据。除了应用 Dataverse 系统的平台外，华东师范大学平台具有数据集、课题、文件 3 个层次的数据包含关系并对应分配用户权限，在通常作为基本单元的数据集层次之下进行了课题子集划分；武汉大学平台在继承 DSpace 组织模型的基础上开放了课题、数据集两个组织层次，用于归集和分级管理数据。

（4）平台合作和数据融合。国外平台体现出较

高的合作与融合程度，体现为合作联盟、系统通用性和数据标准等多种形式。ICPSR、UKDA、4TU.ResearchData 等通过多机构联盟形式建设，形成统一的数据存储和服务，是最直接的合作形式。Figshare 作为商业化云系统，在为不同高校提供定制化服务的同时，通过主平台入口提供了各机构数据的统一揭示。数据标准包括技术层面及关涉语义的数据描述层面，技术层面体现为广泛应用 OAI－PMH、RESTful API 等建立数据接口，通过 DOI 对数据进行统一标识和索引。数据描述包括使用兼容性的元数据元素等。兼顾自然科学和社会科学的平台通常以都柏林核心元数据集 DC 元素为基础；起步于社会科学数据管理的平台多以 DDI 元素为基础，其中 Harvard Dataverse 也通过元数据导出形式对 DC 元素进行了完整映射；不过也有 ICPSR、UKDA 等平台主要根据自身需要设计元数据元素。国内平台目前在这方面还比较薄弱，主要的融合特征体现在技术层面，例如复旦大学、北京大学、华东师范大学的平台均采用 OAI－PMH 作为数据接口，北京大学成熟应用 DOI 标识符并向科学数据注册平台 DataCite 提交数据。在数据描述上，除武汉大学采用自定义的元数据元素外，其他平台主要选用部分 DDI 元素构建元数据、但相互之间选用的基本元素存在差异，其中北京大学也支持元数据导出为 DC 格式、但并未实现元素的完整映射。同时，国内平台主要由各高校独立建设和运营、缺少数据合作，复旦大学和北京大学尽管都应用了 Dataverse 系统，但是尚未在数据互通方面进行尝试。

结合国内外高校平台实践特点，特别是国内平台建设现状，本文将重点关注和探索以下问题：

(1)探索自主研发路径下平台建构的思路和逻辑。国外自主研发及市场化开发的建设路径，形成了围绕需求定位的正向逻辑构建，具备平台不断迭代进化的主动性，而国内以选型应用、二次开发为主，有一定被动性。本文将通过探索自主研发的建设思路，构建平台整体功能框架，丰富国内平台建设路径和思考。

(2)探索能够更好支持研究过程中数据组织和用户协作的平台功能。高校平台主要吸纳机构内的科学数据，提升数据数量和质量需要政策、咨询服务等多方面的支持，而在平台层面则需要结合用户特点做好功能建设、增强用户黏性。高校平台建设已经体现出以数据存储为基础，覆盖数据发布、发现、重用的共性，而同时高校平台与校内研究团队具有服务共生关系，在数据管理方面需要考虑团队内部数据使用与存储的动态性、团队内跨项目的成员组成和项目数据之间的关联性，向数据的团队协作、过程管理和自助管理方面倾斜。这方面国内平台已经关注到团队数据管理并尝试了不同的数据组织形式，但是对协作的理解主要还停留在数据归集和层次管理上，功能易用性也有待商榷，本文将通过对相关功能的进一步探索，强化平台在用户数据过程服务方面的作用。

(3)探索面向数据融合的数据描述方法。单一高校平台面临数据来源有限的局限性，最终需要通过平台间合作建立数据互通和统一揭示体系、扩展数据范围。为了实现这一点，除了政策、机制、资金等方面的保障，在单一平台建设中还重点需要关注面向数据融合的数据标准。其中技术方面 OAI－PMH 接口、DOI 标识符等的应用已经比较成熟，而数据描述方面国内各平台主要基于自身需要进行元数据自定义或者 DDI 衍生应用，还未形成面向数据融合的建设导向和统一规范。在国内高校尚未形成紧密合作机制的背景下，本文将从具有兼容性潜力的数据描述方法切入，进行数据融合的前期探索。

