

DOI: 10.19788/j.issn.2096-6369.190409

# 科学数据共享系统的现状与趋势

李云婷<sup>1,2</sup> 温亮明<sup>1,2</sup> 张丽丽<sup>1</sup> 黎建辉<sup>1\*</sup>

(1. 中国科学院计算机网络信息中心, 北京 100190;

2. 中国科学院大学, 北京 100049)

**摘要:** 数据密集型科研已经成为大数据时代科学发展的新范式, 科学数据开放共享已成科技界的普遍共识。在长期实践中, 科学数据共享形成了以科学仪器、数据平台、数据出版、众包处理、数据交易等为典型代表的不同模式。与之对应, 针对不同的领域和应用场景出现了种类繁多的解决方案, 如仓储型、联邦服务型、数据分发型和按需计算与分析云服务型等。本文在对上述四类主流科学数据共享系统的服务内容、技术特征、应用场景与代表性系统分析比较的基础上, 提出科学数据共享系统未来发展的趋势, 并以中国科学院战略性科技先导专项“地球大数据科学工程”研发的地球大数据云服务平台为典型案例, 进行了深入的剖析。本文认为, 未来的科学数据共享系统将围绕着科学数据全生命周期管理的需求, 形成具备数据获取、存储、分发共享、计算分析、智能服务等功能于一体的融合型云服务系统, 并将实现数据的FAIR化、智能关联和机器可理解, 促进数据共享良性生态的形成。

**关键词:** 科学数据共享系统; 数据共享; 数据融合; 智能处理; 数据生态; 科学数据管理; 科学数据; 数据系统

**中图分类号:** TP315

**文献标识码:** A

**文章编号:** 2096-6369 (2019) 04-0086-12

引用格式: 李云婷, 温亮明, 张丽丽, 等. 科学数据共享系统的现状与趋势[J]. 农业大数据学报, 2019, 01(04): 86-97.

Li Y T, Wen L M, Zhang L L, et al. The Status and Trends of Scientific Data Sharing Systems[J]. Journal of Agricultural Big Data, 2019, 01(04): 86-97.

## The Status and Trends of Scientific Data Sharing Systems

Li Yunting<sup>1,2</sup> Wen Liangming<sup>1,2</sup> Zhang Lili<sup>1</sup> Li Jianhui<sup>1\*</sup>

(1. Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China;

2. University of Chinese Academy of Sciences, Beijing 100049, China)

**Abstract:** Data-intensive research is emerging as a new paradigm for science discovery in the era of big data, and the use of open data has become common in the scientific community. Over time, different models of scientific data sharing have emerged, including scientific instruments models, data platforms models, data publishing models, crowd-sourcing and data market models. Correspondingly, a variety of solutions have emerged for different fields and applications, such as data repositories, data federated services systems, data distribution systems, and on-demand computing and analysis cloud services systems. This paper examines development and future trends in scientific

收稿日期: 2019-10-25

基金项目: 中国科学院战略性先导科技专项(A类)子课题(XDA19020104)

第一作者简介: 李云婷, 女, 硕士生, 研究方向: 科学数据云存储、分布式存储、科学数据管理; E-mail: liyunting@cnic.cn

通讯作者: 黎建辉, 男, 博士, 研究员, 博士生导师, 研究方向: 大数据资源开放共享、大数据管理技术、大数据计算与分析技术; E-mail: lijh@cnic.cn

data sharing systems, using the Big Earth Data Cloud Services Platform as an example. It analyzes and compares the typical services and technical characteristics, using scenarios and representative systems of the above-mentioned four types of mainstream scientific data sharing systems. Our analysis suggests that future scientific data sharing systems will focus on the need to manage the full life-cycle of scientific data and will converge into a cloud service system providing functions such as data acquisition, storage, distribution and sharing, analysis, and intelligent services. By making data FAIR (Findable, Accessible, Interoperable and Reusable), machine understandable and AI-Ready, promote the formation of data sharing eco-systems.

