·“双清论坛”专题:开放科学的实践与政策·

建立权责明晰且能力健全的科学数据开放共享机制
——以高能物理领域为例

齐法制　　陈　刚*　　程耀东

(中国科学院高能物理研究所, 北京 100049)

[摘　要] 高能物理(也称粒子物理)研究一直处于物质科学的最前沿,大科学装置往往是物理基础研究和满足国家战略需求的国之重器,为基础及应用研究提供了重要的平台。大科学装置产生的数据是一座极为重要的科学金矿,而科学数据的开放与共享是科学数据效益最大化的必要条件。前沿物理大科学装置包括面向特定学科的专用研究类装置和服务于多学科交叉前沿的公共服务平台类装置,两类装置科学数据构成大体一致,均包括实验数据、模拟数据以及文档、成果和专利等数据,但在数据主权和共享机制上则存在较大差别。专用研究装置和数据采用合作组模式,合作组有国内外科学家共同参与组成,并在合作组框架下实现科学数据共享和利用。公共服务平台类大科学装置数据管理和共享则仍然没有相应的管理规定和规范,需要结合我国科学数据管理办法和领域特点,进一步开展相关研究、探索和实践。科学数据的开放共享应遵循“尽量共享、不得已才受限”的原则,推动科学数据的效能最大化,并在经费保障、技术研发和人才队伍等方面给予支持。

[关键词] 大科学装置;数据管理;数据共享;数据保存;数据所有权;数据使用权限;数据标识

科学大数据一般由大科学装置和大科学项目产生,具有不可重复性、高维性及计算分析高复杂性等特征。大科学装置是国家科技创新体系的核心单元,是我们国家综合国力的象征,它对国内外高度开放,广泛采取了国际合作的机制。有的大科学装置面向特定学科的专门前沿研究,有的装置属于多学科交叉前沿的公共研究平台。大科学装置规模大,建设和运行周期长,不仅是基础研究和应用基础研究的前沿,也是高技术研发的中心,它们每年产生TB级或者PB级甚至数十PB级的海量数据。这些海量数据已经成为大科学装置的核心资产,是科学研究的第一手资料,也是产生高质量科学成果的金矿,因此科学大数据尤为珍贵。

随着科学数据的不断积累,基于大数据的科学发现已经成为继实验归纳、模型推演、仿真模拟之后的科学研究第四范式,并引起相关国家和科技领域的高度重视。数据开放和共享是发展科学大数据的关键,必须实施可持续发展的科学数据共享,包括重视科学数据出版等数据集成与开放共享机制。

1　背景及国内外现状

为进一步加强和规范科学数据管理,保障科学数据安全,提高开放共享水平,更好支撑国家科技创新、经济社会发展和国家安全,国务院办公厅于2018年3月17日印发了《科学数据管理办法》[1],该办法首次站在国家高度、面向多领域科学数据,提出开放为主的指导原则,具有划时代意义。办法要求科学数据管理遵循分级管理、安全可控、充分利用的原则,明确责任主体,加强能力建设,充分体现“谁拥有、谁负责”“谁开放、谁受益”。同时,我国在2018年还开始遴选和建设高能物理、空间天文、地球科学等一批有重要影响的国家科学数据中心,促进科学数据开放共享。作为《科学数据管理办法》的具体落实,2019年2月,中国科学院办公厅发布了《中国科学院科学数据管理与开放共享办法(试行)》。该文件为进一步加强和规范中国科学院科学数据管理,

收稿日期:2019-02-12;修回日期:2019-02-25

* 通信作者,Email: cheng@ihep.ac.cn

保障科学数据安全，提高科学数据开放共享水平，促进科技创新和经济社会发展提供了具体指导意见。办法规定科学院法人单位是科学数据管理与开放的责任主体，需建立科学数据管理及共享体系。所有科学数据必须向项目指定的科学数据管理机构汇交。科学数据应按照分等级、可发现、可访问、可重用的原则适时向院内外用户提供开放共享。用户使用科学数据时可在确保用户权益的基础上，通过协议的方式开展科学数据的收集和保存等工作。科学院对科学数据中心建设运行进行评估，依据评估结果给予相应支持。同时该管理办法还对数据安全体系建设进行了明确规定。

