doi:10.3772/j.issn.2095-915x.2017.04.009

科学数据管理应用模式的研究

中国科学技术信息研究所 北京 100038

张迎 张志平 梁冰

摘 要 随着科学研究进入第四范式，科学研究过程中产生了大量的科学数据，对科学数据的管理越来越受到科研人员的关注。为了实现科技文献和科学数据的关联服务，促进科学数据的开放获取、共享和再利用，在文献管理的基础上，大量研究不断探索科学数据管理的方法。本文在相关理论研究基础上，研究并构建科技文献和科学数据一体化科学数据管理应用模式，根据该模式搭建出科技文献和科学数据一体化管理系统，然后利用高能物理领域的实例数据验证模式和原型系统的可行性。

关键词： 科学数据，科学数据管理，数据管理应用模式，数据管理生命周期

中图分类号： G350

开放科学（资源服务）标识码（OSID）

Research on Scientific Data Management Application Model

Institute of Scientific and Technical Information of China, Beijing 100038, China

ZHANG Ying ZHANG ZhiPing LIANG Bing

Abstract With scientific research entering the fourth paradigm, scientific data has been produced in the process of scientific research, and the management and application of scientific data are paid more attention by researchers. In order to realize the related service of literature and scientific data and promote the open access, sharing and reuse of these data, researches explored the method of scientific data management on the basis of document management constantly. On the basis of relevant theoretical research, this paper studied and constructed the scientific data management model of scientific and technological literature and science data integration, built scientific and technical literature and scientific data integration management system according to this model. Moreover, this study also used the practical data validation model and applied the system in the field of high energy physics.

Keywords: Scientific data, scientific data management, data management application model, data management lifecycle

作者简介： 张迎（1990–），女，硕士研究生，中国科学技术信息研究所，主要研究方向：信息检索与数据库建设；张志平（1963），男，中国科学技术信息研究所，信息技术支持中心，研究员；梁冰（1974–），男，中国科学技术信息研究所信息技术支持中心主任，高级工程师。

doi:10.3772/j.issn.2095-915x.2017.04.009

1 引言

科学研究是人类探求科学真理、认识世界的过程，是推动人类进步的关键。随着人类科学研究进入第四范式[1]，数据呈现爆炸式增长，科学研究模式转变为由仪器收集或仿真计算产生数据，然后由软件处理数据，再由计算机存储信息和知识，科学家通过数据管理和统计的方法分析数据和文档。科学家正在积极探索将数据集整合进学术信息交流中，科学研究第四范式的目标是让人们拥有一个科技文献和科学数据都在线且能够彼此交互操作的世界。

目前在对科技文献和科学数据的管理上倾向于单独管理，然而科技文献和科学数据之间的管理是客观存在的。为了满足用户的需求，使文献与科学数据产生关联，拓展科技文献的开放获取服务，实现对科学数据的开发获取、共享和重用，我们需要对科技文献和与之相关的科学数据进行管理，让人们在阅读文献的同时，可以查看他们的原始数据，甚至可以对数据做分析；在查看某些数据时，可以查看关于这一数据的文献。

国家科技图书文献中心（NSTL）是SCOAP3（Sponsoring Consortium for Open Access Publishing in Particle Physics，粒子物理学开放获取出版赞助联盟）中国集团牵头机构，2013 年代表中国加入了 SCOAP3，实质性参与 SCOAP3 的管理和运作。SCOAP3 致力于将高能物理领域的学术论文转为开放出版，科研机构和科研人员希望在阅读高能物理文献的同时，可以查看与之相关的科学数据；在查看科学数据时，可以查看关于这一数据的文献，同时对科学数据进行再利用。本文在对科学数据管理进行理论研究的基础上，以高能物理领域科技文献和科学数据为例，构建一个科技文献和科学数据能够彼此交互操作的科学数据管理应用模式，为下一步科学数据管理应用系统的构建做好铺垫。

2 科学数据

2.1 科学数据定义

OECD(Organization forEconomicCo-operation and Development，世界经济合作与发展组织）在《OECD 关于公共资助科学数据获取的原则和方针》[2] 中提出：科学数据是指从科学研究中事实记录中得到的，科研团体或者科研工作者均认为对研究结果有用的数据，例如实验数值、图像等，目前这种定义方式最受业界认可。科学数据的表现形式有：文档、数据文件、问卷、模型、算法、软件或代码、图片、音视频资料等。

科学数据具有以下特点[3]：(1) 数据量巨大；(2) 连续性：每年都会产生大量新的数据；(3) 数据类型多样；(4) 共享性：数据被不同用户使用，并在出版物中被引用；(5) 数据交换性：和领域其他数据平台进行数据交互收割；(6) 与数据支撑的科技文献之间有所关联。

