空间科学数据产品组织模型的应用研究

纪 珍[1,2] 佟继周[1,2] 胡晓彦[1,2] 邹自明[1,2] 马福利[1,2] 熊森林[1,2]
（1.中国科学院国家空间科学中心，北京 100190；2.国家空间科学数据中心，北京 100190）

摘要：科学数据产品规范化组织是科学数据管理过程中的重要环节，对空间科学领域多学科多类型数据资源的统一组织是发挥数据价值，支撑科技创新的有效保障。本文基于国家科学数据中心构建的科学数据产品组织模型，以空间科学先导专项为例，从数据汇交与处理、管理与归档、发布共享3个数据活动关键环节，系统地介绍其在空间科学数据管理活动中的应用，总结科学数据产品组织模型实际应用效果。

关键词：科学数据；科学数据管理；科学数据组织模型；空间科学数据
DOI：10.3772/j.issn.1674−1544.2022.01.010 **CSTR：**15994.14.issn.1674−1544.2022.01.010
中图分类号：P35；P17 **文献标识码：**A

Application of Data Organization Framework of Space Science

JI Zhen[1,2], TONG Jizhou[1,2], HU Xiaoyan[1,2], ZOU Ziming[1,2], MA Fuli[1,2], XIONG Senlin[1,2]
(1.National Space Science Center, CAS, Beijing 100190; 2.National Space Science Data Center, Beijing 100190)

Abstract: The standardized organization of scientific data products is an important step in the process of scientific data management. The unified organization of multi-disciplinary and multi-type data resources is an effective guarantee for data value mining and technology innovation in the field of space science. Based on the scientific data product organization framework constructed by the national space science data center, taking the Strategic Priority Program on space science as an example, this paper systematically introduces the application of the organization framework in the space science data management activities from three key steps of data management activities: data collection and processing, management and archiving, release and sharing, then summarizes the achievements of organization framework application.

Keywords: scientific data, scientific data management, space science data organization framework, space science data

0 引言

空间科学是具有前沿性、拓展性及发展急迫性的交叉学科领域，主要对发生在日地空间、行星际空间乃至整个宇宙空间的物理、天文、化学以及生命等自然现象及规律进行研究，涵盖了空

作者简介：纪珍（1982—），女，中国科学院国家空间科学中心副研究员，研究方向为科学数据管理、空间科学信息系统、空间物理；**佟继周**（1976—），女，中国科学院国家空间科学中心研究员，研究方向为空间科学信息系统、计算机应用技术、数据处理和管理（通信作者）；**胡晓彦**（1987—），女，中国科学院国家空间科学中心副研究员，研究方向为空间科学信息系统、数据组织模型、空间物理；**邹自明**（1971—），男，中国科学院国家空间科学中心研究员，研究方向为空间科学信息学、日地空间大数据处理与应用；**马福利**（1986—），男，中国科学院国家空间科学中心副研究员，研究方向为空间科学信息系统、计算机应用技术、数据处理和管理；**熊森林**（1987—），男，中国科学院国家空间科学中心助理研究员，研究方向为空间科学数据处理、空间物理。

基金项目：中国科学院“十四五”网络安全与信息化专项“中国科学院空间科学数据中心能力建设项目”（WX145XQ07−06）。
收稿时间：2021 年 9 月 30 日。

间物理、空间天文、太阳物理、空间地球科学、微重力科学及空间生命科学等学科领域。在空间科学大数据时代[1-2]，科学研究呈现全球化、链条式及定量化等特点，对科学数据分析挖掘与综合利用的依赖性日益增强。利用天基、地基一体化探测网络所产生的数据资源，综合不同学科数据资源的物理要素，开展空间科学前沿问题的智能化研究，探索数据驱动的研究方法，深入挖掘数据的科学价值，为空间科学数据管理提出了新的挑战。