在高能物理领域，为了促进数据长期保存和开放共享，欧洲核子研究中心（CERN）、美国布鲁克海文国家实验室（BNL）、德国电子同步加速器中心（DESY）、意大利国家核物理研究院（INFN）、法国国家核物理和粒子物理研究所（IN2P3）、日本高能加速器研究机构（KEK）、中国科学院高能物理研究所（IHEP）等数十家国际高能物理研究机构于 2008 年共同发起了 DPHEP（Data Preservation in High Energy Physics）组织。2012 年，DPHEP 发布高能物理数据长期保存蓝图[2]，阐述了高能物理数据长期保存的使用场景、模型、技术及相关建议等。该蓝图还定义了高能物理数据共享的四个等级，包括第一级（Level-1）公开文档或者论文数据、第二级（Level-2）用于教育或科普的简单格式数据、第三级（Level-3）用于完整科学分析的重建数据、（Level-4）第四级所有原始数据及相关条件数据和软件等。2015 年，DPHEP 又发布了扩展蓝皮书[3]，更加详细地描述了高能物理数据长期保存的愿景、案例及项目等。

目前，国际上有多个项目实现对不同级别的高能物理数据的管理和开放。HEPData[4] 在 20 世纪 70 年代建立，收集了粒子物理领域数千种出版刊物上论文中的图表数据，形成独特数据库，实现完全的开放共享。根据 DPHEP 定义，HEPData 实现了 Level-1 级数据的共享开放。CERN Open data[5] 由欧洲核子研究中心于 2014 年发起，用来保存和开放大型强子对撞机 LHC 产生的数据，开放 Level-2 和 Level-3 的数据，以及相关条件数据、软件、文档等，如图 1 所示。CERN Open data 还保存了 LHC 实验以外的实验，比如 OPERA 中微子实验国际合作组的数据。CERN Open data 采用完全开放的协议，并使用数字对象标识符（DOI）来进行数据标注。

图 1 CERN Open data 开放数据集示例

CERN Open data 在 2015 年开始上线运行，2017 年第一篇基于开放数据集的论文发表[6]，该论文作者是非 LHC 国际合作组成员，说明了数据开放的科学价值。

EuDAT[7] 项目于 2011 年 8 月启动，由欧盟第七框架计划和地平线 2020 计划先后资助，其目标是建设一个跨欧洲的协作式科学数据管理基础设施，并针对科学数据的整个生命周期，为科研团体及个人、数据存储管理者提供一整套服务，包括科学数据的同步与交换、存储与共享、安全备份、数据分析、查询及利用。有些高能物理实验，比如 ALEPH 在 EuDAT 上进行长期保存和开放[8]。

2 高能物理及数据

高能物理（也称粒子物理）是研究比原子核更深层次的微观世界中物质的结构、性质，以及在很高能量下这些物质相互转化规律的基础科学，一直处于物质科学的最前沿。高能物理一般于对撞机、地面及地下宇宙线及中微子观测站以及空间卫星等重大科技基础设施开展研究。高能物理科学数据的产生、获取、保存、共享和分析研究贯穿于科学研究的全过程，这些数据是高能物理领域乃至国家的重要战略资源。

2.1 高能物理数据

产生高能物理科学数据的大科学装置包括用于高能物理学科研究的大型专用实验装置以及由高能物理实验技术衍生出来的服务于多学科交叉前沿研究的大科学装置。

我国高能物理领域专用实验装置一般是由多家机构甚至多国参与的大型国际合作，密切面向高能物理领域一个或者多个科学问题开展科学实验。其中北京正负电子对撞机、大亚湾中微子实验装置、高海拔宇宙线观测站为专用实验装置。而服务于多学

科交叉前沿研究的大科学装置则依托高能物理实验技术设计和建设，面向包括材料科学、生命科学等在内的多学科研究提供公共实验技术平台，如上海同步辐射光源、中国散裂中子源、上海自由电子激光装置以及北京高能光源等。由于装置规模、服务对象及科学目标具有较大差别，难以使用统一的分类方法来涵盖高能物理领域所有科学数据的构成。

2.2　专用类科学数据

我国高能物理领域中，大型专用实验装置一般有多家机构甚至多国参与的大型国际合作，此类高能物理科学数据可以从数据来源、数据产生过程、数据承载内容和属性以及实验结果等多个维度进行分类和描述。