2.2 科学数据的界定

科学数据具有不同的结构层次，相应的管理手段也不同。

在高能物理领域，科研人员按照金字塔的形状将科学数据分为四个层次[4]，从下到上依次为：①第 4 级数据：位于金字塔的底部，是实验的原始数据，他们来自探测器。②第 3 级数据：第 4 级数据重建及仿真后得到的数据、以及对第4级数据进行科学分析所用到的软件。

③第 2 级数据：针对第 3 级数据，通过分析代码来简化，结果变成了简化数据格式后的第 2 级数据。在进一步的处理步骤中，第 2 级数据是准备进行发布的数据集。④第 1 级数据：在论文中出现的图片、表格数据，以及支撑论文中图表数据的原数据。

在这里，我们对与科技文献相关联的科学数据进行管理，具体包括以下三种形式：①出现在文献中的图表数据；②支持文献中图表数据的与文献相关的辅助数据，我们将辅助文献的数据称为数据集，一个数据集包含多个数据文件；③文献中的文字提及到的数据，即文献中引用到的数据集，在文献里一般在参考文献中以链接的形式给出，通过链接可以找到文献中引用的数据集。

3 科学数据管理

3.1 科学数据管理内涵

科学数据管理，国内外对其定义略有不同。在国外，科学数据管理被称为 Science Data Curation[5]，美国伊利诺伊大学图书馆与信息科学研究生院将其定义为一种主动的、持续的，随着对学术、科学和教育的兴趣及用途的生命周期而进行的数据管理；日本工业标准委员将其定义为从数据产生开始，对其进行管理和完善，确保数据集的持续性补充和更新，以实现数据的随时使用、再发现、再利用。2007 年，微软首席研究员、计算机图灵奖获得者 Jim Gray 提出，数据管理活动涵盖了制定标准、元数据创建、数据映射到不同仓储、语义注释、文献链接等广泛的活动 [6]。在国内，科学数据管理在国内的含义也是多种多样，有的学者将其定义为科学数据策管 [7]，有的将其定义为科学数据监护 [8]，有的将其定义为科学数据监管 [9]，还有的将其定义为数字保存和数据掌管 [10]。

综合国内外可知，科学数据管理是利用计算机硬件和软件技术对科学数据进行有效的收集、存储、处理和应用的过程。

3.2 科学数据管理生命周期

数据管理是对数据的整个生命周期，从数据被创建到存储管理、复用的循环过程进行研究。科学数据管理生命周期核心要素包括：数据管理计划、数据收集、数据处理、数据分析、数据保存、数据共享和数据再利用等 [11]。

为了帮助科研工作者更好地管理科学数据，不同研究机构的学者给出了不同的数据生命周期模型，最为典型的是英国数据监护中心 (Digital Curation Center，DCC) 创建的 DCC 数据监护生命周期模型（DCC Curation Lifecycle Modle）[12]，将科学数据管理生命周期按照顺序分为八个阶段：概念化、创建或接收数据、评估和选择、摄取、长期保存、存储、访问和再利用、转化。

本文的科学数据管理应用系统的管理流程是从数据获取到数据重用，为数据管理生命周期的一部分，具体分为五个阶段：①获取科学数据；②描述科学数据，生成元数据；③存储科学数据和元数据；④发布科学数据；⑤重用科学数据。由于本文搭建的科学数据管理系统是针对科技文献和科学数据两种数字资源的，因此，本文的科学数据管理包括科技文献和科学数据的获取、描述、保存、发布和重用。

3.3 OAIS参考模型

OAIS（Open Archival Information System，开放档案信息系统）是一种在构建开放性数字信息

doi:10.3772/j.issn.2095-915x.2017.04.009

系统时所遵循的基本框架，遵循的标准由美国空间数据系统咨询委员会（CCSDS）制定，于 2003 年作为国际标准组织（ISO）颁发[13]。最近几年，OAIS 模型被广泛采用，例如很多图书馆和机构知识库正在更新他们的系统以达到 OAIS 兼容。

因此本文的科学数据管理应用模式参照 OAIS 标准模型进行构建，将科学数据管理分为 SIP（Submission Information Package）、AIP（Archive Information Package）、DIP（Dissemination Information Package）三个阶段[14]。SIP 即提交信息包，是生产者传送到本系统的信息包；AIP 即存档信息包，是信息被接收到系统后，经过信息描述等数据管理相关流程，在系统存储后的信息包；DIP 即分发信息包，是将信息发布给用户的信息包。