同时，科学数据是国家科技创新和经济社会发展的重要基础性战略资源，也是大数据时代开发利用潜力最大的科技资源，对其的收集、保存、共享和利用是国家科技投入效益的直接体现。在确立大数据国家战略的大背景下，2018年3月，颁布了首个国家层面的数据管理政策《科学数据管理办法》[3]，在加强科学数据全生命周期管理、保障数据安全、建立共享交流的审查机制、充分发挥科学数据的重要作用等方面，为国家科学数据中心开展科学数据管理与共享工作确定了行动纲领。

在空间科学数据管理全过程中，数据产品规范化组织是数据汇交与管理阶段的重要环节，对多学科多类型的数据进行统一管理与标准组织是其中急需解决的问题之一。国内外学者大多是对空间科学数据的时空组织模式进行研究，而对数据产品组织模型设计的探讨较少。如美国国家航空航天局（National Aeronautics and Space Administration，NASA）联合国家空间物理研究所提出的空间物理档案搜索与提取系统SPASE模型[4]和行星科学数据系统PDS[5-7]模型等数据模型。这些模型为空间物理和行星科学领域的数据组织存储与统一归档进行指导与约束，给出了用于资源描述的元数据规范，但其应用程度及范围均局限于单一学科领域。空间科学领域需要建立一套适用于所有学科领域数据资源的产品组织规范，实现对多学科数据的统一管理、关联发现及综合应用。

作为我国空间科学领域的首个数据中心，国家空间科学数据中心（以下简称“数据中心”）立足于科学数据全生命周期过程及空间科学数据过程管理模型，充分吸纳SPASE和PDS等国际数据模型优势，综合各学科领域数据特点，构建了空间科学数据产品组织模型[8]，实现对数据资源的分层次实体管理，并按照不同层级定义相应的辅助性文件，提升了数据应用的便捷性，为数据开放共享活动奠定了坚实基础。同时，数据中心将空间科学数据产品组织模型应用于中国科学院信息化专项等项目[9-11]的数据管理过程中，并在实践基础上对空间科学数据产品组织模型进行进一步简化与修正。科学数据产品组织模型考虑了学科特色与数据应用的共性特征，定义了数据资源的层级结构与广义元数据，从而实现对各层级数据资源进行统一管理与发现；设计数据说明等辅助性资料，对数据资源的质量、内容等详细信息进行描述，同时将与数据处理应用相关的软件工具与数据进行关联管理，进一步提升了数据的应用便捷性与科学易用性。

作为数据中心的指导性规范，科学数据产品组织模型贯穿于空间科学数据标准化组织、规范化管理等数据活动全过程，并应用于空间科学先导专项、国家子午圈计划等重大科技任务及国家重点研发计划中。本文将结合国家重大空间科学任务的数据管理实践活动，系统地介绍科学数据产品组织模型及其在空间科学数据管理中的应用，总结模型应用的效果。

1　科学数据产品组织模型

科学数据产品组织模型描述了学科数据实体资源、数据描述资源和数据标注资源的组织层次关系结构，是空间物理、空间天文、行星科学等学科领域科学数据产品组织的基本框架。对数据组织、数据描述、数据发现和共享应用都具有重要的指导意义。

1.1　模型框架

科学数据产品组织模型划分为3个层级：数据产品文件、数据集及数据卷（图1）。针对不同层级的数据实体进行不同粒度的组织与管理。

（1）数据产品文件：是由一个或多个数据对象配合数据标签组成，是数据产品组织管理的最小粒度实体，也是数据集的最小组成单元。

（2）数据集：是具有相同的起源、处理过程、应用价值或相互关联的一系列数据产品文件的集合，并配备学科元数据、核心元数据以及相关辅助文档和软件工具，以支持在数据集层次的统一管理、检索和释义。数据集是数据管理与发布共享的主要形态。