按照数据产生来源，高能物理领域科学数据包括实验数据和模拟数据。实验数据包括采集于实验装置及后续处理产生的数据。这些数据的属性由实验或者大科学装置自身情况决定，如我国主导的实验，其建设和运行由中国占据支配地位，则由中国拥有实验数据的所有权，实验依托单位拥有对所有实验数据的完整拷贝，并负责其日常维护。同时，参与实验的国际合作组依据协议拥有数据的使用权。模拟数据指基于物理模型通过计算机模拟计算得到的数据。模拟数据在高能物理实验设计优化、软件开发测试、物理模型验证等很多方面被频繁使用，对实验物理学家同样具有重要意义。以合作组名义统一发布的大规模模拟数据，其数据产生所需的大量计算资源如果由中国提供，则其所有权和使用权同实验数据一致。个别合组作成员或课题组为特殊科研目的产生的小规模模拟数据，一般由产生者自主管理和支配。

根据数据产生的不同阶段划分，高能物理数据包括原始数据、重建数据、分析数据等。原始数据指实验运行过程中从探测器直接获取的数据。原始数据直接依赖于实验设施及其运行，其产生代价高昂，且是所有后续科研活动的起点。因此，原始数据的安全性尤为重要。从数据中心的维护角度看，原始数据需要纳入防灾备份等措施的整体考量；从数据访问角度来说，原始数据只向用户开放读权限，杜绝任何形式的篡改行为。重建数据指经过了标准算法处理后，可供物理学家直接进行科研分析的数据。用于重建的标准算法会不断演进优化，重建数据会存在多个版本，具体管理策略由合作组视需求决定。由合作组发布的重建数据，普通用户也只具有读权限。分析数据则指基于用户个体智力活动，经由重建数据进一步处理后得到的尚未正式发表的数据。分析数据更多体现了个体贡献，在研究进展阶段由研究主体支配。

按照数据承载内容及属性，高能物理科学数据可分为物理数据和条件数据。物理数据泛指直接包含物理反应过程信息的数据，其权属参照前文内容。条件数据则指实验装置参数及其运行状态、软件环境等数据处理过程中必要的所有其他数据。条件数据类目繁杂多样，权属规则不尽相同，具体内容由合作组内部约定。

2.3　公共平台类实验数据

基于大科学装置等条件建立的面向多学科前沿研究的公共服务平台是国家级实验设施，这类平台面向各类研究领域向国内外科学家用户提供实验手段开展多学科前沿研究。公共平台实验产生的数据有些可以直接用于科学研究，有些数据需要经公共平台提供的计算资源进行处理后才能提供给用户使用。由于用户使用了平台的实验资源，这两类数据的所有权应由公共平台和用户共同拥有。

公共平台类实验数据同样可以分成原始数据和用户数据。原始数据是利用实验样品进行实验产生的数据，包含采集的实验数据、实验装置控制数据及监视数据。这两类数据都是科学研究不可缺少的数据资源。

过去相当一段时间以来，国内外公共平台类设施在科学数据的保存及共享中缺乏相应的技术手段和驱动力，如实验数据格式、原数据管理技术等没有形成标准和规范，科学数据的协同处理和分析难度较大。随着合作的加强以及公共平台设施的快速发展，国际上已经在开展相应的技术准备和合作，我国多个此类设施也在积极推动科学数据格式标准化、数据管理规范化研究以及科学数据处理软件框架的合作。

除了实验直接产生的科学数据之外，在大科学装置建设、运行和研究过程中产生的大量文档、成果和专利等内容也是科学数据的重要组成部分。其中仅限合作组内部交流使用的技术文档等数据由合作组自主负责。利用实验数据获得的研究成果等内容，其公开发表过程必须符合合作组规章要求，并以合作组整体名义发表，合作组内部承认该项成果的研究主体为主要贡献者。

3　高能物理科学数据共享

3.1　数据共享模式

高能物理科学数据依托高能物理实验和大科学

装置产生，在高能物理实验周期内产生的所有实验数据，包括原始数据和模拟数据等都是实验的重要资产，需要保证数据长期安全可靠的保存和使用以及基于数据利用规范的共享。

针对依托以我国为主开展的高能物理实验，如果其实验设施是由我国投资并由我国科学家主导建设的专用大科学装置，数据主权属于大科学装置承担机构，原始数据只能在高能物理科学数据中心进行存储和处理。高能物理实验数据对合作组内部完全开放，非合作组成员无法访问。高能物理实验管理采用合作组模式，合作组由国内外科学家共同参与组成，所有签订协议并参与合作组的国内外单位具有实验数据的访问权，并利用数据开展科学研究。