3　建设思路和方法

3.1　总体框架

浙江大学开放数据平台(以下简称浙大平台)建设由依托浙江大学图书馆成立的浙江大学信息资源分析与应用研究中心实施，通过自主设计平台架构和具体需求、委托软件开发企业进行系统开发的路径落地实施。在学校推进“双一流”建设的背景下，浙大平台以推动浙江大学科学数据的开放、交流与共享为总目标，以服务 3 类对象为具体目标：面向数据生产者提供科学数据的存储、管理和发布服务，创建面向研究过程、易于协作的数据共享环境；面向数据使用者提供科学数据的发现、浏览和访问服务，创建信息揭示清晰、访问通道顺畅、倡导学术规范的数据使用环境；面向管理者提供与数据开放过程相适应的合规审核和系统管理功能，支持平台的持续有序运行。

在现状调研和问题分析的基础上，平台设计明确了“以数据存储为基础、服务开放科学数据周期，重视研究过程动态性和团队协作性，重视学科通用

性和跨平台互操作性”的特征定位，形成如图 1 所示架构。其中功能应用层（以下简称功能层）是指直接面向用户和数据活动的功能实体，数据生命周期和过程服务主要在这一层次体现；协议规范层（以下简称协议层）是指功能背后的标准规范，数据服务在跨学科、跨平台方面的通用性和互操作性主要在这一层次体现。基础要素是指平台应该具备的基本特性，对应于开放科学数据活动中存储、发布、发现、重用各环节的基础性需求及定义平台架构逻辑的核心层次，完成基础层建设即可实现平台的可用；拓展要素是指能够进一步提升用户体验、平台质量的需求，可以在平台建设迭代中逐步加强和优化。

同时，浙大平台采用分期建设策略，以逐步扩大服务范围为原则，划分馆内、校内、校际 3 个建设阶段。第一阶段主要是结合馆内科学数据活动相关的实际场景，进行系统设计、基础层开发和平台试用，积累具体经验；同时以小范围社会科学学科用户为试运行目标对象，对涉及多学科的功能在设计方面保持框架完整、在落地实施方面参照试运行学科的具体需要进行原型验证性开发。第二阶段进行平台优化和拓展层开发，并将平台在校内全面推广使用。第三阶段是逐步开展校外服务，同时通过校际合作探索不同高校平台间的互通对接和数据统一揭示。

3.2 围绕协作过程的功能层建设重点

（1）数据组织结构

将数据存储和发布过程与数据生产者在研究过程中的协作活动结合起来，并提供与之相适应的数据组织结构，是浙大平台选择自主开发模式的重要动因。具体地，浙大平台在数据基本单元的定义上延续了现有通常做法，即将数据集作为检索、访问、管理的基本单元，据此组织数据文件和元数据。但同时重点设计了数据集的父层级，以承担多人协作和共同管理的高级功能（见图 2）：父层级包含实体层（空间和项目）和虚拟层（文件夹）。

