DOI:10.13530/j.cnki.jlis.2020047

基于 FAIR 标准的科学数据融合体系研究*

张文萍　宋秀芬　魏银珍　李立睿

摘　要　本文在分析科学数据 FAIR 标准——可发现、可访问、可互操作、可重用的基础上，阐述这四项标准在实现条件上的依次递进、包含与被包含的层级关系，以此形成从低到高、包含四个层次的科学数据融合体系。基于这个层次结构，分别从“技术基础”以及“研究文化和数据管理制度”两方面探讨实现科学数据融合的体系架构，包括数据描述模型、数据服务模型，以及为规范数据融合实施过程和步骤而必需的数据管理计划、实施标准和评价体系。该体系架构可为实现跨组织、跨系统环境下的科学数据融合提供多层面的参考和借鉴。图 4。表 1。参考文献 20。

关键词　科学数据　融合体系　数据对象　FAIR

分类号　G203

FAIR-based Framework for Scientific Data Harmonization

ZHANG Wenping, SONG Xiufen, WEI Yinzhen & LI Lirui

ABSTRACT

Scientific data curation activities have started to realize the reuse of scientific data since decades ago and thus enhance the value of scientific data utility. However, there is no systematic and comprehensive summarization of the scientific data governance and effective use of scientific data. The architecture developed in this study integrating existing implementable solutions or proposing new ones from multiple levels is to clarify the multi-level problems that need to be solved in the process of scientific data governance from a global perspective, with the purpose to assist in scientific data curation and governance.

In order to achieve this goal, this study takes the FAIR principles (Findable, Accessible, Interoperable, Reusable) as scientific data curation basis and developed Scientific Data Harmonization System accordingly, which consists of a four-tier stacked structure with each consisting of specific requests of scientific data respectively, based on the explicit analysis of the relationships between four sub-items of the FAIR principles, which embodies related internal linkages and dependency. Based on this systematic framework, as well as the outcomes of implementation, management and research activities imposed on the field of

* 本文系国家自然科学基金青年项目“移动学术社区科研用户微知识持续协作行为及协同创新机制研究”(编号:71804153)和武汉科技大学人文社科高水平培育项目“知识网络的语义结构与社会结构的共演化研究”(编号:W201709)的研究成果之一。(This article is an outcome of the youth project “Study on Micro-knowledge Continuance Collaboration Behavior and Synergy Innovation Mechanism Oriented to Researchers in Mobile Academic Community” (No. 71804153) supported by National Natural Science Foundation of China and the project “Study on Co-evolution between Semantic Structure and Social Structure of Knowledge Network” (No. W201709) supported by Wuhan University of Science and Technology.)

通信作者:张文萍,Email zwp.wust@qq.com,ORCID:0000-0002-4807-3490(Correspondence should be addressed to ZHANG Wenping, Email:zwp.wust@qq.com, ORCID:0000-0002-4807-3490)

scientific data management to achieve or promote the achievement of scientific data management, this study first explores the data description model which is necessary for the fairlization of scientific data and optimization and improvement of the data service model from the technical perspective, and then discusses the data management plan and various standards and evaluation metrics that should be followed during the implementation of the data fairlization from the perspective of research culture and data management system; finally we discuss various environmental factors that may affect the implementation of data fairlization from the perspective of the external environment.

Besides functionalizing in data fairlization, the proposed system framework maybe be used to improve the operational practice of scientific data management for the departments of scientific research management and researchers from different domains; it can also be used in other data management fields, such as social media, national public service departments etc. In addition, practically, this study may be served as reference for the construction of regional or disciplinary open platforms for sharing scientific data.

The limitation of this study is that the investigation of the research basis for establishing framework of Scientific Data Harmonization System is not comprehensive enough, thus the resulted structure is relatively abstract and with less detailed practical procedures. Our later research work will focus on this direction to specify the operable procedures for each levels of data curation within the system. Meanwhile, the optimization and implementation of service models, and the impact of various other environmental and cultural factors on the integration of scientific data in the process of scientific data management are also on our research schedule. 4 figs. 1 tab. 20 refs.

KEY WORDS

Scientific data. Harmonization ecosystem. Data objects. FAIR.

0 引言

科学研究活动过程中产生的科学数据(Scientific Data)是科研传承的纽带和基石。科学数据,无论其表现形式或编码格式有何异同,都应该被严格而详细地加以分类、标注、索引、存储,以供其他研究者参考利用。然而,随着时间的推移,科学数据数量显著增长,数据类型多种多样,表示格式也更加复杂。同时,由于科学研究组织、机构的个体独立性以及科学研究者研究时长的有限性,科学数据呈现出横向组织层面以及纵向时间层面的分割片段保存状态,这就使得科学数据难以为后续同类研究提供支持和借鉴。

当前,众多国际研究合作组织针对科学数据的有效管理和利用提出倡议,其中最引人注目的是欧洲开放科学云(European Open Science Cloud,EOSC)提出的科学数据治理的 FAIR 标准[1],即可发现(Findable)、可访问(Accessible)、可互操作(Interoperable)、可重用(Reusable)。为了促进科学数据的共享和有效利用,实现科学数据可发现、可访问、可互操作、可重用的新目标和新共识,必须从包括科学数据创建、监护、存储等环节的整个数据生态系统出发,对科学数据描述、注册、服务等进行广义的管理和监护,使之能在大范围内,甚至跨组织、跨学科的环境下屏蔽不同数据源的差异,实现科学数据互访、互操作及重用。为实现这一目标,就需要跳出任意单一学科的规则和框架,站在整个科学研究领域的高度,对科学数据的表示格式、语义标注、身份识别、元数据以及各类数据服务,进行跨学科、高抽象层次的统一规范和部署,以实现科学数据的深度融合。