4 科学数据管理应用模式构建

由于 NSTL 加入了 SCOAP3，本文是以 SCOAP3 中科技文献为研究对象产生的后续成果，主要是为了拓展高能物理领域科技文献的开放获取服务，实现与科技文献相关联的科学数据的开放获取、共享和重用。因此，本文在科学数据管理应用模式中，以高能物理领域的科技文献和科学数据为例，来进行元数据方案设计、数据和文献关联和数据组织，并以此数据来进行实例验证。

4.1 科学数据管理应用模式

科学数据管理应用模式结合数据管理生命周期，参照 OAIS 参考模型进行构建，如下图 1 所示。

第一是 SIP 阶段，也就是数据获取阶段。首先系统获取到科技文献的原文，以及与文献相关的科学数据，然后生成文献元数据和科学数据的元数据，元数据验证成功后，科技文献、科学数据和元数据一起被摄入科学数据管理应用系统。第二是 AIP 阶段，在此阶段科技文献、科学数据和文献元数据、科学数据元数据会被一起存储到科学数据管理应用系统中，系统管理员可以对文献元数据和科学数据的元数据进行修改，进而更新元数据；然后索引元数据，实现文献和科学数据的检索与浏览。系统可以通过持久标识符来实现数据的关联。最后是 DIP 阶段。用户可以访问并获取到已发布的数据，并对数据进行重新使用，重用包括对其进行浏览、下载和数据引用等。

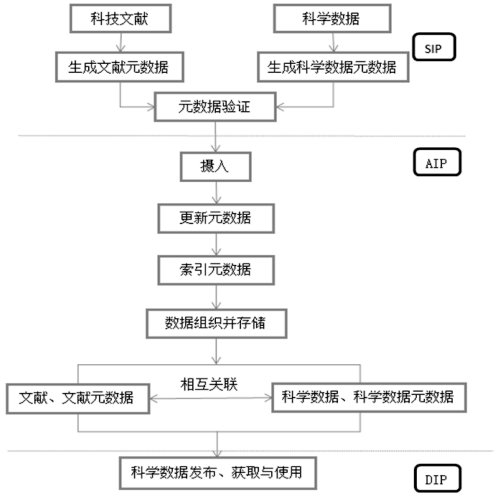

图1 科学数据管理应用模式图

4.2 科学数据管理应用模式解析

科学数据管理应用模式中包括元数据管理、数据和文献关联以及数据组织方式。根据高能物理领域科技文献和科学数据特点，列出元数据定义表、数据和文献关联元数据表，并根据数据的特点进行了数据组织。

(1) 元数据管理

元数据通常解释为“关于数据的数据”，是对数据进行组织和处理的基础。元数据被定义为关于数据的一个或多个方面的数据提供信息，

它用于总结关于数据的基本信息，可以方便跟踪和使用特定数据。

元数据方案设计是构建科学数据管理应用模式需要考虑的核心问题之一。本文在描述科技文献时基于 DC 元数据标准，包括标题、作者、出版年份、文献类型等元数据描述元素；在描述科学数据时基于 DataCite 元数据方案，针对高能物理领域科学数据的特殊性进行有效扩展，定义一些和基本元素不重复的新元素，形成自有的元数据规范。

科技文献的元数据方案设计分为必备项和可选项两部分，具体形式如下表所示。

(2) 数据和文献关联

科学数据和科技文献的关联研究目前成为学术界关注的热点，二者的关联有助于在当前的信息社会实现信息服务，促进知识发现并完善 E-science 环境，从而促进科学数据的发现、重新利用和引用。

表1 科技文献元数据定义表

元数据名称	英文名称	说明	约束
标题	Title	文献的标题	必备
作者	Authors	文献的作者	必备
合作单位	Collaboration	文献作者所在单位	必备
DOI	DOI	文献的持久唯一标识符	必备
主要分类	primaryClass	文献分类信息	必备
摘要	Abstract	文献的简要描述信息	必备
关键词	Keywords	文献关键词	必备
文献出处	Literature Sources	有关文献出处的信息	可选

表2 数据集元数据定义表

元数据名称	英文名称	说明	约束
标题	Title	数据集的标题	必备
作者	Authors	数据集的作者	必备
合作单位	Collaboration	数据集所在单位	必备
数据集描述	Dataset description	对数据集的详细描述	必备
数据集 DOI	DOI	数据集的持久唯一标识符	必备