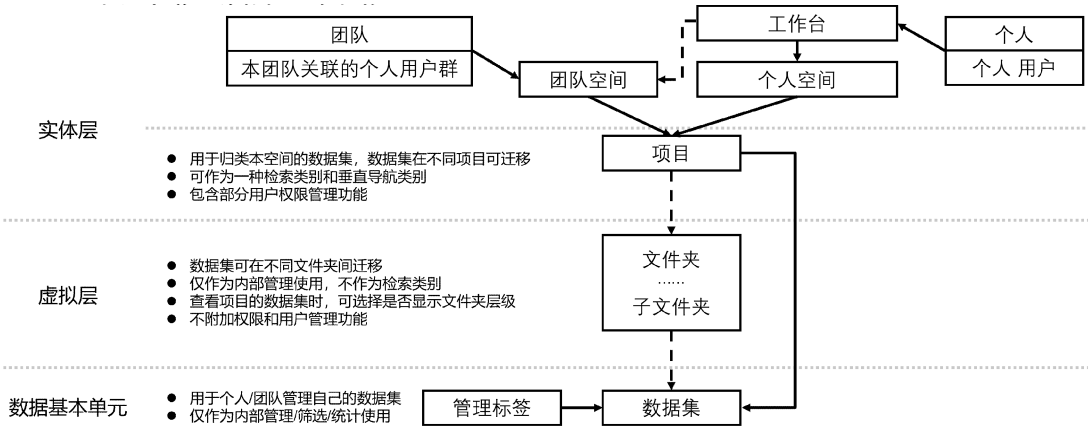

图 2 浙江大学开放数据平台数据组织结构

实体层的“空间”反映了一组数据集的成果归属属性，它不同于 Dataverse 中层层嵌套的空间概念，而是仅划分两种类别，以对应现实中的数据生产者实体，即研究者个人或研究团队。对于研究者个人，他能够直接创建和管理自己所拥有的研究数据，而个人空间是展示和汇聚这些数据集的权属集合。对于研究团队，能够基于个人用户组成关系稳定的群组，通过内部角色分工创建和维护团队共有的数据集；并且在数据集发布前，团队空间可作为研究过程数据的管理集合供团队内部查看和编辑，实现高校数据平台将存储功能由长期保存到过程性存储的延伸。实体层的“项目”反映了一组数据集的内容关联

属性，便于规模相对较大的研究团队既能将有关联的数据集分别组织起来，也能将不同分工的成员进一步组织起来。

虚拟层的“文件夹”，是为研究者管理大量数据集而设置的灵活分类组织层次。文件夹介于项目和数据集之间，最多允许创建两层子文件夹，它本身不附加权限和用户管理属性，仅作为可选形式对成员展示，以便利团队内部管理和使用中对不同数据集的标记、区分和整理。在面向数据使用者时，文件夹的结构则不会呈现出来，以保持数据集作为基本单元的层次清晰性、降低用户认知成本。

(2)权限控制

开放数据不等同于无条件开放获取。已发布的数据应是始终公开、可访问的，但它可能要求支付费用、接受法律协议，或者得到数据生产者的允许[12]。相应地，需要对数据访问进行权限设置。同时，当引入团队协作功能之后，基于团队分工和管理的需要，也需要对数据权限进行控制。

为了实现多场景权限控制的灵活性和细粒度，浙大平台从两个维度定义了权限体系。其一是依附于空间的权限控制(见表1)。这一维度主要面向团队协作活动，空间内部赋予用户高度自主的管理权，可以通过空间中的角色分组实现日常数据活动的批量授权管理。其二是依附于数据集的权限控制(见表2)。由于数据集被定义为数据管理的基本单元，所以基于数据集设定的权限具有最高的优先级和最小的粒度。其中，已发布的数据集意味着可被发现和查看，对它的权限控制主要在于数据文件的可访问性，即根据用户为数据集设置的公开程度，控制相应数据文件可供下载的用户分组。未发布的数据集则不对外公开，需要通过邀请用户加入协作组的形式赋予有限的合作编辑功能。协作组的形式也为跨团队数据合作提供了灵活性。

在平台与用户的权限关系方面，则降低系统干预度、增加用户自主权。例如平台管理员仅对团队创建和数据集发布环节进行合规审核，不直接干涉用户的数据权限；平台为团队空间设定默认角色权限模板，但是允许用户自主调整角色权限定义。

(3)计划任务和数据统计

团队协作中对数据的管理最终要落实到对数据活动的管理，这种管理不仅是对成员分工的管理，也包括对数据工作计划任务的分配和对数据工作状态及成效的统计。其中计划任务功能主要是以待办提醒的形式实现，允许团队管理者添加任务和发送提醒，并且可以设置周期任务定时发送提醒消息，以方便团队开展长期的、周期性的数据采集工作。数据统计功能主要是以用户行为为基础，以人员、时间、学科、标签等为维度，客观描绘数据工作、数据使用相关的指标。除了常见的访问量、下载量、数据集容量等指标外，还着重纳入了与数据活动过程相关的数据工作量、成员参与度等指标，以方便用户追踪数据工作轨迹、评估数据影响力。