本文以 FAIR 标准为理论基础,以科学数据跨系统、跨学科的融合为目标,分别从技术基础角度、管理制度及文化角度,发现影响科学数据共享与再利用的多层面问题。换言之,以科学数据融合为目标,探讨科学数据治理与监护过程中的基础性、共性问题,从而服务于跨组织、跨学科的科学数据开放与共享。

1　科学数据融合的发展历程——大数据生态系统发展

科学研究的产出形式多样,如学术论文、科研数据、软件产品、专利以及行业标准或规范等。已有研究表明,尽管科研产出以丰富多样的形式存在,但主要形式有两大类:一是研究过程中产生的科学数据记录,二是科研结果的出版物,如学术论文和著作[2]。本文将科学研究过程中的产出物——科学数据作为研究对象,某些情况下亦称之为"数据对象",或更加概括地称之为"数据"。当然,必须将其与"元数据"区分开来,即一切描述"科学数据"或"数据对象"相关信息的数据,但事实上元数据本身也被包含在广义"数据"范畴之内。

目前,国内对科研产出成果的管理主要集中于科研出版物(纸质或电子形式)的管理与发布,如各类文献、专利数据库,而对于科学数据的共享、开放及重用的实践活动则鲜有涉足。与此同时,国外同行在这一领域的研究及实践活动则风生水起,最典型、最直观的体现是美国及欧洲多国建立了大量的开放科学数据平台,用于支持科学研究活动。

从科学数据研究所属学科的角度来看,目前关于科学数据监护以及如何实现其有效融合的研究和实践活动,大多集中于医学、生物(命)科学、物理、化学、材料科学等自然科学领域,已建成并运行的开放数据平台也几乎都是由这些学科研究者和研究机构倡议而创建,数据集的主体部分也主要来源于这类学科。目前已建成了较为专业的、聚焦于某一学科或者某一类相关学科的数据发布与开放访问平台,如主要致力于生物医学领域数据存储与共享的 caCORE[3] 和 TAPIR[4],主要存储有机化学蛋白质分类数据的 UniProt[5]。然而,随着大数据学科的发展以及对科学数据价值的重新认识,自然科学领域以外的人文社会科学也开始重视数据的治理、开放与重用。由此,一些综合性数据平台逐渐出现并致力于全学科数据的共享与访问服务,比如 CODATA、RDA、GO FAIR[6]、World Data System 等。专业化的数据平台由于其数据来源学科高度集中,所以数据集的异构化程度较低,数据集的描述性元数据和语义性元数据的标准更容易统一。这类平台通常更容易实现专业化、多功能、分类细致的查询功能和服务界面。与此相比较,全学科的综合型数据平台由于数据来源学科广泛,导致数据结构、格式及其元数据格式差异较大,平台很难提供精细化的搜索服务,从而增加了数据发现的人力成本[7]。

从科学数据研究或实践所着眼的问题来看,目前这一领域已有的研究或实践活动中,有的致力于解决跨平台数据访问的技术手段或工具,如 DATA TAPIR 平台制定并使用领域专用的查询语言,并要求所有加入平台的数据库都必须能够响应该语言发出的访问请求[3,4]。有些研究则从系统实现角度,通过定义满足特定需求的访问接口来实现用户数据访问,如 caBIO 平台通过提供基于 SOAP 技术的应用编程接口,实现跨数据集之间的数据整合和访问的通用界面[8];myGrid 和 BioMoby 平台利用规范化的领域本体描述互操作访问的服务接口和服务流程,从服务标准入手改善跨平台、跨数据集之间的数据访问[9,10];Figshare 平台提供开放的应用编程接口 API,支持用户基于其 API 进一步开发自动的数据集成服务。这些活动一定程度上实现了跨资源的数据发现和整合,但是由于实施的复杂性,对于加入平台的数据提供方而言,需要承担额外的负担;另外因为数据源所属专业的集中性,使得这两种实施方案并不适用于通

用数据平台的数据处理,缺乏普遍适用性。此外,还有些研究力图从实现数据融合的最底层——数据表示或描述以及数据模型的角度来规范数据的格式,从而在初始阶段即消除不同数据集间融合的障碍,便于跨数据集间的数据融合,如 SSWAP 通过使用语义网技术 RDF 和 OWL 描述数据集,以提高数据集内容的机器可识别度[11]。

尽管上述各类研究活动所针对的具体问题有所不同,但其最终目标无外乎立足于解决科学数据发现、访问、互操作以及重用过程中各个环节可能存在的问题。本文在仔细分析 FAIR 标准中的四项基本要求——可发现、可访问、可互操作及可重用之后,发现这四个方面并非各自独立、自成一面,彼此之间是层层递进与关联的关系。数据发现是整个体系的基础与前提,只有建立有效的数据发现机制,才能进一步实现对所需数据的访问;只有实现对所需数据的访问,才能在此基础上实现对数据的操作。可见,数据发现是基础,只有采取有效的发现机制,才能进一步访问数据;数据可访问是数据可互操作的基础,通过建立符合规范的数据访问机制,才能实施下一步的数据互操作,最终实现数据重用的终极目标。如图 1 所示,四项基本要求之间表现为包含与被包含的关系,内层是其相邻外层的实现基础,外层的实现必然包含内层实现所需的一切条件。故而本文将实现这四个层次的科学数据处理称之为“科学数据融合”,其包含了四个递进关系的科学数据处理阶段,体现了科学数据融合从低层到高层的融合深度。

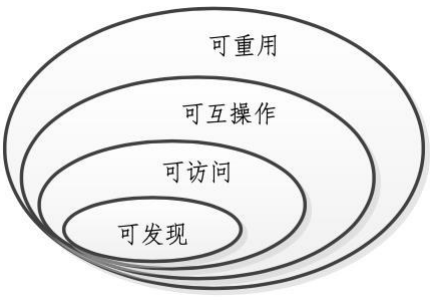

图 1 FAIR 标准中四项基本要求之间的相互关系

综上所述,本文以 FAIR 标准所包含的四项标准及其递进层次关系为科学数据融合体系的主轴,详细探讨实现此四项标准所涉及的各个对象及其属性、各项服务标准及技术要求,力图构建整个科学数据融合的体系架构,以期为跨系统、跨平台的科学数据共享与重用提供实施基础。