表3 数据文件元数据定义表

元数据名称	英文名称	说明	约束
标题	Title	数据文件的标题	必备
位置	Location	数据文件对应文献中数据的位置	必备
数据文件注释	Datafile comment	对数据文件的详细描述	必备
反应关键词	reaction keywords	物理反应中的关键词	必备
可观察到的关键词	observable keywords:	物理反应中可观察到的实验关键词	可选
数据文件 DOI	DOI	数据文件持久唯一标识符	必备

国内外针对科学数据和科技文献的关联进行了研究：理论方面主要是科学数据和科技文献的关联方式及关联技术，例如涂勇、彭洁 [15] 建立了基于 DOI 技术的科学数据与科技文献融合的模型。实践方面主要是构建了科学数据管理平台，这些平台可以实现科技文献与科学数据的关联，以及科学数据的共享和重用，例如 Figshare 与 Willey 建立合作关系发布期刊数据，促进数据的共享 [16]。

本文基于元数据描述，利用 DOI 技术，实现科学数据与科技文献融合，具体方式为在科技文献元数据中加入数据集 DOI，在数据集元数据中加入科技文献 DOI，如下表所示。

表4 科技文献关联元数据表

元数据名称	英文名称	说明	约束
相关数据集 DOI	Related Dataset DOI	与文献相关联的数据集的唯一标识符	必备
关系类型	relation Type	文献与数据集之间的关系类型	必备

表5 科学数据关联元数据表

元数据名称	英文名称	说明	约束
相关文献 DOI	Related Literature DOI	与数据集相关联的文献的唯一标识符	必备
关系类型	relation Type	数据集与文献之间的关系类型	必备

举例来讲，文献与数据集之间的关系类型可以为文献与支撑文献结论的文献外表格形式

doi:10.3772/j.issn.2095-915x.2017.04.009

的数据集之间的关系。

(3) 数据组织

数据组织是指按照一定的规则和方式对数据进行归并、存储和处理的过程。科学数据组织是科学数据管理和科学数据服务的前提。科学数据组织的基本原理是用一定的知识组织方法把数据客体中的知识元素和知识关联揭示出来，并编排成序，形成易于利用的知识体系结构[17]。

本文在进行数据组织时定义五个实体：成果空间、馆藏、条目、数据流包、数据流。成果空间是一个虚拟的容器实体，可以嵌套，即成果空间下可以包含子成果空间,还可以包含多个馆藏。例如，可以用成果空间表示学科和研究领域，则父成果空间可为“物理学”，子成果空间可为“高能物理”。馆藏也是一个虚拟的容器实体，不可以嵌套，它依存于某个成果空间，由多个条目组成。例如，用馆藏表示期刊，则 *Journal of High Energy Physics* 可以作为一个馆藏，它位于“高能物理”成果空间之下。条目为资源对象（如文献、数据集），它依存于某个馆藏，包含一个或多个数据流。数据流包是一组逻辑上相关的数据流组成的集合，如本文高能物理领域中一个数据文件及其在文章中对应的图形可以组成一个数据流包。数据流即单个数据文件。具体数据组织结构示例如下图所示。

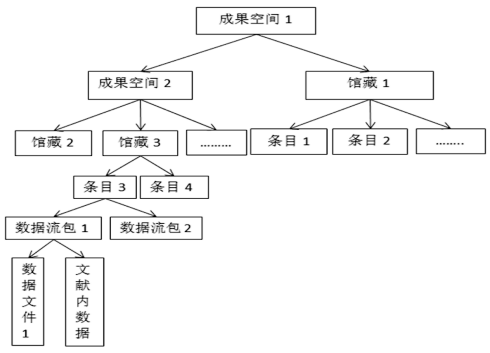

图2 数据组织结构图

4.3 实例验证

根据科学数据管理应用模式，搭建了科学数据管理应用系统，并利用高能物理领域的科技文献和科学数据，验证模式和原型系统的可行性。

科学数据管理应用平台的功能需求包括两部分：用户功能需求和后台管理功能需求。用户功能包括用户对数据和文献的浏览和检索、对数据的引用及下载。后台管理功能包括数据获取、元数据管理、编辑数据、发布数据、数据在线分析和可视化展示。