（3）数据卷：是由一系列具有关联关系的数据集组成，并配置卷编目与卷索引支持数据卷的定位与解析，是科学数据长期归档存储的组织形态。

1.2　广义元数据

广义元数据包括组织模型中定义的所有数据描述资源和数据标注资源，用于对不同粒度的数据属性信息的描述，其功能与内容也有所区别。各层级数据实体必须配置相应的元数据，主要包括数据标签、核心元数据、学科元数据、辅助性文档、卷编目及卷索引。

（1）数据标签：是对数据产品文件中的数据对象组织形式进行细致描述，以关键字、文件头等形式出现，用于产品文件的自解释。

（2）核心元数据：是面向数据集查询而设计的共性属性，包括数据集名称、摘要描述、观测来源等基本信息，可以跨学科使用。

（3）学科元数据：是依据学科规范对数据集的详细描述，包括数据集的承继关系、质量信息、参量信息等，可用于学科精细化检索与应用。

（4）辅助性文档：是辅助数据集释义与应用的技术文档，如数据说明、质量报告、处理报告等。

（5）卷编目：是数据卷摘要性信息与组织结构信息，包括卷主题、质量信息、内容描述等。

（6）卷索引：分为数据集索引和数据卷，主要描述数据集或数据卷的目录结构、文件路径等信息，用于对数据集或数据卷内容的快速定位。

2　空间科学数据系统设计

在空间科学数据管理活动中，依据科学数据组织模型，数据生产者负责生产规范化数据产品文件，并与数据中心共同制定学科元数据规范，编制元数据与辅助性资料，进一步根据使用惯例组织成数据集。数据生产者将数据产品文件或数据集后汇交至数据中心，数据中心负责对汇交的

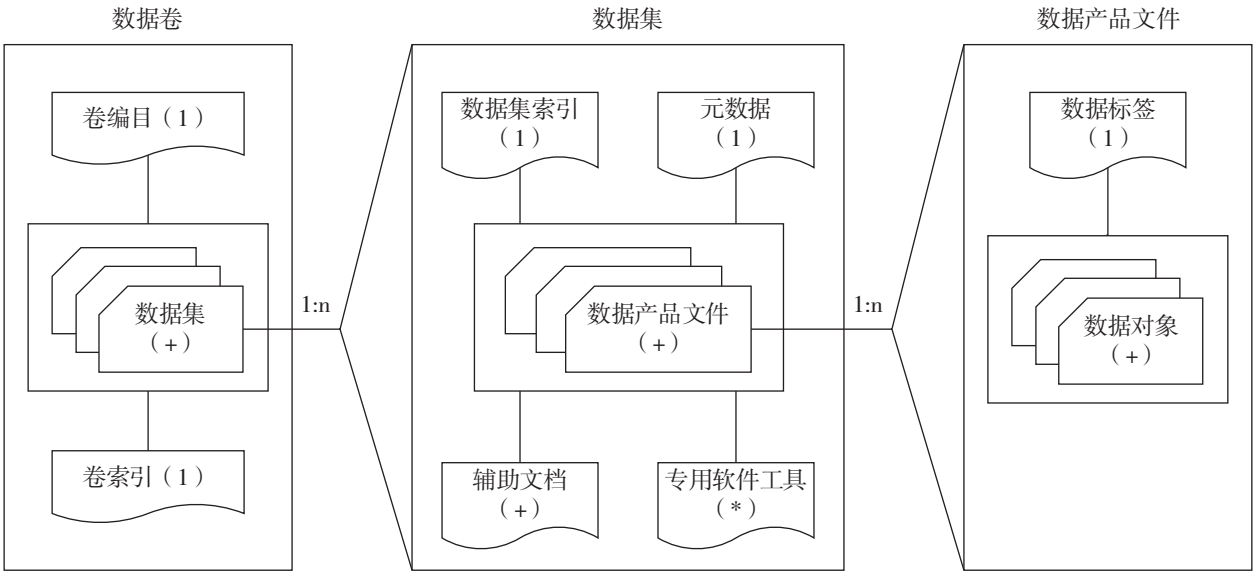

图 1　科学数据产品组织模型框架

数据产品文件或数据集进行校验、质量复核；配置所需的元数据信息、索引信息等，将数据产品组织成数据集；根据数据分级分类管理要求，最终制作成标准数据卷，开展长期安全存储管理；根据数据共享范围及途径要求进行数据集的开放共享。