面向多学科交叉研究的公共服务平台类大科学装置科学数据管理中，我国目前还没有统一的数据主权相关规定，但国外的通常做法则是该类科学数据由实验用户及平台共同所有，在数据保护期内，实验用户拥有该类数据使用权，公共服务平台有义务对数据安全性进行管理，数据保护期之后，公共服务平台则有权使用和开放该部分科学数据。

针对我国科学家参与的国际高能物理实验，如果实验数据产生于国外的重大科技基础设施，则参照高能物理合作组制度，在合作组之内的我国科研人员具有科学数据的使用权。

高能物理成果数据作为物理成果文章发表时，需要通过合作组内部相关委员会审核，并在合作组学术研讨会上报告，由合作组指定的相关审稿人审阅后，才可投稿发布。物理成果属于整个合作组，即实验组所有成员都会在文章署名。

到目前为止，我国的高能物理实验数据对公众的开放仍然有限。此类数据对公众（特别是用于教育与科学普及）的开放共享需要政策及经费的支持。

3.2 数据共享技术

高能物理数据具有明显的学科特征，包括大科学装置多、数据来源广泛、数据体量大、质量要求高、数据服务周期长、服务用户广等。这对数据共享技术提出巨大的挑战，目前主要的数据共享技术包括数据标识、数据仓库、数字图书馆、软件及运行环境等几个部分。高能物理共享的大致流程如图 2 所示。

3.2.1 数据标识

长久以来，由于缺乏统一的数据标识，导致高能物理数据独立出版和引用非常困难，因此所有开放的高能物理数据集都需要有一个永久的唯一标识

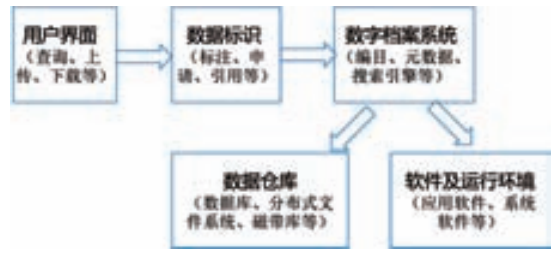

图 2 数据开放共享流程及组件

符，方便数据出版、引用和访问。数字对象标识符（Digital Object Identifier，DOI）、句柄系统（Handles）、统一资源名称（URN）、开放链接（OpenURL）等是目前应用较多的数字标识符，其中尤以 DOI 的应用和研究最为广泛[9]。DOI 是用于识别数字环境下对象的知识产权的字符串，是国际标准化组织（International Standard Organization，ISO）“信息与文献”领域的一项标准，广泛应用于数字化图书、期刊、数据等内容类型的学术出版。DOI 具有唯一且永久标识、永久定位、出版源头注册、点击即链接、动态更新、版本更迭等特点，在科学数据出版中具有跟踪、引用、集成、关联等的多重价值，因此便于实现对数据出版物的原文获取、引文链接、数字版权管理及永久标识等功能，解决数据多重链接和知识产权问题。

CERN Open Data 采用了数据对象标识符 DOI 进行数据标识，并遵循 FORCE11（The Future of Research Communications and e-Scholarship）的“数据引用原则联合声明”（JOINT DECLARATION OF DATA CITATION PRINCIPLES—FINAL）[10] 进行数据引用。为了方便数据引用，CERN Open Data 还为每个数据集提供了推荐引用格式（比如，BibTex 等）。论文引用的数据允许第三方平台（比如 INSPIRE）来跟踪这些数据集的引用情况，并评估其影响因子。

EUDAT 的数据标识采用句柄系统，称为 EUDat B2Handle。B2Handle 是一个分布式的服务，为数据提供全局透明唯一的永久标识符（Persistent Identifier，PID）。PID 用来可靠标识、发布、保存和引用数据对象，也可以用来访问数据。B2Handle 服务包含了标识符空间管理、策略和业务流、句柄系统运行以及用户友好的开发库等。

3.2.2 数字档案系统

数据的开放与共享需要数字档案系统（Digital Library）支持，实现数据存储、分类、发现、引用、访问以及归档等需求。Invenio[11] 是 CERN 开发的一