2 科学数据融合的系统架构

现有实践经验表明,由于不同学科产生的数据或数据对象在语义及语用方面存在差异,数据共享及互操作的需求主要发生在学科范围内或相邻学科之间,因此,学科间差异越大,共享的需求往往就越低[12]。可见,数据的规范化表示和标识以及在此基础上实现的共享和重用,首先应该在学科范围内达成一致,而相邻学科之间可能存在少量共享及互访的需求,数据只需要在一个较为基础的层面实现格式和描述的统一即可。

EOSC 工作组在 2017 年提出了科学数据治理的 FAIR 标准,并给出了这四个标准的详细解释,但是并未给出具体的实施细则,即如何处理科学数据或者采用何种操作机制或服务模式,使得不同来源的数据集在用户界面上能够按照用户的使用需求无缝地集成并呈现(可发现);当数据满足用户需求时,用户端可以通过何种机制定位并连接到数据集(可访问);数据集及其元数据采用何种表示格式,以便于人机操作处理(可互操作);数据集应该包含何种附加信息,以降低使用过程中的歧义(可重用)。上述四个层次的实现需要技术基础层面的多方面支持与实施,例如,如何描述和标识数据,如何在网络空间中定位数据,数据和元数据如何做到规范化的统一,等等。然而,技术基础层面的顺利实施必然离不开数据管理制度与标准上的重要转变,如政策层面的数据管理计划、标准和评

价机制辅助并配合技术基础层面的实施,从而实现科学数据的共享与重用。因此,本文以 FAIR 标准的四项递进需求为主线,分别从技术基础层面和数据管理制度与标准层面探讨科学数据融合的体系架构,从而为实现科学数据融合构建全局视图。

2.1　实现科学数据融合的技术基础支持

根据前文分析可知,实现 FAIR 标准的四个子项(可发现、可访问、可互操作、可重用)需要具备的条件是层层递进、彼此关联的。在已有的关于科学数据的研究中,研究的问题虽然广泛涉及整个科学数据领域的不同方面,但均致力于直接或间接地实现这四个子项中的一个或多个,并在各自研究中取得一定的成果。无论是解决科学数据开放与共享中的哪一个子项,解决问题的途径总是围绕两个方面展开:一是通过采用特定的治理与监护标准,赋予原始数据丰富的附加信息以及更加规范的外在表示形式,从而便于外界的识别、发现或者访问、操作,即从丰富和完善数据描述模型的角度为实现不同层次的数据融合提供基础信息准备;二是采用更合理的服务访问模式,使得无论是人工处理还是机器自动处理系统的数据都变得更容易、便捷,即数据服务模型的优化。下文将从这两方面探讨实现数据融合的数据描述模型和数据服务模型。

2.1.1　面向融合的数据描述模型

为实现数据融合而进行的数据治理与监护工作的主要内容是赋予原始数据多方面、多角度的特征值,这些值分别描述了原始数据的各项属性与特征,可称之为数据画像(Data Profile)。由于涉及的描述面较广,一个完整的数据画像可能包含多种类型的元数据。对于被治理与监护的对象而言,一个完整的数据画像的组成结构形成了该类对象的数据描述模型,不同对象或不同应用领域对数据描述模型的结构需求不尽相同,但是一个完善的数据描述模型应该包含足够多的描述面(Facets),从而使得被描述对象附带必要的信息,以被外界充分了解并获取和利用。由于科学数据治理的最高目标是实现数据的共享与重用,因此本文考察科学数据为满足 FAIR 标准的四项要求分别需要哪些描述面的信息,即哪些描述信息的存在使得科学数据更容易被发现、访问、互操作和重用。

(1)可发现。为了使湮没在数字海洋中的任何数据对象(包括数据集、研究成果等)容易被发现,首先,数据对象需要一个可以将它与其他任何对象区分开来的唯一身份标识符(Identifier),这样的标识符命名规则目前并未在全因特网范围内达成统一,通常 Web 页面使用 URL 作为标识符,数字出版物采用 DOI 或 URN,而有些数据管理机构则使用自有的方法命名其数据对象,比如使用机构的域名加上机构内部规定的编号或名称作为资源的标识符。不管使用何种方式,重要的是做到全网唯一不重名。其次,这样的身份标识符通常只是一个字符串,并不包含任何有关数据集来源学科或研究方向的信息,因此数据对象需要被赋予尽可能丰富的描述其所属知识分类的元数据,以使来自同领域的其他研究者更容易了解其来源学科背景和知识类目。这样的信息以机器可处理的格式向外界发布,可使得该数据对象容易被外界发现。综上,这类描述数据对象基本信息的元数据在构成上完全可以在全学科范围内达成一致,即统一的元数据标准,也称之为描述性元数据。

(2)可访问。采取上述描述机制后,数据对象可被外界发现,无论这样的发现过程是由用户经过浏览器发起或者由网络中的 Web Services 进程在一定的条件下自动发起。在可发现的基础上要实现数据对象可访问。首先,数据对象需要被赋予至少包含以下信息的管理型元数据:数据对象的标识符、存储位置、访问权限说明、下载及传播的许可条例等信息,不论是描述型元数据还是管理型元数据,都需要在数据管理系统中注册和索引,以便于外界利用并实现访问;其次,数据对象所在系统应该能被外界使用通用网络协议访问,如 HTTP 协议,这样的