表1　浙江大学开放数据平台依附于空间的权限设计

功能		角色默认权限				
		团队空间				个人空间
		创建者	管理员	编辑员	只读成员	所有者
团队管理	编辑团队信息	√				
	增删自定义角色组、自定义角色权限	√				
	管理成员(增删、分组成员)	√	√			
数据集管理	新建数据集	√	√			√
	编辑元数据	√	√	√		√
	发布数据集	√	√			√
	编辑数据集权限	√	√			√
	查看未发布数据集	√	√	√		√
文件管理	上传	√	√	√		√
	替换	√	√	√		√
	下载	√	√	√	√	√
	删除	√	√	√		√

表 2　浙江大学开放数据平台依附于数据集的权限设计

	功能	数据集公开程度	平台全局用户		数据集用户分组	
			游客	注册用户	下载组	协作组
已发布数据集	查看数据集元数据、文件列表、历史版本等信息		√	√	√	√
	下载数据集文件	完全公开	√	√	√	√
		部分公开		√	√	√
		申请访问(可直接在平台线上发起申请加入下载组)			√	√
		限制(平台不直接提供申请方式，需通过邮件等其他方式申请访问)			√	√
未发布数据集	编辑元数据					√
	发布数据集					
	编辑数据集权限					
	查看未发布数据集					√
	上传数据集文件					√
	替换数据集文件					√
	下载数据集文件					√
	删除数据集文件					√

3.3　面向数据融合的协议层建设重点

(1)元数据

元数据是用于描述数据的数据，浙大平台以数据集作为数据管理和发现的基本单元，相应的元数据建设主要围绕数据集层次展开，并将元数据的兼容性作为主要出发点，设计了元数据层次和元素。

元数据层次与跨学科维度有关，高校平台具有学科综合性，需要兼顾不同学科在数据特征方面的差异。从数据融合的视角看，关系到数据可发现和可访问的部分处于相对底层，容易在跨学科大范围内实现统一，关系到数据可互操作和可重用的部分处于相对顶层，只能在特定学科小范围内统一[13]。因此，浙大平台规划了“基本元数据＋专业元数据”的二级元数据层次。其中，基本元数据学科特征不显著，是描述数据集的基础，而专业元数据用于增强不同学科数据集专业性、差异化的元数据表示能力。

元数据元素与跨平台维度有关，单一高校平台的数据主要来自本机构的数据生产者，因此高校平台有融入广泛的数据集成体系的内在要求，需要考虑基本元数据跨平台交换的语义对应关系。浙大平台将基本元数据的元素划分为 3 个部分，其中基本信息部分主要包含数据集的通用描述信息，服务于数据集的发现和访问；方法部分主要包含数据集的采集、处理信息，服务于数据集的重用；权利部分主要约定数据所有权和使用限制方面的条款。具体元素综合考虑了 DC 元素、DDI 元素和平台调研中记录的常用元素，并明确了相互间的映射关系。其中对 DC 元素能够完全覆盖，而 DDI 元素非常丰富，这里主要参考了 ICPSR、UKDA 选取的常用元素和 DDI 精简集提示的重要元素[14]。对于各个元素分别给出英文名称、中文名称、元素说明并定义了必填项和录入格式(见表 3)。

(2)引用格式

引用格式是数据描述的另一维度。恰当引用科学数据是加强学术规范、扩大传播影响力、提高重用价值的重要环节，也是促进数据在不同学术成果形式间关联融合的媒介。开放数据平台自动生成和展示信息明确的数据引用格式，有助于引导用户进行合理引用和跨平台传播。尽管目前开放数据活动中还没有就引用方式形成统一的规范，但是国内外关于数据引用格式已经形成了一些标准和实践方面的探索，浙大平台对国内《GB/T 7714－2015 信息与文