套开源的综合数字档案系统软件，提供了一系列的框架和工具，能够很方便的构建按需定制的数字档案系统。Invenio 软件提供的技术涵盖了数字档案管理的各个方面，它遵循 OAI-PMH（Open Archives Initiative metadata harvesting protocol）即“开放文献元数据收割协议”协议，采用 MARC21 作为数据编目标准。Invenio 能够管理各类文章、书籍、论文、照片、视频等多媒体的档案信息。由于其灵活性和性能，Invenio 为中大型数字档案系统提供了完整的解决方案，广泛应用于高能物理领域，包括 CERN Document Server、INSPIRE、ILC Document Server、Indico、CERN Open Data、EUDATA、Zenodo 等。

INSPIER[12] 也称为 INSPIER-HEP，是全球高能物理领域的文档、文献及数据等信息的聚合地，能够一站式查询相关信息。INSPIER 基于 Invenio 构建，由欧洲核子研究中心（CERN）、德国电子同步加速器中心（DESY）、美国费米国家加速器实验室（Fermilab）、美国斯坦福直线加速器中心（SLAC）、中国科学院高能物理研究所（IHEP）联合运行，并与多家文献和数据机构紧密合作，包括 arXiv. org，NASA-ADS，PDG，HEPDATA 等。

Zenodo[13] 是一个多学科研究成果储存库，支持各种内容，包括刊物、演示文稿、论文集、项目、图像、软件（包括与 GitHub 的集成）以及所有语言的数据，由欧洲核子研究中心 CERN 维护。它对数据格式没有任何限制，最多可以存储 50 GB 的数据。此储存库中数据可终身保存，可以采用封闭（只要未授权都无法访问）、开放或禁止（禁止期内无法访问）状态储存。Zenodo 受到欧盟第七框架计划和地平线 2020 OpenAIRE 等项目的资助。

EUDAT 项目数字档案系统 B2Share[14] 为研究人员、科学领域及个人提供界面友好可靠以及安全的管理方式，具有存储、长期保存及分享数据等功能。B2Share 不仅仅是一套管理系统，同时包含存储等基础设施，能够以“云”的方式提供服务。

3.2.3　*数据仓库*

数字档案系统中的数据最终要保存到数据仓库中，特别是 Level2 以上级别的数据，容量非常大，达到 PB 级以上，同时还要求长期保存数据。数据仓库技术包括数据库、分布式文件系统、磁带库管理等。在数字档案管理系统中使用了关系型数据库及文档型数据库等。小型数字档案系统直接把文件存储在本地文件系统中，但是对于像 CERN Open Data 这类大型的系统需要将数据存储在分布式文件系统上，甚至磁带存储中。CERN Open Data 使用 EOS 和 CASTOR 系统。EOS 是 CERN 开发的新型 EB 级存储系统[15]，基于 Xrootd 协议进行数据访问，满足 LHC 数据存储及分析的需求，目前已经在 CERN 管理了 200PB 以上的 LHC 实验数据。需要长期保存的数据将从磁盘备份到磁带库中。磁带存储具有容量大、保存时间长、成本低、安全可靠等特点，在高能物理领域内广泛使用。CERN 开发了开源的软件 CASTOR 来管理磁带库及磁带存储[16]。目前，在 CERN 的 CASTOR 系统已经存储了超过 300PB 的数据。分布式文件系统和磁带存储系统之间通过管理工具来进行备份和数据迁移。

图 3 描述了分布式文件系统与磁带库存储之间的关系。对于经常变化的关键数据，比如用户个人数据、数据库文件等依据管理员定义的规则通过备份软件，进行定期的增量备份和全备份。备份软件通过扫描分布式文件系统的元数据获得需要备份的文件列表，然后通过磁带存储系统的接口 API，将数据写入带库。关键数据丢失时，可以通过备份软件将数据从磁带库中找回。对于大量的实验数据，则通过一个专用的数据传输工具在磁带库和磁盘文件系统之间迁移。迁移策略服务器通过扫描分布式文件系统的 CHANGELOG 或者与分布式文件系统元数据交互，根据迁移策略，将需要迁入/迁出的文件写入队列。数据传输工具从迁移策略服务器提供的迁移队列中查找迁入和迁出任务，完成数据移动后，向文件系统注册数据的位置。