访问协议应该是开放且被广泛使用的,便于多数据平台之间的数据交流与共享。

(3)可互操作。根据 FAIR 规则对跨系统的不同数据集间的互操作定义,可以发现这样的互操作需要从两个层面展开:语义层面的互操作和技术层面的互操作,语义层面的互操作是为了实现数据集内容的有效共享和利用,而技术层面的互操作则是为了保障语义层面的互操作的实施。为实现语义层面的互操作,数据集需要采用领域权威的本体及受控词表来进行语义标注,其元数据结构也必须采用领域权威的元数据标准(Metadata Schema),从而实现对数据集的精准描述和概念定义;技术层面的互操作则需要数据集本身及其元数据在表示层面使用标准的编码格式进行有效的知识表示,以方便数据处理系统之间无障碍的数据存取。为实现互操作而赋予数据对象的附加信息称之为操作性元数据。

(4)可重用。数据对象的重用是最终发挥其价值的阶段。数据对象在实现了可发现、可访问和可操作的情况下,决定其最终能否满足用户需求、发挥其价值与功能的关键在于该数据对象的来源信息(Provenance)、使用情境信息(Contextual Information)是否契合用户需求。这就要求对数据的来源有详细的信息记录,包括数据是遵循何种调查协议或实验过程得来的,或者借助于何种技术设备或特殊工艺得来的,数据是原始数据还是经过某种处理或变换过程获得的,以及明确说明的数据使用许可协议。而数据使用情境信息则应该明示使用该数据应具备的前提条件和注意事项,以避免数据误用带来的错误性结论。因此,这类被赋予数据对象的信息被称之为功能性元数据,有助于数据使用者判定获得的数据是否可用于其研究过程中。

总结上述四个部分,可得到对应于不同融合层次的数据描述结构模型(见图 2)。

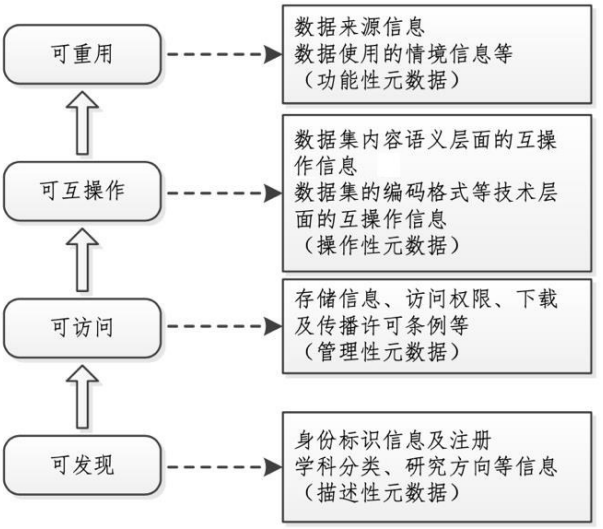

图 2 面向融合的数据描述模型

分析上述为实现不同层次数据融合而设计的数据描述模型,可以发现,越是处于较低的融合层次,融合的实施需求越具有共性,越容易在较大范围内实现实施标准的统一。因为底层实现的是数据资源的标识和定位,即便是来自不同领域的数据资源,在元数据结构上很容易达成统一。越是高层越难以形成统一,因为高层解决的问题是如何操作和利用数据集

的内容,来自不同领域的数据集在数据内容的描述方面很难达成统一标准,即使来自相同领域却由不同研究机构创建的数据资源也常常因使用的学科词表或本体不统一而出现语义上的分歧。由此可见,在图 2 所示的层次结构状的数据描述模型中,底层的共性和高层的个性决定了数据描述模型从下至上无法全部实现大范围内统一的实施标准,其必然遵循底层大范围内统一、高层小范围内自治的模式,这样才能最大程度地兼顾数据融合的需求以及学科的自由性。

2.1.2　面向融合的数据服务模型

数据描述模型为实现科学数据融合做好了必要的前期准备,即在数据治理与监护过程中,数据可携带通用的标识和各类丰富的元数据,这些丰富的附加信息使数据需求者能够更容易地发现、访问、操作和利用数据。经过高质量处理后的数据对象存储于网络数据平台中,用户通过何种机制发现自己需要的数据,数据平台如何最大程度地协助用户发现需要的数据,用户、服务、数据资源平台之间如何基于互联网协调实现数据的有效访问,基于这些问题,下文将探讨更易于实现数据发现及访问的数据服务模型。

为了使研究机构产生的数据更容易被发现,需要改变过去用户逐个访问每个数据平台搜索数据或者依赖搜索引擎定位到某一专有数据平台的服务模式,即传统的数据服务模式。如图 3 所示,任一研究机构产生的科学数据及其元数据(使用机器可读的 XML 格式表示)一起存储于各机构专有的数据平台上,各平台的元数据被搜索引擎爬取并加以索引,这样用户可以通过搜索引擎找到这些数据集并分别访问;或者用户直接访问某一专有数据平台,从而获得其存储的科学数据,当然这建立在用户对所访问的专有数据平台充分了解的前提下。这种模式下,用户最终需要分别访问每个专有数据平台(即便通过搜索引擎发现多个专有数据平台,也需要一一访问之),以获取自己需要的数据。