为了实现空间科学数据的全生命周期管理，数据中心设计了包含基础设施、业务应用及公众服务3个层次的数据系统（图2）。

2.1　基础设施层

基础设施主要用于支持数据中心各业务系统的网络资源、计算资源、存储资源及基础软件资源。其中，网络资源主要包括国际科技网、中国科技网、重大项目数据专网等，计算资源包括CPU集群、GPU集群等，存储系统包括NAS存储系统、磁带库、光盘库等，基础软件资源包括数据库系统、基础设施综合管理系统等。这些资源共同构成了数据中心一体化的基础运行环境。

2.2　业务应用层

业务应用层是面向数据全生命周期的各个环节设计，包括数据汇集系统、处理系统、管理系统、档案系统及发布系统。数据汇集系统与重大专项的数据系统底层对接，并为科技计划提供在线服务，支持自动与人工两种数据汇集模式，并支持对汇集的数据资源进行标准化与质量复核。数据处理系统主要完成元数据编目、数据集/数据卷制备及质量审核，确保元数据信息齐全，数据集/数据卷内容完整、连续，格式规范。数据管理系统对数据集进行管理，形成按照标准数据集库，对数据集采用“在线—近线—离线”三级存储机制进行统一管理，支持对数据集进行常规管理，如出入库、检索、浏览及统计分析等。数据档案系统是数据卷进行统一管理并进行分级存储、本地备份及远程灾备。数据发布系统是面向数据中心门户提供数据分发与推送服务，保证数据共享的实时性、灵活性及便捷性。同时，整个系统还具有相应的安全保障体系与数据标准规范体系，为业务系统设计与建设提供技术、接口、数据标准等保障，并对数据、用户、系统等进行权限管理与实时监控，为业务系统的信息安全及稳定运行提供支持。

数据质量是数据价值的重要体现，只有经过数据质量评估的数据才能具备应用价值，才能

图2　空间科学数据系统架构

在科研创新活动中做出应有的贡献。为此，数据中心专门数据管理系统中设计了数据质量审核软件，针对不同领域或项目数据的规范性、完整性、标准性制定不同的质量审核策略，利用新技术、新算法研发系列审核工具。如在空间科学先导项目中，针对卫星遥测数据采用了大数据异常识别算法，对数据中的异常值进行定位，并对异常发生原因进行初判，为数据异常情况时的快速处置提供辅助性依据。

2.3　公众服务层

数据中心建设了数据门户网站（https://www.nssdc.ac.cn），面向公众提供数据共享、软件工具应用、动态信息及科普宣传服务，并提供基于工作流的数据应用环境，极大地节省了用户的数据查询、获取及应用的时间成本，有效地提升了科研创新活动的效率。

3　应用实践

在中国科学院战略性空间科学先导专项实施过程中，国家空间科学数据中心依据数据全生命周期过程管理指南，深入参与项目数据管理活动，指导项目各方开展数据汇集与处理、管理、归档及发布等关键数据活动，积极推动科学数据组织模型在各环节的应用。

第一，在数据汇集与处理阶段，根据数据的处理程度将数据产品进行分级分类，制定标准的数据产品分级定义与格式说明；对各级各类数据产品文件的数据标签内容、产品存储格式、数据对象组织方式等进行明确描述，确保数据产品能够具备数据解析与检索的必要信息。