EUDAT 使用 CDI（Collaborative Data Infrastructure）来保存、查询和访问数据。CDI 定义了一系列数据模型和技术标准规范。来自欧洲 14 个国家的 20 多个研究机构和科学数据中心通过这些标准规范建立了分布式数据基础设施，包括数据仓库。数据上传到 EUDAT 以后，将通过数据传输工具保存到不同的数据中心里。这些数据中心是服

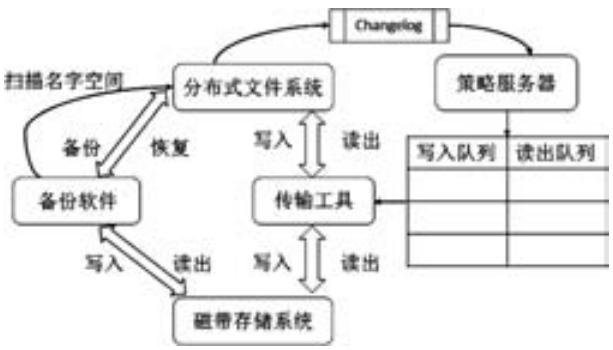

图 3　数据迁移系统

务提供者(Service Provider:SP),具有海量的磁盘存储和磁带存储系统。目前,CDI 是欧洲最大的综合数据服务系统之一。随着欧洲开放科学云(European Open Science Cloud: EOSC)[17]的实现,EUDAT CDI 将成为 EOSC 的主要资源提供者。EOSC 旨在为科学数据的存储、管理、分析和再利用提供开放和无缝的服务。

3.2.4 软件及运行环境

高能物理数据需要特定的软件才能读取和分析,因此数据开放共享必须要提供软件和运行环境才有意义。然而,高能物理软件往往非常复杂,包括应用软件、大量的公共库文件,还有编译器、操作系统依赖等。比如,分析 ATLAS 实验的数据需要 Athena 软件以及依赖的 ROOT 库、GEANT 库、Gaudi、OPENMP、GCC 等,同时不同的数据可能需要不同的软件版本,所以用户自己从头安装配置软件和运行环境是不太可能的。为了方便用户使用,在开放数据的同时,还将该数据依赖的软件和运行环境通过虚拟机的形式发布。目前,高能物理领域广泛使用 CernVM[18]。CernVM 将操作系统、系统软件、应用库和应用软件打包在一起,形成一个虚拟的应用程序,并做了针对性的优化,一个虚拟机所占的空间最小的只有 100MB。CernVM 是一个完整、便携和易于配置的用户环境,可以在本地或者商业云(Openstack、Amazon EC2、Google Compute Engine)上开发和运行高能物理数据分析,并且独立于操作系统和硬件平台(如 Linux、Windows、MacOS 等)。此外,CERN Open Data 等项目还将数据、虚拟机和云计算连接起来,当用户访问数据时还可以直接启动相应的运行环境,然后在云端上完成数据读取、分析等工作,不用下载到本地处理。

3 数据安全

科学数据是高能物理科学研究的重要资产,数据安全非常重要,需要保证数据长期安全可靠的保存、使用以及杜绝非法访问。高能物理科学数据安全包含两方面的含义,即数据访问安全和数据存储的安全。数据访问安全主要是指采用技术手段对实验产生的原始数据和用户数据进行主动保护,如数据加密、权限控制、双向强身份认证等;数据存储安全主要是采用现代信息存储手段防止实验数据意外丢失,如通过分级存储、异地容灾等手段保证数据的存储安全。

我国高能物理科学数据安全面临的主要风险是缺乏相应的资源、人力和软硬件基础设施投入,导致在数据安全与开放共享过程中缺乏人力和经费支持。虽然我国于 2018 年发布了科学数据管理办法,明确了科学数据管理和共享的原则,但受制于种种因素,当前面向多学科交叉研究服务的重大科技基础设施平台产出的科学数据管理中,仍然缺乏可操作的科学数据安全和数据主权相关的法律法规以及管理制度支持,需要在设施建设、运行和服务过程中涉及到的多方主体进一步讨论、细化相关管理内容。

4 讨论与建议

2018 年 4 月欧盟发布了关于科学信息的访问与保存的 2018 第 790 号建议书(COMMISSION RECOMMENDATION (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information)[19]。尽管建议书对成员国没有法律约束力,但将促进各成员国立法。该建议书对科学数据的开放共享提出了具体要求。建议书要求成员国应以国家行动计划的方式就公共经费支持的研究所获得的科学数据建立和实施明确的管理及开放共享政策。政策和行动计划的目标是建立数据管理计划,保证科学数据的可发现、可访问、可互操作和可再用。数据要做到尽可能的开放,除非必要才不开放。成员国须向学术机构提供公共经费用于支持数据管理与开放共享。