表 3　浙江大学开放数据平台基础元数据设计及与 DC、DDI 元数据的映射关系

	英文元数据项	中文元数据项	说明	是否必填	DC 元数据对应项	DDI 元数据对应项
基本信息	Title	标题	数据集名称	是	Title	＜titl＞ 2.1.1.1
	Description	描述	描述数据集重点信息、摘要	是	Description	＜abstract＞ 2.2.2
	Keyword	关键词	概括性描述数据集内容的词语	是	Subject	＜keyword＞ 2.2.1.1
	Topics	学科	数据集内容所属教育部一级学科	是		＜topcClas＞ 2.2.1.2
	Author	负责人	数据集的著作者或负责人，及其所属机构	是	Creator	＜AuthEnty＞ 2.1.2.1
	Collector	采集人	数据集的采集人，及其所属机构	是	Contributor	＜othId＞ 2.1.2.2
						＜dataCollector＞2.3.1.2
	Contact	联系邮箱	可接收联系信息的电子邮箱	是		＜contact＞2.1.4.2
	Sponsor	资助信息	资助机构、资助基金等信息	否		＜fundAg＞2.1.3.6
						＜grantNo＞2.1.3.7
	Publisher	出版者	数据集的出版发布机构：浙江大学开放数据平台	系统自动获取	Publisher	＜producer＞ 2.1.3.1
	Version	版本号	数据集的版本号	系统自动获取		＜version＞2.1.6.1
	Version Date	当前版本发布日期	数据集当前版本发布的日期	系统自动获取		
	Production Date	首次发布日期	数据集首次发布的日期	系统自动获取	Date	＜prodDate＞ 2.1.3.3
	Type	类型	文档的类型：数据集	系统自动获取	Type	＜dataKind＞ 2.2.3.10
	Format	文件格式	数据集中所包含文件的格式及数据文件大小	系统自动获取	Format	＜fileType＞ 3.1.5
	Identifier	唯一标识符	数据集唯一标识 ID，暂未定义	系统自动获取	Identifier	＜IDNo＞ 2.1.1.5
	URL	统一资源定位符	数据集页面地址	系统自动获取		＜ holdings location = "" callno="" URI=""＞ 2.1.8
方法	Language	语种	数据信息所用语言的语种	是	Language	
	Temporal Coverage	时间范围	数据集覆盖的时间范围	是	Coverage	＜timePrd＞ 2.2.3.1
	Date of Collection	采集时间	采集数据的时间范围	是		＜collDate＞ 2.2.3.2
	Geographical Coverage	地理范围	数据集覆盖的空间范围	否		＜geogCover＞ 2.2.3.4
	Frequency	频率	采集、更新数据的频率	否		＜frequenc＞2.3.1.3
	Collection Resouces	采集来源	数据来源	是	Source	＜sources＞ 2.3.1.8
	Method of Data Collection	采集方式	数据的采集方式	是		＜collMode＞2.3.1.6
	Observation Units	分析单元	数据的观察分析单元	否		＜anlyUnit＞2.2.3.8
	Time Dimensions	时间组织方式	数据采集的时间组织方式，如纵向时间、截面时间等	否		＜timeMeth＞2.3.1.1
	Sampling Procedures	抽样方法	数据抽样方法	否		＜sampProc＞2.3.1.4
	Response Rates	应答率	数据应答率	否		＜respRate＞2.3.3.1
	Weight	权重	数据的权重设定说明	否		＜weight＞2.3.1.12
	Reference	参考文献	数据集所引用的内容，建议采用标准引用格式列举	否		＜othRefs＞2.5.4
	Related Content	相关内容	与数据集有关联的研究内容、文档，如相关文献、出版物、网站等	否	Relation	＜othrStdyMat＞ 2.5
	Additional Info	其他信息	需要补充的其他信息	否		＜notes＞2.6
权利	Access	访问权限	数据集可以公开获取的程度	是		＜restrctn＞2.4.2.3
	Copyright	版权及所有权声明	数据集权利声明，可填写通用的版权协议或在法律框架内自主声明	是	Rights	＜copyright＞ 2.1.3.2

献 参考文献著录规则》《GB/T 35294－2017 信息技术 科学数据引用》，国外 Dataverse 引用格式[15]、ICPSR 引用格式[16]进行了比较，最终基于《GB/T 7714－2015 信息与文献 参考文献著录规则》中规定的“电子资源”引用格式，选取主要元素、并增加版本元素形成了引用格式（见图 3），在数据集页面进行生成和展示。