数据中心负责原始数据的处理，针对常用的数据存储格式，对数据标签基本信息提出了要求，包括但不限于数据名称、数据级别、数据时间及生产机构、数据校验码等 10 余项属性。在上述规范性文档的指导下，数据中心将数据处理生成规范的数据产品文件，并配置相应的数据产品文件处理报告，便于数据管理者及使用者掌握数据处理程度。

科学应用系统负责标定级数据的处理与汇集，与数据中心共同制定数据归档计划，明确数据产品汇集的方式、内容及时频等，并持续开展数据汇集工作。依据数据分级定义及格式说明，对数据进行标定处理，生成具备详尽标签信息的数据产品文件。并根据学科使用惯例及组织模型，将数据产品文件按照一定规则组织成文件集合，如按时间序列、观测号、实验号等，再按照数据归档计划的约定将数据提交至数据中心。数据中心负责对汇交的数据产品文件进行校验与复核，对数据产品文件进行形式审查，包括但不仅限于数据集目录结构、数据产品文件的标签信息与存储格式等。

第二，在数据管理阶段，数据中心对经过审核的数据产品文件进行标准化编目形成数据集，并制定元数据规范，完成核心元数据、学科元数据、数据说明等广义元数据的编制。数据中心将同一设备产生的同级同类同版本的数据产品文件按照规范的数据集目录（图 3）进行组织与管理，将数据产品文件、广义元数据、相关的软件工具及其他示意性资料（数据样例、缩略图等）作为整体进行统一管理。

制定核心元数据和数据说明规范，并遵循学科元数据标准开展数据集与数据卷的制备。面向空间科学先导专项数据资源的多学科、多级别、多类型等特点，数据中心设计了涵盖数

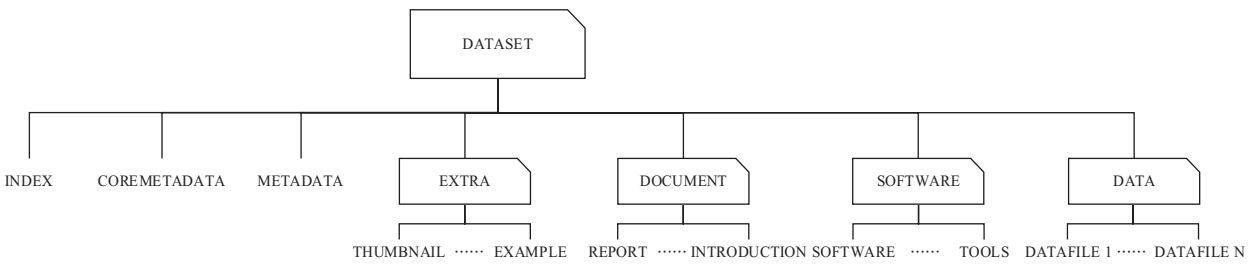

图 3　空间科学卫星数据集目录结构示意

据基本信息、生产信息、共享信息、机构信息、来源信息等31项内容的核心元数据，并通过核心元数据实现跨学科数据查询。同时，依据不同卫星所属的学科领域，制定标准元数据及其数据字典规范，对物理要素、观测设备、溯源与承继关系、管理发布机构等进行详细描述，对数据内容进行详细描述，从而支持数据集的参数检索与关联应用。制定了数据说明模板，对数据的处理方法、数据质量、使用要求、权益声明等信息进行细致描述，并根据实际情况自由扩展模板提纲，辅助数据使用者在无需获取数据集的情况下，快速、系统及全面地了解数据集的基本信息。