2018 年国务院发布的《科学数据管理办法》是科学数据开放共享的纲领性文件,为我国科学数据的开放共享提供了法律依据。尽管相关科研及教育机构依照国务院的管理办法制定了相应的科学数据管理实施办法,但是在具体操作及条件保障方面仍有许多需要细化和落实到地方。为了促进和保障科学数据的开放共享及其可持续的发展,应从以下几个方面落实相应的措施。

(1) 推动科学数据效能最大化。科学数据开放共享的目的是使数据的效能最大化。科学数据绝大部分来自国家财政支持的科学装置和科学研究项目。根据国务院科学数据管理办法的精神,国家和科学家对数据共同拥有所有权。对于参与外方为主的大科学国际合作,应在遵循我国科学数据管理办法的前提下,建立数据使用与开放共享协议,确保中国及中国科学家在数据上的权益。为保障参与数据生产的科学家在数据使用上的优先权,可以建立数据保护期,期限 2—3 年,让科学家能在第一时间利用数据开展科学研究。对急需使用实验数据的外部

用户，可以与数据拥有者签订合作共享协议，让用户在保护期内也能及时利用数据进行科学研究。

（2）落实“尽量共享、不得已才受限”的数据开放共享原则。就国家层面而言，数据的开放共享应遵循“尽量共享、不得已才受限”的原则。除可能影响国家安全或技术保密等不可回避的原因外不得以不合理的借口阻碍数据的开放。对确实需要不公开的数据，应由专门的机构组成专家组对理由进行充分的评估，确定数据开放共享的程度与方式。

（3）加强对数据开放共享技术和软件环境的支持。为了确保数据的开放共享可持续，需要在技术上提供强有力的支撑。首先必须要建立可靠的数据保存系统，确保在数据的生命周期内甚至永久期限内能否被发现、被访问和被利用。仅做到这些还不够，还需要为科学数据建立完善的解释文档，让数据用户了解数据的结构、性质以及使用方法，确保用户高效正确的使用数据。科学数据开放共享需要技术保障，需要建立和完善数据保存、数据访问及数据加工处理的软件环境。软件环境包括操作系统、数据管理及传输、用于科学研究及教育与科普的软件工具等等。另外，为保障科学数据的正确性和有效性，需建立数据审计验证系统对数据进行验证，确保数据在保存及迁移的过程中不被损坏或篡改。

（4）加强对数据人才队伍的培养和建设。现代科学数据的规模和复杂度都是空前的。数据的使用，包括数据分析的过程非常复杂，任务量巨大，需要一批专业数据专家和计算机专家对于数据分析数据分析的算法和软件进行开发，为相关科学领域的科学家以及数据用户提供服务。大部分领域科学家，特别是青年科学家及学生在计算机技术方面的训练不够，对科学数据分析处理工具、软件编程等不够熟悉，因此数据科学家及计算机专家就显得尤为重要。国家应建立相应的政策，鼓励数据科学家及计算机科学家参与科学数据工作，同时鼓励领域科学家参加计算机技术的培训，提升科学家计算机水平和数据分析的水平，促进科学产出。建立合理的机制对数据科学家和计算机专家的工作给予肯定，并在职业晋升、待遇等方面给予足够保障，吸引高水平的人才稳定从事科学数据工作，包括计算机软件开发，运行维护等工作。为了对数据的拥有者的工作给予肯定，数据用户利用科学数据取得并发布成果时应该在论文中对数据科学家和计算机专家的工作进行致谢。这对有效评价和认可数据科学家的贡献尤为重要。

（5）加强对数据管理的经费支持。数据保存、加工、管理及开放工作需要巨大的成本，这是不可回避的问题。没有稳定充足的经费支持，科学数据开放共享工作将成为空中楼阁。数据服务系统的稳定运行是共享开放的基本保障。国家应该建立完善的经费支持保障机制。经费支持机制的重要任务之一是对数据开放共享绩效的评价，应建立合理客观的评价体系，对科学数据开放共享的绩效进行评价，用于经费支持的决策。

科学数据是国家财富，是科学发现的基础。开放共享有利于科学数据的效能最大化，国务院和科学院分发布的《科学数据管理办法》及《中国科学院科学数据管理与开放共享办法（试行）》恰逢其时，是我国科学数据开放共享的基础。我国及科学院科学数据的开放共享工作才刚刚开始，还有很长的路要走。科研机构、科学家与政府部门应通力合作，尽快完善科学数据开放共享环境，为科技创新作出共享。

参　考　文　献

[1] 中国政府网，2018-04-02. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.