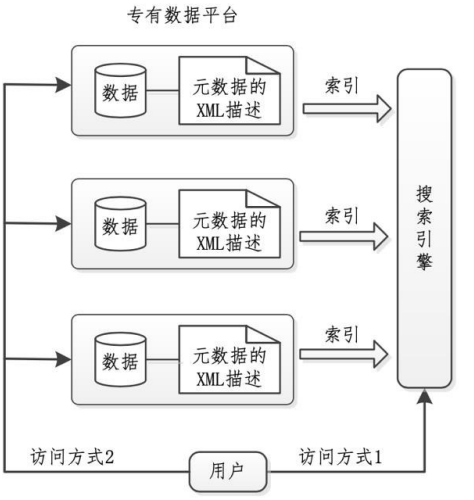

图 3　传统的数据服务模型

与传统的数据服务模式不同,在新的数据服务模型中,为实现数据资源更容易被发现和访问,各专有数据平台上数据资源的描述信息需要彼此之间实现互访、获取并存储备份,形成各自的元数据目录,以便于任何客户端,无论是搜索引擎还是某个综合性数据平台或者各专有数据平台,都能够获取元数据信息,存储在索引表中或者形成元数据目录。如图 4 所示,在面向融合的数据服务模型中,用户发现数据的方式有三种:一是通过访问某个综合性数据平台上的元数据目录,可以实现一个用户界面访问多个专有数据平台的数据;二是通过搜索引擎发现某个综合性数据平台;三是发现某个专有数据平台,由于专有数据平台间的数据目录共享,同样可以获得极为丰富的可用数据集。这种新的服务模型极大地拓展了数据发现与贡献的渠道,使数据的共享与重用得以最大程度地实现,但需要解决的核心问题是元数据获取的实现机制以及元数据目录的形成机制和组织结构。当下元数据获取使用的协议或应用编程接口常用的是 OAI-PMH 或 RESTful API,二者均基于 HTTP 协议实现跨系统的元数据获取。数据提供方通过协议或接口向外界呈现自己的元数

据目录,数据请求方通过发起服务请求来获取元数据,然后提供方将元数据发送到请求方,请求或响应进程均通过封装在 HTTP 协议中发送。

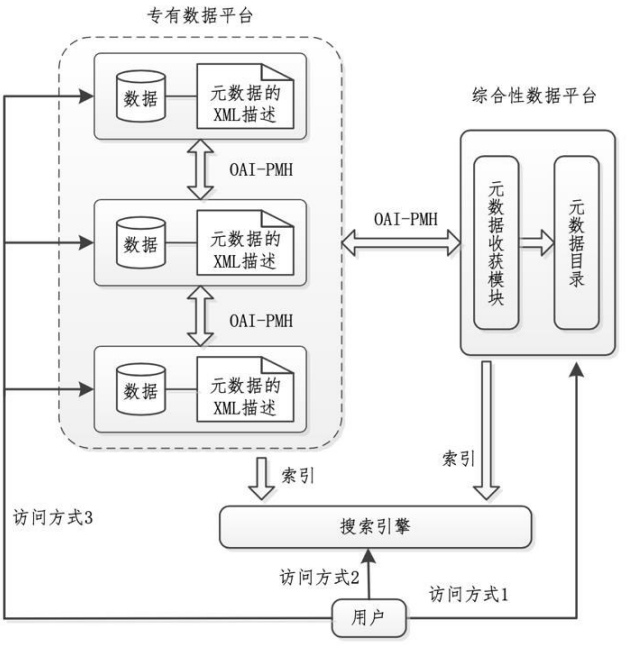

图 4 面向融合的数据服务模型

2.2 面向融合的数据管理制度与标准

为了更好地实现科学数据融合的目标,除了上述技术基础层面的各项实施要求之外,在整个数据生命周期内,科学数据治理、监护、共享与互操作的实施环境与管理制度,系统中所有参与者对科学数据融合的意义和价值的理解与认识,数据融合系统各组成部分及各项操作需要遵循的标准和评价体系等,对于实现科学数据的融合同样具有重要的影响。

近年来,世界各国尤其欧美发达国家的学术领域兴起“开放数据”“开放科学”以及“开放获取”等关于科研数据和科研成果共享的潮流,越来越多的大学、科研机构以及政府智库创建各类非营利性数据与资源共享平台。与此同时,越来越多的研究基金资助方要求被资助的研究组织在提交基金申请时必须提供详细的科研数据开放与共享计划书[13]。在这样的时代背景下,科学数据开放共享的文化氛围逐渐形成,同时与之配套的数据管理制度和各类评价标准也日趋成熟。

2.2.1 数据管理计划

由于科学数据治理与监护的复杂性,科学数据从产生之初就需要专业管理人员按照事先制定的数据管理计划,有步骤地执行有效的数据治理与监护工作,以规范这项工作的执行过程。以前这项工作一直被认为是可有可无,甚至是个负担。现在越来越多的研究组织与机构意识到数据开放与共享的重要性,以及为实现数据共享与重用而实施的科学数据管理的重要性。因此,围绕科研过程中如何制定和实施有效的数据管理计划(Data Management Plan,DMP)也显得尤为重要,并成为科研项目启动前的必要事项。

根据“欧盟研究与创新”(EU Research and Innovation)资助项目 Horizon 2020 的规定,数据管理计划(DMP)应详细规定为实现数据重用在

数据生产/收集、处理等环节中应该完成的一系列操作任务[14]。这一系列操作任务使得数据对象被赋予大量的基础信息(元数据),包括数据本身的基本信息,数据的生产、维护和保存的情境信息,为实现重用而赋予数据的各类附加信息,等等。DMP 为整个数据生命周期中的治理与监护提供了指导依据和实施步骤。

自数据"开放获取"(Open Access)思想诞生以来,很多公共基金资助的项目被要求提交关于研究数据的管理计划。对于数据管理计划(DMP)具体应该包含的内容,不同基金提出的要求不尽相同,但其宗旨几乎是一致的,即公共基金资助的研究项目产生的数据和成果应该对外界开放,并且应该以容易被发现、访问、操作和重用的形式对外发布[8]。