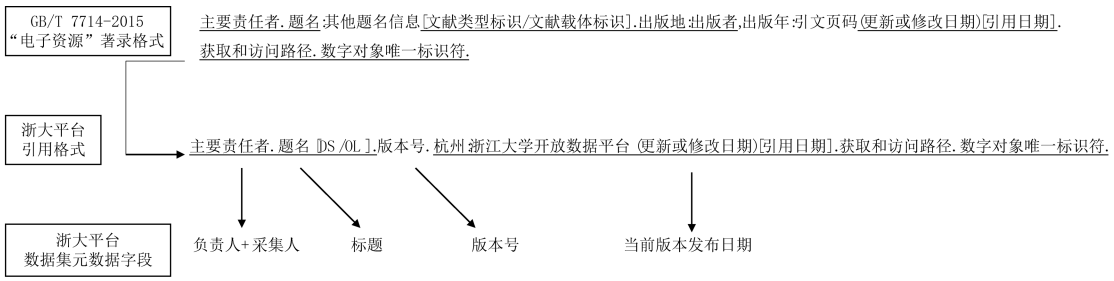

图 3　浙江大学开放数据平台的引用格式

在格式选型设计中主要考虑了两方面因素。一是相关格式涵盖的信息量。刘小宇等通过调查我国科学数据平台的被引状况，发现平台提供的引用声明不够具体是造成引用不规范的重要原因，只包含平台名称和平台网址，不包含数据创建者、数据名称和数据解析地址的引用格式，很难回溯获取相应科学数据[17]。通过对比引用格式中包含的元素信息（见表 4），可见 4 种引用格式都包含了访问科学数据的必要元素，其中《GB/T 7714－2015 信息与文献参考文献著录规则》提供的元素最丰富，其他 3 种格式都包含了额外的版本元素。二是相关格式的应用潜力。Dataverse 格式在部署该系统的平台中集成，有一定的平台传播效应，例如北京大学、复旦大学平台采用了该格式；ICPSR 格式主要是该平台自定义的标准；《GB/T 35294－2017 信息技术 科学数据引用》是专门针对科学数据制定的国内标准，但是尚未出现有影响力的应用；而《GB/T 7714－2015 信息与文献 参考文献著录规则》中“电子资源”引用格式是国内广泛应用的文献著录规则的一个子集，易于与现有的参考文献应用场景相融合，同时国内开展数据论文出版的学术刊物《中国科学数据》和《图书馆杂志》给出的数据引用格式也复用了该标准，在影响力和易用性方面相对具有较大潜力。

表 4　科学数据引用格式元素对照

来源	GB/T 7714－2015	GB/T 35294－2017	Dataverse	ICPSR
可用元素	主要责任者	作者	Author	Author
	题名:其他题名信息	名称	Dataset Title	Title
	文献类型标识/文献载体标识			
	出版地			
	出版者	传播机构	Data Repository	Distributor
	出版年	传播时间	Year	
	引文页码			
	更新或修改日期			Date
	引用日期			
	获取和访问路径	解析地址		
	数字对象唯一标识符	唯一标识符	Global Persistent Identifier	Persistent Identifier
		版本	Version	Version
		创建机构		
		创建时间		

注：为方便对比差异，表中元素排列顺序不代表引用格式中的实际展示顺序。

4 成效和思考

4.1 建设与应用成效

浙大平台在2022年初完成一期建设并开展了试运行(网址：ciraa.zju.edu.cn/opendata)，实现了基础层的功能、协议方面主要设计构想，具备以下应用特点：(1)覆盖开放数据活动周期，提供丰富的权限控制体系；(2)面向研究过程，适应个人工作与团队协作，可通过建立团队赋予成员多样的权限，将平台作为团队数据协作工具，编辑和维护研究团队共有的数据集；(3)具备通用性的元数据元素、自动生成引用格式；(4)提供多样的检索发现路径，支持关键词检索和垂直学科导航；(5)支持高宽容度数据文件上传和数据安全保护。

试运行期间，平台主要面向受邀用户开放注册，截至2023年10月已收录9个团队空间、295个数据集。通过试用，建设团队也同步收集了用户反馈的功能建议，并结合分期规划确定了后续建设要点。