第三，在数据归档阶段，遵循科学数据组织模型，将数据集按照其关联关系以规范的目录结构组织成标准数据卷，进行长期存储管理、本地备份和异地灾备。科学卫星在轨观测时会同时产生科学数据、工程辅助数据，并依据卫星的科学目标及有效载荷的性能情况开展仪器定标、地面对比实验、仿真模拟等活动，进而产生定标数据、对比实验数据、仿真数据等。因此，数据中心提出了数据卷命名规则、卷索引及卷编目格式要求，将这些不同种类的数据集进行关联编目，作为整体存储在同一目录下，形成标准数据卷（图4）。卷索引主要包含数据卷中各类文件的位置信息，便于数据系统快速定位数据卷的内容；卷编目则对数据卷摘要信息及目录结构进行描述，主要包括数据卷的生产信息、版本信息、与其他数据卷的关联关系等。

第四，在数据发布阶段，数据中心制备完成的数据集通过国家空间科学数据中心门户网站和各卫星任务数据网站进行线上及线下共享。所有核心元数据同步发布在中国科技资源共享网。用户可以在发布页面上通过元数据进行数据的查询、浏览与下载，并获取在线共享数据集的目录结构与数据说明信息，有效地提升了数据获取效率。

在空间科学先导专项的实施过程中，数据中心将科学数据组织模型贯穿数据处理、汇集、管理、归档与发布的全过程，并制定了一系列

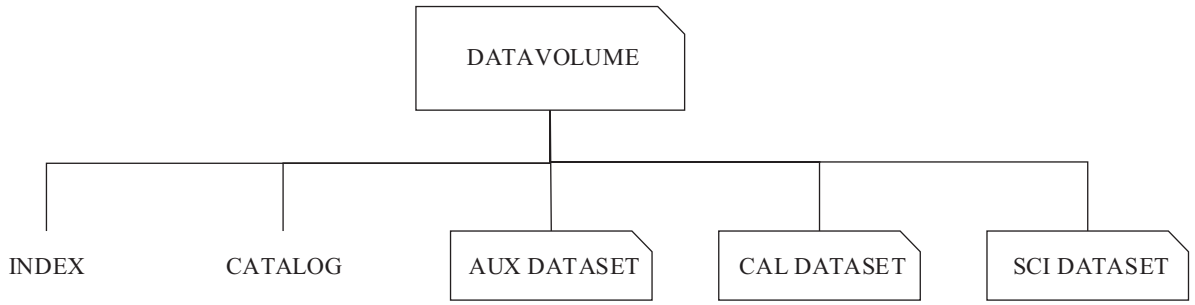

图4　空间科学卫星数据卷目录结构示意

的元数据规范、格式要求及技术文档，为“悟空”“慧眼”“墨子”等在轨科学卫星的数据规范化生产、标准化管理及高效共享提供了有力支持，实现了跨学科数据的统一管理。

4　结语

本文从空间科学数据管理系统设计及实践活动入手，系统地介绍了科学数据产品组织模型在空间科学数据管理过程中的应用。应用结果表明，从数据产品文件粒度对数据资源进行规范化有序管理，能够有效地降低科学数据管理的复杂程度，对保证数据资源的规范性、完整性、易操作性具有重要意义。科学数据产品组织模型能够较好地满足国家空间科学数据中心对多类型、多来源数据管理与共享的业务需求，较好地在项目数据汇交过程中发挥指导性作用。同时，在学科资源交叉管理方面，组织模型不存在明显的学科