早期的数据管理计划(DMP)更多关注研究数据以及学术论文的存储,并不关注数据是否便于被外界用户使用,或者为方便用户的使用该对数据进行何种处理,也不注重元数据以及数据处理过程中使用规范与标准的说明[15],这对于后续的数据重用极为不利。各国的公共基金管理者意识到这个问题后,开始细化数据管理的要求,如英国政府基金会要求研究机构提交的数据管理计划中明确数据的类型和格式、元数据的标准、数据的存储方案、数据安全与共享政策[9];美国国家科学基金会(National Science Foundation,NSF)要求数据管理计划(DMP)中详细描述数据、数据存档信息、访问与共享的细节、数据文件的分布、元数据信息、知识产权信息、隐私与伦理、数据存储方案等[13];美国能源部对其资助的研究项目要求在数据管理计划中提供以下对象的信息:研究数据、元数据及其标准、数据格式、数据存档及标注、数据共享策略与限制、数据共享所需要的资源、知识产权、敏感信息的保护等[16];澳大利亚研究委员会对其资助的研究数据和成果也制定了开放访问的要求,包括存储平台的开放性,为外界访问和重用提供充分的信息、策略和工具,等等[17]。综合分析多个国家和部门的公共基金会对于其所资助项目的数据管理计划的要求,以及 Horizon 2020 制定的数据管理计划框架[18,19],发现数据管理计划(DMP)中要求提供的信息以及施加于数据之上的操作,按其作用或实现功能划分,均可归属于满足数据融合四个层次中的一个或多个层次(见表 1),最终数据管理计划(DMP)实施过程中,赋予数据的附加信息共同促进了科学数据的可发现、可访问、可互操作及可重用的深度融合。

2.2.2　科学数据融合体系中各类标准、规则与评价体系

数据管理计划(DMP)是为了规范数据创建者或管理者治理与监护数据的操作任务与流程,明确需要施加在数据对象上的处理任务以及被赋予的附加信息。然而,这些处理任务应遵循什么标准,使得被处理的数据更容易被重用,数据的附加信息应该满足何种标准,以利于数据的重用。数据管理计划(DMP)的实施条目在执行过程中,有些条目可以遵循现成的通用执行标准,这些标准并不区分数据来源的学科性质,具有普适性;有些条目拥有现成的行业执行标准,可以依照执行,不同的学科有不同的行业标准或规则;还有些条目在执行时没有可依据的执行标准或规则,但存在一定的评价体系或指标,可以用来对执行结果加以评定。无论是前期的执行标准或规则,还是后期的评价指标或体系,对科学数据的治理与监护都起到至关重要的作用。

本文无意一一列举数据管理计划(DMP)中每项实施条目对应的执行标准或评价指标,仅从通用化和专业化两个角度考察表 1 中"具体实施条目"的执行标准或评价指标。显而易见,在数据处理过程中,选用业界接受程度越高的指标或标准,数据越容易实现共享与重用。数据管理计划(DMP)中第一部分"数据基本信息",即描述数据集基本信息的基础元数据,对于来自任何学科的科学数据对象而言,这部分信息的构成完全可以在全学科范围内达成统一结构并形成通用标准,部分分支信息如"学科信

表 1　数据管理计划实施内容分析

	具体实施条目	条目详解
数据基本信息	项目信息	数据所属项目的基本信息,如项目编号、项目标题、项目描述(关键词)、项目资助方、项目受助方、起止时间等
	学科信息	数据来源项目所属的学科信息
	数据类型	文本、数字表格、图片等
	数据获得途径	说明数据获得的方法,如观测、实验、仿真、推导或编译等
	数据获得工具	数据获得过程中借助的工具,如软件程序、实验设备器械及耗材、仿真算法等
	数据收集目的	数据在研究项目中的具体作用,如为了测定或检验某个指标
数据可发现实施细则	数据集标识符	制定标识符分配规则,选择标识符注册、发布方式和平台
	描述型元数据	赋予数据集体现其含义的语义元数据;赋予数据集所属项目及学科的项目关键词及学科分类元数据
	元数据标准	数据集的元数据是遵循何种元数据标准创建的
数据可访问实施细则	数据存储方式	数据集、相关元数据及数据集描述文档的存储平台选择
	数据访问地址	数据集的 URL,用于实现访问定位
	数据访问规则、限制	针对用户身份制定不同级别的数据访问权限;针对不同数据集制定不同的开放访问级别
	数据访问工具和访问模式	取决于数据集存储的系统环境能够提供的访问方式
数据可互操作实施细则	数据格式	选择数据编码格式,如 txt\doc\pdf\XLSX\csv\jpg 等
	数据项词表	应来自专业化的标准词表或本体,实现不同标准词表映射
数据可重用实施细则	数据精度与时效	说明数据量化的标准、精度以及使用时效
	使用情境	说明该数据集可被重用的研究情境或使用场景,以及使用注意事项
	数据使用许可	明确数据引用标注,对于营利性和非营利性用途的规定

息”,不同国家有成熟的标准,如美国有杜威十进制分类法;“数据类型”则仅有有限的几种通用类型,可以做到一一列举;“数据获得途径”和“数据获得工具”也有可识别的、能在大范围内达成普遍接受的选项集合。第二部分“数据可发现实施细则”中“数据集标识符”目前常见的是使用 DOI、URI、URL 以及 URL 结合标题或文件名的组合,任何一种均可用于任意学科数据对象的标识;至于“元数据标准”,目前无论一般性基础元数据还是专业性元数据,均有大范围内或者至少学科领域内普遍接受的元数据标准;而对于数据集内容的“描述型元数据”,在选定的元数据标准下,用于描述数据集的词表或本体通常也有常用的标准,并且能在领域范围内实现标准化。第三部分“数据可访问实施细则”中“数据访问地址”现今几乎都是客户端浏览器可访问的 URL 地址,“数据工具与访问模式”几乎都是基于客户/服务器模式的 HTTP 访问模式,这些几乎构成了业界的统一标准。第四部分“数据可互操作实施细则”中“数据格式”目前包含若干种较为通用的格式,且均为跨学科通用标准;“数据项词表”则只能在专业领域