4.2 建设路径的决策因素

国内外平台建设实践呈现出了多种技术应用路径，浙大平台选择自主研发路径，主要基于对商业方案、开源方案和国内开发生态的适用性分析。购买商业化平台的系统服务能够实现平台的快速落地、减轻后续运维负担，国外以Figshare为代表的商业方案经过长期迭代在功能上日臻完善，但是其在本地化、定制化上的不确定性以及云服务的数据安全风险仍然是主要障碍。以Dataverse为代表的开源产品专门面向开放科学数据活动，覆盖了主要的需求场景，但是其功能集成耦合程度高，存在二次开发难以触及核心功能的问题，不利于实现在架构层面强化数据协作功能的开发设想；同时存在版本升级与定制功能兼容适配困难的问题，为平台后续迭代升级工作带来不确定性。与此同时，近年来国内信息产业发展快速，头部企业技术实力增强也带动了相关企业的数量增加和质量提升。在前期调研中，有来自图书馆技术服务、高等教育信息化、综合性信息服务等不同领域的多家企业展现了合作意愿，为推进自主方案的落地提供了成本可控、技术可行的市场条件。

4.3 实践中的馆员参与机制

在平台建设中，定义产品经理和数据用户的馆员角色，有助于避免创新型功能开发可能产生的预期偏差。一方面，不同于通常采取的需求方提出意向和框架、由承建单位负责需求细化和实施的开发模式，浙大平台实践通过设置产品经理角色，预先开展详细的系统设计，对业务流程、功能节点、页面交互、术语规范等方案进行完整构建，并对项目全流程进行跟踪把控，为落实平台自主开发路径创造了过程可控的条件。另一方面，图书馆从事学术情报数据工作的团队是平台的直接用户群体，将相关馆员作为数据用户纳入团队，全程参与需求调整和数据工作流程测试，以直面用户的方式完善平台设计，为平台的用户体验和细节控制提供了保障。

4.4 高校平台的建设与合作

实现未来高校平台数据融合和体系构建，需要在数据描述方法、技术方法和合作机制方面结合现实条件因地制宜、多点发力。在数据描述层面，面对国内目前缺乏统一标准的局面，浙大平台研究应用了注重兼容性的元数据和基于“参考文献著录规则”的引用格式，以期提出有参考性的解决方案。在技术层面，结合国内平台将DOI用于数据集标引和永久访问、将OAI－PMH用于元数据收割等的应用经验[18－20]，浙大平台倾向于沿用成熟技术方案，降低对接成本。在合作机制层面，浙大平台通过中国高校研究数据管理推进工作组①的工作会议分享了平台建设的方案和思考。目前这一机制仍有较大的潜力等待挖掘，通过开展更紧密的合作研究，特别是在元数据元素选用、学科主题词表建设方面推进相关国内共识的建立，将有助于为更高层次的开放数据合作创造条件。

5 结语

对高校及高校图书馆来说，建设作为开放数据活动基础设施的系统平台是应对科学数据开放趋势和支持数据驱动研究的重要步骤。本文通过阐释浙大平台的建设重点和实施过程，论述其中的探索和思考，或许能为相关实践提供一些参考。

同时要注意到，一方面尽管开放科学数据平台的建设有许多功能和标准方面的共性需求，但是各建设主体仍要结合自身条件确定建设路径、提出特定思考、明确建设重点，这不仅是更好服务核心用户

① 2014年成立，由复旦大学图书馆、北京大学图书馆、浙江大学图书馆等9家单位组成。网址：http://society.library.sh.cn/node/2611.

的需要，也是现阶段丰富国内平台建设实践的需要。另一方面，从科学数据开放活动的长期发展来看，高校平台中的数据需要依赖于更高层级的统一发现和揭示服务来促进存量数据的利用和融合。为达成这一目标，在建设运营好各自平台的基础上，还需要各建设主体通力合作，建立科学数据开放生态，逐步搭建起存储型、索引型、导航型平台相互协同的开放数据平台体系。