（下转第96页）

印发[EB/OL]. (2012−02−11)[2021−10−08]. https://www.cas.cn/sygz/ 201902/t20190221_4679910.shtml.
[16] 科技部, 财政部.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科[EB/OL] . (2019−06−05) [2019−06−10]. http://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/qtwj/qtwj2019/201906/t20190610_147031.html.
[17] 中国科学数据 [EB/OL]. [2021−10−08]. http://www.csdata.org/p/static/33/.
[18] 王卫军, 李成赞, 郑晓欢, 等.全球科学数据出版发展态势分析: 基于 Web of Science 数据库的调研[J].中国科学数据:中英文网络版,2021,6(3):262−280.
[19] MARTONE M.Data citation synthesis group: joint declaration of data citation principles[M]. San Diego CA: FORCE11, 2014.
[20] 斯普林格数据共享政策[EB/OL]. [2021−10−08]. https://www.springernature.com/gp/authors/research−data−policy/research−data−policy−types.
[21]《数据分析与知识发现》编辑部支撑数据提交要求[EB/OL]. [2021−10−08]. http://manu44.magtech.com.cn/Jwk_infotech_wk3/fileup/2096−3467/NEWS/20161213090914.pdf.
[22] 中华外科杂志编辑部通知[EB/OL]. (2017−08−15) [2021−10−08]. http://zhwkzz.yiigle.com/notice/index.htm.
[23] Generalist Data Repository Grid[EB/OL].(2020−06−04)[2021−10−08].https://fairsharing.org/collection/GeneralRepositoryComparison.
[24] re3data.org [EB/OL]. [2021−10−08]. https://www.re−3data.org/browse/by−country/.
[25] re3data.org [EB/OL]. [2021−10−08]. https://www.re−3data.org/browse/by−subject/.
[26] 孔丽华, 邵明玥.科学数据出版内容与案例分析[J].科研信息化技术与应用, 2018, 9(6): 39−46.
[27] 黄国彬, 王舒, 屈亚杰.科学数据出版模式比较研究[J].大学图书馆学报, 2018, 36(1): 34−40, 33.
[28] 魏奉思.《国家重大科技基础设施子午工程专题》卷首语[J]. 中国科学数据(中英文网络版), 2021, 6(2): 6.

（上接第88页）

壁垒，能够便捷地推广至其他学科领域，并且通过调整元数据的体系设计和标准规范保证了数据的专业性与学科特色，为数据资源关联发现、综合分析与应用提供保障。

在应用实践活动中，使用了国家空间科学中心公共技术服务中心空间科学数据融合计算平台提供的计算服务。同时，感谢国家科技资源共享服务平台—国家空间科学数据中心（https://www.nssdc.ac.cn）为本文中的实践活动提供的支持。

参考文献

[1] 邹自明, 胡晓彦, 熊森林, 等. 空间科学大数据的机遇与挑战 [J]. 中国科学院院刊, 2018(8): 877−883.
[2] 邹自明, 佟继周, 熊森林, 等. 大数据时代空间科学领域的科研信息化实践与成果 [J]. 大数据, 2016(6): 83−96.
[3] 国务院办公厅.科学数据管理办法[EB/OL]. (2018−04−02)[2018−06−30]. http://www.gov.cn/zhengce/content/2018−04/ 02/content_5279272.htm.
[4] KING T, THIEMAN J, ROBERTS D A.SPASE 2.0: a standard data model for space physics[J]. Earth science informatics, 2010, 3: 67−73.
[5] MCMAHON S K.Overview of the Planetary data system [J]. Planetary and space science, 1996, 44 (1): 3−12.
[6] Jet Propulsion Laboratory (JPL). Planetary data system standards reference [M]. California: JPL California Institute of Technology, 2009.
[7] 郑岩, 邹自明, 佟继周, 等. 行星科学数据系统(PDS)标准规范的研究 [J].科研信息化技术与应用, 2009(1): 1−8.
[8] 熊森林, 邹自明, 胡晓彦, 等. 空间科学数据产品组织模型 [J]. 农业大数据学报, 2019 (4): 30−36.
[9] 佟继周, 邹自明, 傅衍杰, 等. 空间科学e−Science应用: 空间科学虚拟观测台 (VSSO) [J]. 科研信息化与应用, 2011, 2(1): 61−68.
[10] 姜旭, 佟继周, 崔辰州, 等.基于虚拟天文台的HXMT卫星数据检索发布系统设计与实现[J]. 天文研究与技术, 2014, 11(4): 378−387.
[11] 熊森林, 郑岩, 闫振中, 等. 南美空间天气信息系统设计与初步实现 [J]. 科研信息化技术与应用, 2017(1): 47−57.