内实现标准化,并需要解决不同标准之间的词表映射问题。第五部分“数据可重用实施细则”中无论哪一项都很难形成跨学科通用标准,均与数据来源学科息息相关。

从以上分析可以发现,在整个数据融合体系中,越是处于底层,针对数据的基本操作越容易实现标准化;越是处于上层,针对数据内容的处理和使用操作越难以实现全学科范围内的标准化,往往只能在领域内部形成标准,甚至无法达成一致标准。然而作为科学数据管理的任何参与方,需要注意的是经标准化处理的数据对象更易读、易理解、易被共享及利用,无论对人还是对机器来说均是如此,因而尽可能遵循统一标准或评价体系是至关重要的。

3　科学数据融合的环境因素

实现科学数据融合,不仅需要一系列技术操作层面的数据治理与监护工作、新型数据服务模型的支持,还需要为了更好地实施数据的治理与监护而制定的数据管理计划、标准和评价指标等。与此同时,诸多科学研究环境因素也影响着科学数据的有效管理与利用,如学界对于科学数据开放使用可能带来的利益前景的预期,以及由此对数据治理与监护工作动力的影响,科学数据共享平台长期可持续的运营模式,数据管理与维护活动中相关人员的知识与技能,等等。

3.1　科学数据融合的制约因素

为实现科学数据重用而实施的数据治理与监护工作给科学研究机构带来了额外的工作负担,因而并非所有研究机构都热衷于数据的善后处理工作。有些研究机构对此无任何作为,有些则遵循长期以来形成的一套符合行业内部需求的数据管理标准和制度,使得数据的大范围重用受到一定限制,当然也有些研究机构认识到数据重用的价值,遵循 FAIR 标准并开始实施这一工作。

当前在科学数据领域内,大范围遵循 FAIR 标准治理数据的制约因素有三个。一是对于科学数据重用价值的认识程度。针对这个问题,目前很多知名数据管理平台和论坛开展了对数据重用实践案例的整理,并依赖平台广泛促进跨组织、跨项目间共享数据的合作研究。二是数据治理工作的复杂性、实施要求和步骤的明确性和易操作性。FAIR 标准只是定义了数据可发现、可访问、可互操作及可重用的含义,并未明确实现四个子项时该遵循何种标准,现有研究的焦点往往分散于整个体系中的某一子项,因此需要在大范围内形成数据治理标准的共识,使得数据治理与监护工作有章可循。此外,为减轻数据管理任务的繁琐性,数据管理平台在智能化和工具化方面有很大的发展空间,简而言之,就是提高软件管理系统辅助数据处理的参与度,减轻工作人员负担,相应地也就提升了科研机构数据治理与维护的意愿。三是数据治理工作价值的社会认可度。当前学术机构用代表研究成果的论文和专著的发表及引用来评价学术成就,而数据治理与监护工作的成果难以量化,其价值亦难以认定,对此已有研究者和相关机构提出关于科学数据引用规则的联合申明[20],从而提升学术界对科学数据价值的认同。

3.2　科学数据共享平台的可持续运营模式

目前世界多数发达国家的大学、专业研究中心、企业或政府智库在“开放科学”(Open Science)思想的指导下纷纷建立开放数据平台,向外界提供开放的数据访问与利用服务。由于这样的工作属于非营利性质,并不能带来直接的收益,因此如何实现数据平台的可持续运营需要合理而又经济的规划。

这类开放数据平台有综合性数据平台和专业性数据平台。综合性数据平台通常由国家层面机构或者跨国联合机构负责出资创建,如欧洲开放科学云平台(EOSC)、欧洲十国共同创建的通用数据平台 OpenAire、英国国家基金会资助的非结构化数据平台 Figshare。专业性数据平

台通常由数据来源学科的联盟出资创建,平台创建的资金通常来源于各类科学研究基金,而平台建成后长期不断的数据维护以及各类数据服务的持续提供还需要源源不断的资金支持。除了接受来自各类机构、组织的基金资助之外,找到一种可持续的运营模式,对数据平台的长期发展至关重要。综合现有数据平台的运营模式,以及借鉴各类网络知识平台的运营模式,可总结以下常见的运营策略:①从研究项目资助基金里预留出一部分,作为该项目研究产出数据的专用维护费用,如果数据存储在第三方平台,这部分费用将支付给该平台;②数据资源提供者通常也是数据资源需求者,鼓励资源提供者负责其所提供数据的前期治理与监护,以换取其他数据资源的使用权限,从而减少平台管理方在数据后期的治理与监护中的支出与花费;③采用付费模式,由数据提供方向平台支付一定的存储与管理费用,或者由数据使用方支付一定的使用费;④采用平台会员制,明确会员的权利与义务,鼓励会员参与平台数据的管理和维护,实现数据资源的共享与共治。

以上运营策略主要源于 Web 2.0 的资源共建共享的思想,旨在最大程度地降低平台后期的运营维护成本,以实现科学数据平台的可持续运营和发展。当然,由于众多数据平台的种类及所有权的差异,具体选择何种运营模式还需要根据自身的运营环境做出决策。

3.3 科学数据管理与监护的技能需求

科学数据服务于科学研究的前提是数据被有效地管理与监护,即从数据创建之初的原始数据,到以通用格式表示、被赋予丰富的描述及管理信息,到存储在管理规范、服务完备的数据平台、以易发现的形式发布其元数据,这个过程需要具有专业技术知识和技能的人员来操作完成。因此,管理人员的知识与技能结构深刻地影响着数据管理与监护的质量。

首先,数据管理人员需要具有与 IT 技术相关的信息或数据处理技能,能熟练地使用各种数据管理工具和软件,如数据库管理系统、数据处理软件,了解基于万维网的数据服务模式,具备初步的服务器端数据库管理能力。其次,数据管理人员还需要具备一定的数据管理理论知识,如数据组织与数据结构,知道如何构建数据表结构以形成结构合理的数据对象,了解一些通用的知识组织系统,能够实施有效的数据分类管理并赋予数据以丰富的元数据信息,甚至还涉及数据管理标准的制定以及数据服务模式的设计等工作。最后,数据管理人员的学科背景知识对于其工作胜任力有很大影响,即在掌握上述技能的基础上,如果管理人员掌握数据来源学科的专业知识越深厚,他对于数据对象的语义、结构以及表示格式就会有越精准的把握,数据管理工作的质量也就越高。