参考文献

1 Data FAIRport Initiative. Data FAIRport[EB/OL].[2023－04－10].https://www.datafairport.org/.
2 Wilkinson M D, Dumontier M, Aalbersberg I J, et al. The FAIR guiding principles for scientific data management and stewardship [J]. Scientific Data, 2016, 3(1): 1－9.
3 Data Documentation Initiative. DDI 3.3 (2020) documentation [EB/OL]. [2023 － 04 － 10]. https://ddi-lifecycle-documentation.readthedocs.io/en/latest/User%20Guide/Introduction.html.
4 League of European Research Universities. LERU statement on open research data[EB/OL]. [2023－07－23]. https://www.leru.org/publications/leru-statement-on-open-research-data.
5 Office of Science and Technology Policy. Public access to results of FDA-Funded scientific research[EB/OL]. [2023－07－23]. https://www. fda. gov/science-research/about-science-research-fda/public-access-results-fda-funded-scientific-research.
6 UK Research and Innovation. Open research[EB/OL]. [2023－07－23]. https://www.ukri.org/what-we-do/good-research-resource-hub/open-research/.
7 国务院.国务院办公厅关于印发科学数据管理办法的通知[EB/OL]. [2023－04－10]. http://www.gov.cn/zhengce/content/2018－04/02/content_5279272.htm.
8 温亮明，李洋，宋佳.《科研数据北京宣言》解析[J]. 图书情报研究，2022，15(3)：90－97.
9 崔旭，赵希梅，王铮，等. 我国科学数据管理平台建设成就、缺失、对策及趋势分析——基于国内外比较视角[J]. 图书情报工作，2019，63(9)：21－30.
10 Re3data.Re3datasearch[EB/OL]. [2023－07－23]. https://www.re3data.org/search.
11 中央人民政府.科技部 财政部发布国家科技资源共享服务平台优化调整名单[EB/OL]. [2023－07－31]. https://www.gov.cn/xinwen/2019－06/11/content_5399105.htm? eqid＝ad105a8e000211b3000000036456e0b1.
12 Kratz J, Strasser C. Data publication consensus and controversies[J]. F1000Research, 2014(3): 94.
13 张文萍，宋秀芬，魏银珍，等. 基于 FAIR 标准的科学数据融合体系研究[J]. 中国图书馆学报，2020(6)：41－54.
14 Data Documentation Initiative. DDI lite-for DDI codebook version 2.0[EB/OL].[2023－04－10]. https://ddialliance.org/sites/default/files/ddi-lite.html.
15 Harvard Dataverse. Data citation－the Dataverse project[EB/OL].[2023－04－10]. https://dataverse.org/best-practices/data-citation.
16 ICPSR. Citing data-ICPSR[EB/OL].[2023－04－10]. https://www. icpsr. umich. edu/web/pages/datamanagement/citations.html.
17 刘小宇，纪雪梅，王芳. 我国科学数据共享平台的被引用状况调查与分析——以国家科技基础条件平台为例[J]. 情报杂志，2020，39(9)：168－175.
18 罗鹏程，崔海媛，聂华，等. 高校图书馆持久标识符应用研究[J]. 大学图书馆学报，2017，35(5)：108－116.
19 Science Data Bank. OAI 2.0 request results-SciDB[EB/OL].[2023－04－10].https://www.scidb.cn/oai? verb＝Identify.
20 张计龙，殷沈琴，张用，等.社会科学数据的共享与服务——以复旦大学社会科学数据共享平台为例[J]. 大学图书馆学报，2015，33(1)：74－79.

作者单位：浙江大学图书馆，浙江杭州，310058
收稿日期：2023 年 4 月 25 日
修回日期：2023 年 10 月 14 日
（责任编辑：关志英）

Exploration on the Development of Universities' Open Scientific Data Platform

—Taking Zhejiang University Open Data Platform as an Example

Yao Xiangyu　Huang Chen　Ge Hang

Abstract: The open scientific data platform is the information infrastructure for open data activities, construction of relevant platforms in universities is an important step in strengthening scientific data services. By investigating the current situation both domestically and internationally, the paper proposes issues to be explored in the construction logic, data collaboration functions and data harmonization standards of domestic platforms. Based on the practice of Zhejiang University, it explains the overall framework of independent platform developing path, the major issues of construction about collaborative processes and data harmonization, as well as the effectiveness and reflection of the practice, in order to provide reference for domestic universities to carry out relevant practices and promote the further implementation of domestic open scientific data activities.

Keywords: Open Data; Scientific Data; Research Data; Data Platform