系统业务流设计图如下图所示。

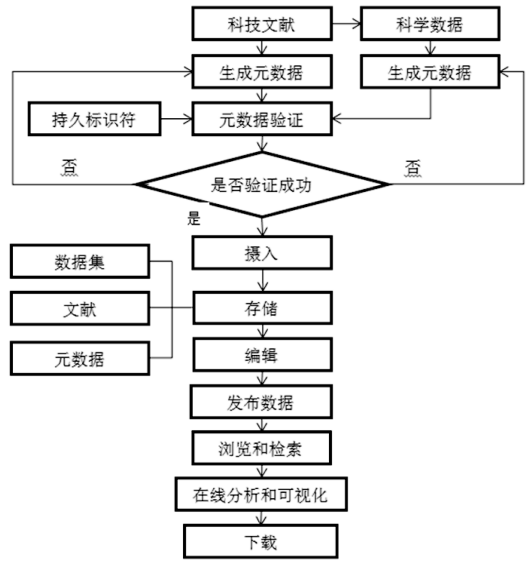

图3 系统业务流程图

从 INSPIRE[18] 中获取到科技文献，从 INSPIRE 关联网站 HEPData[19] 中获取到科学数据，生成科技文献元数据和科学数据元数据，进行元数据验证，验证成功后将元数据、文献和科学数据一起摄入到系统中进行存储。用户可以对数据进行编辑，以完成数据的更新。用户可以浏览和检索科技文献和科学数据，并可以进行下载。用户可以进入科技文献的元数据

界面，找到与文献相关联的科学数据；反之，用户也可进入科学数据的元数据界面，找到相关出版物字段，通过字段中的链接找到与科学数据相关联的科技文献。系统还提供科学数据的在线分析和可视化功能。

5 结语

数字化科研环境下，科学研究产生的科学数据直接影响着科学研究的发展变化。有效并科学地对科学数据进行管理，为科学研究活动提供足够的支持，方便科研工作者共享和再利用科学数据，是目前研究人员关注的焦点，更是当前亟待解决的问题。

本文从科学数据管理角度出发，设计了元数据管理方式、关联方式和数字组织方式，研究并构建了一个科技文献和科学数据一体化管理模式，并根据此模式搭建了科学数据管理应用系统，利用高能领域实验数据进行了实例验证。然而这样的方式具有一定的局限性，适用范围相对较小，有待以后进行更为深入的研究。

参考文献

[1] 邓仲华, 李志芳. 科学研究范式的演化——大数据时代的科学研究第四范式[J]. 情报资料工作, 2013, 34(4): 19-23.

[2] DIRK P. OECD Principles and Guidelines for Access to Research Data from Public Funding[J]. Data Science Journal, 2007, 6: 4-11.

[3] 殷沈琴, 张计龙, 张莹, 等. 社会科学数据管理服务平台系统选型研究——以复旦大学社会科学数据平台为例[J]. 图书情报工作, 2013, 57(19): 92-96.

[4] Herterich P, Tiessen S D. Data Citation Services in the High-Energy Physics Community[J]. D-Lib Magazine, 2016, 22(1/2).

[5] 李文文, 陈雅. 国内外Data Curation研究综述[J]. 情报资料工作, 2013(5): 35-38.

[6] Hey T. The Fourth Paradigm – Data-Intensive Scientific Discovery[M]. E-Science and Information Management. Springer Berlin Heidelberg, 2012.

[7] 于霜, 姚占雷. 数据监管: 图书馆科研数据共享新服务[J]. 中国科技资源导刊, 2013(6): 45-50.

[8] 杨鹤林. 数据监护: 美国高校图书馆的新探索[J]. 大学图书馆学报, 2011, 29(2): 18-21.

[9] 沈婷婷, 卢志国. 数据监管在我国高校图书馆的应用展望[J]. 图书情报工作, 2012, 56(07): 54-87.

[10] 张智雄. 如何长期保存数字资源[J]. 中国教育网络, 2006(4): 26-28.

[11] 杨林, 钱庆, 吴思竹. 科学数据管理生命周期模型比较[J]. 中华医学图书情报杂志, 2016, 25(11): 1-6.

[12] Higgins S. The DCCCuration Lifecycle Model[J]. 2008, 3(1): 453-453.

[13] 吴江华. 开放性档案信息系统:背景、职责及功能[J]. 图书情报知识, 2006(05): 85-87.

[14] Lavoie B F. The Open Archival Information System Reference Model: Introductory Guide[J]. Microform & Imaging Review, 2008, 33(2): 68-81.

[15] 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.

[16] 现代图书情报技术编辑部. Wiley与Figshare合作促进数据共享[J]. 现代图书情报技术, 2016(2): 24-24.

[17] 樊俊豪. 基于科学数据管理的图书馆知识服务实现研究[D]. 上海: 上海大学, 2014.

[18] INSPIRE.INSPIREProject Overview. [EB/OL]. [2016-10-26]. http://inspirehep.net /info/general/project/index.

[19] HEPDATA.ABOUT HEPDATA.[EB/OL]. [2016-10-26]. https://hepdata.net/about.