4 总结

科学数据的共享与重用在科学研究中的重要作用已得到各学科的高度认同,也基本在世界范围内达成共识。但是,受制于各学科不同的研究文化以及各研究组织和机构关于科学数据不同的处理理念和处理方式,科学数据共享与重用面临重重障碍。欧洲开放科学云平台(EOSC)制定的 FAIR 标准指明了实现数据共享与重用的具体实施标准,本文在分析这四个标准的内在联系及实施要求的基础上,阐述其层层递进及包含与被包含的关系,并在此层级结构上确立从低到高的多层次数据融合体系。同时,进一步从技术基础的角度探讨构建数据融合的描述模型,改进数据服务模型;从研究文化和数据管理制度角度探讨了实施数据融合所必需的数据管理计划以及实施过程中应该遵循的各类标准和评价指标;从外围环境角度讨论了可能影响数据融合实施的各种环境因素。

建立科学数据融合体系架构是为了从全局角度认识科学数据共享与重用过程中需要解决的问题,并整合现有的、可实施的解决方案,探

讨寻找新的解决途径,以期为科学数据的治理与监护提供全局视图。后续研究将专注于数据融合过程中不同层次的技术实现及其效用检验和评价,服务模型的优化与实现,以及各类环境和文化因素对科学数据融合的影响和作用,等等。

致谢:武汉大学信息管理学院邓仲华教授给予了研究选题指导,武汉科技大学恒大管理学院涂静和张凌老师提供了研究方法建议和协助,湖北省图书馆刘伟成教授提供了关于数据科学领域发展及研究动向的咨询和建议,在此一并致谢!

参考文献

[1] 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.

[2] Asmi A,Cordewener B,Goble C,et al. Report on data interoperability:findability and interoperability[R/OL]. [2019-08-20]. https://eoscpilot.eu/content/d63-1st-report-data-interoperability-findability-and-interoperability.

[3] Covitz P A,Hartel F W,Schaefer C F,et al. caCORE:a common infrastructure for cancer informatics[J]. Bioinformatics,2003,19(18):2404-2412.

[4] De Giovanni R,Copp C,Hobern D,et al. TAPIR-TDWG access protocol for information retrieval[R/OL]. [2019-08-20]. http://www.tdwg.org/standards/449.

[5] UniProt[EB/OL]. [2019-10-02]. https://www.uniprot.org/.

[6] European Commission Expert Group. Turning FAIR into reality[R/OL]. [2019-08-15]. https://ec.europa.eu/info/sites/info/files/turning_fair_into_reality_1.pdf.

[7] Wilkinson M D,Verborgh R,Clark T,et al. Interoperability and FAIRness through a novel combination of web technologies[J]. PeerJ Comput. Sci.,2017,3:e110.

[8] DiPersio D,Cieri C,Jaquette D. Data management plans and data centers[EB/OL]. [2019-07-11]. https://www.aclweb.org/anthology/L16-1396.pdf.

[9] Digital Curation Centre. Funders' data plan requirements[R/OL]. [2019-06-21]. http://www.dcc.ac.uk/resources/data-management-plans/funders-requirements.

[10] BioMoby [EB/OL]. [2019-10-20]. http://biomoby.open-bio.org/.

[11] Gessler D D,Schiltz G S,May G D,et al. SSWAP:a simple semantic web architecture and protocol for semantic web services[J]. BMC Bioinformatics,2009,10:309.

[12] Wise J,DeBarron A G,Splendiani A,et al. Implementation and relevance of FAIR data principles in biopharmaceutical R&D[J]. Drug Discovery Today,2019,24(4):933-938.

[13] National Science Foundation. NSF data sharing policy[R/OL]. [2019-08-20]. http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/aag_6.jsp#VID4.

[14] Grootveld M, Leenarts E, Jones S, et al . OpenAIRE and FAIR data expert group survey about horizon 2020 template for data management plans[R/OL]. [2019-08-12]. https://doi. org/10. 5281/zenodo. 1120245.

[15] Lee D J, Stvilia B. Practices of research data curation in institutional repositories: a qualitative view from repository staff[J]. PLoS One, 2017, 12(3): e0173987.

[16] Office of Science, US Department of Energy. Statement on digital data management, suggested elements for a data management plan[EB/OL]. [2019-08-20]. http://science. energy. gov/funding-opportunities/digital-data-management/suggested-elements-for-a-dmp/.

[17] Australian Research Council funding rules for schemes under the discovery program (2015 edition) [EB/OL]. [2019-08-20]. https://www. legislation. gov. au/Details/F2015L01468.

[18] Monika K. Guidance for creating a data management plan in Horizon 2020 projects[EB/OL]. [2019-08-20]. http://dx. doi. org/10. 14279/depositonce-7199.

[19] European Commission Directorate-General for Research & Innovation. Guidelines on FAIR data management in Horizon 2020 [EB/OL]. [2019-11-20]. https://ec. europa. eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en. pdf.

[20] Joint declaration of data citation principles-final[EB/OL]. [2019-10-29]. https://www. force11. org/datacitationprinciples.

张文萍 武汉科技大学恒大管理学院副教授。湖北 武汉 430080。

宋秀芬 湖北警官学院讲师。湖北 武汉 430034。

魏银珍 黄冈师范学院副教授。湖北 黄冈 438000。

李立睿 西南大学计算机与信息科学学院副教授。重庆 400715。

(收稿日期:2019-12-28;修回日期:2020-05-08)