[2] Akopov Z, Amerio S, Asner D, et al. Status report of the DPHEP Study Group: Towards a global effort for sustainable data preservation in high energy physics. arXiv preprint arXiv:1205.4667, 2012.

[3] Amerio S, Barbera R, Berghaus F, et al. Status Report of the DPHEP Collaboration: A Global Effort for Sustainable Data Preservation in High Energy Physics. arXiv preprint arXiv:1512.02019, 2015.

[4] Maguire E, Heinrich L, Watt G. HEPData: a repository for high energy physics data//Journal of Physics: Conference Series. IOP Publishing, 2017, 898(10): 102006.

[5] CERN Open Data Portal, http://opendata.cern.ch/.

[6] Tripathee A, Xue W, Larkoski A, et al. Jet substructure studies with CMS open data. Physical Review D, 2017, 96(7): 074003.

[7] Reetz J. EUDAT-Open Data Services for Research//Science Operations 2015: Science Data Management-An ESO/ESA Workshop, held 24—27 November, 2015 at ESO Garching.

[8] Berghaus F. The case for preserving our knowledge and data in physics experiments. arXiv preprint arXiv: 1712.01276, 2017.

[9] Paskin N. Digital object identifiers for scientific data. Data science journal, 2005, 4: 12—20.

[10] Wimalaratne S M, Juty N, Kunze J, et al. Uniform resolution of compact identifiers for biomedical data. Scientific data, 2018, 5: 180029.

[11] Caffaro J, Kaplun S. Invenio: A modern digital library for grey literature. 2010.

[12] INSPIER-HEP Portal: http://inspirehep.net/.

[13] ZENODO Portal: https://zenodo.org/.

[14] Ardestani S B, Hakansson C J, Laure E, et al. B2share: An open escience data sharing platform//2015 IEEE 11th International Conference on e-Science (e-Science). IEEE, 2015: 448—453.

[15] Peters A J, Janyst L. Exabyte scale storage at CERN. Journal of Physics: Conference Series. IOP Publishing, 2011, 331(5): 052015.

[16] Presti G L, Barring O, Earl A, et al. CASTOR: A Distributed Storage Resource Facility for High Performance Data Processing at CERN. MSST. 2007, 7: 275—280.

[17] Giannoutakis K M, Tzovaras D. The European Strategy in Research Infrastructures and Open Science Cloud. International Conference on Data Analytics and Management in Data Intensive Domains. Springer, Cham, 2016: 207—221.

[18] Buncic P, Sanchez C A, Blomer J, et al. CernVM—a virtual software appliance for LHC applications. Journal of Physics: Conference Series. IOP Publishing, 2010, 219(4): 042003.

[19] Commission Recommendation (EU) 2018/790 of 25 April 2018 on access to and preservation of scientific information, C/2018/2375. http://data.europa.eu/eli/reco/2018/790/oj.

Platform of Science Data Open Access and Sharing: Science data management and sharing for High Energy Physics

Qi Fazhi Chen Gang Cheng Yaodong

(*Institute of High Energy Physics, Chinese Academy of Sciences, Beijing* 100049)

Abstract High energy physics (also known as particle physics) research has always been at the forefront of fundamental science. In most cases, large scientific facilities are national essential facilities for fundamental physics research and the nation's strategic infrastructure. In recent years, China has increased its investment in the construction and operation of large scientific facilities for front edge physics. These facilities have been becoming important platforms for basic and applied researches. The data generated from large scientific facilities are the important resources for scientific researches. The open access and sharing of scientific data is essential for maximizing the benefits of scientific data. Therefore the open access and sharing of data have attracted worldwide attention. This paper analyzed the data's characteristics of large scientific facilities for high energy physics, and briefly discussed the data sharing, long-term preservation and reuse. The strategy of data scientists is also discussed. The purpose of the paper is to provide example of data sharing of large scientific facilities in China.

Key words large facilities; data management; data sharing; data preservation; data ownership; data identification