**Keywords:** scientific data sharing system; data sharing; data fusion; intelligent processing; data ecology; scientific data management; scientific data; data system

## 1 引言

随着科学数据采集能力的不断提升,科学数据类型和规模激增,数据生产者与数据消费者的供需矛盾日趋突出,传统的“自给自足”式科研理念已无法支撑当前科研活动的需求<sup>[1]</sup>。为此,推动科学数据开放共享成为现代科学研究的必然诉求<sup>[2-3]</sup>。与此同时,随着大数据与云计算技术的发展,科学数据管理面临着动态化、流水线化、智能化等方面的挑战<sup>[4]</sup>,传统的单场景式科学数据共享系统难以满足当前超大规模关系数据管理、多源数据关联和知识发现、实时高效数据处理等科研环境的需求<sup>[5]</sup>。为突破传统数据管理系统在性能、安全、效率等方面的瓶颈,现代科学数据共享系统需满足分布式、多功能、多场景需求。

诸多数据密集型科研都面向领域科学数据的开放共享提供了种类繁多的系统性解决方案。例如, re3data.org 将注册在库的数据库分为领域型、机构型和其他三类。Grossman<sup>[6]</sup>将基因数据共享系统分为数据湖、数据云和数据共同体三类。全球生物多样性信息网络(GBIF)致力于生物数据资源的集中服务;世界三大基因数据库(GenBank、EBI、DDBJ)则实现了全球范围内相关资源集中与三大库间的协同服务;此外,以多点数据收割为特色的数据(如地球科学观测类数据)服务则常采用分布式、网格化、云化系统架构等等。国内也颇多特色实践,例如数据出版方面的 ScienceDB 系统、全球变化科学研究数据出版系统(GCdataPR),空间天文科学领域的具有高用户覆盖度的空间观测虚拟观测台(VSSO),数据交易方面有数据堂等,以及我国首个获批筹建的国家级综合性基因库 CNGB(China National GeneBank)等。这些系统为特定场景的科研数据开放提供了定制化的解决方案,皆为科学数据共享系统发展的重要代表。本文从

主要服务、技术特点、使用场景和代表性系统等维度入手,分四大类依次介绍主流科学数据共享系统的发展现状,并对科学数据共享系统的未来发展趋势进行分析判断和例证。

## 2 主流共享系统的技术内容、特征与实践

所谓科学数据共享系统,即为推动科研活动中数据开放共享而提供的以 ICT 技术和软硬件环境为一体的系统性解决方案。科学数据共享系统是科研领域之中的计算机系统,它们承载着数据资源的管理与开放共享,也是科学数据开放共享政策与实践的展示平台。所有面向科研数据管理的系统皆可成为科学数据系统。但考虑到面向开放服务趋势下,限制性访问(如离线、近线)数据库仍大量存在,这里笔者提出“科学数据共享系统”,指代一类以开放服务为理念的、围绕科研数据的管理与共享的数据库系统。

### 2.1 主流共享系统的分类

为了更好地阐释主流共享系统的发展情况,以下结合开放科学数据的主要模式以及科学数据管理系统的技术特征与服务特色,对主流科学数据共享系统进行分类归纳。

科学数据开放共享在实践中逐步形成了适合当前共享环境需求的一些模式,如科学仪器模式、数据平台模式、数据出版模式、众包处理模式、数据交易模式等<sup>[7,9]</sup>,一系列与共享模式相匹配的共享系统也应运而生。结合主流数据共享模式的数据来源、数据开放程度、数据服务特征等,并考量各数据共享系统中的数据流、组织形态、技术应用等,本文认为当前主流

的科学数据共享系统可以归纳为四种主要类型,即仓储型系统、联邦服务型系统、数据分发型系统和按需计算与分析的云服务型系统,这些系统在一定程度上解决了数据共享的现实困境。

## 2.2 仓储型系统

### (1) 技术内容与特征

仓储型系统,即收集与汇聚用户所提交的数据文件,并对所有数据资源进行统一存储与管理,对外主要提供数据集的共享与发布服务。系统主要面向科研人员,收集最新科研数据,以促进数据的传播、引用与重用,同时也面向机构、项目组、出版商等服务。就共享服务而言,科研人员可通过门户网站等方

式随时随地接入系统,方便地完成数据提交、管理、共享、检索、发现等操作。就发布服务而言,仓储型系统能支持图像、视频、代码、大文件等数据,极大地丰富了当前科学数据的出版方式。

仓储型系统的一般技术框架如图 1 所示。系统的数据来源于用户,用户可通过门户网站、提交工具或 API 等方式向系统提交数据文件。系统汇聚用户提交的数据资源,并进行统一管理。一般而言,使用云存储或数据库集群等技术提供安全可靠的数据存储管理,为每个数据集分配全局唯一的标识符,通过系统抽取、用户录入等方式形成丰富的元数据描述信息,最终形成数据集产品,对外进行发布与出版,用户可通过门户网站搜索、过滤、发现所需数据,并对数据进行引用。

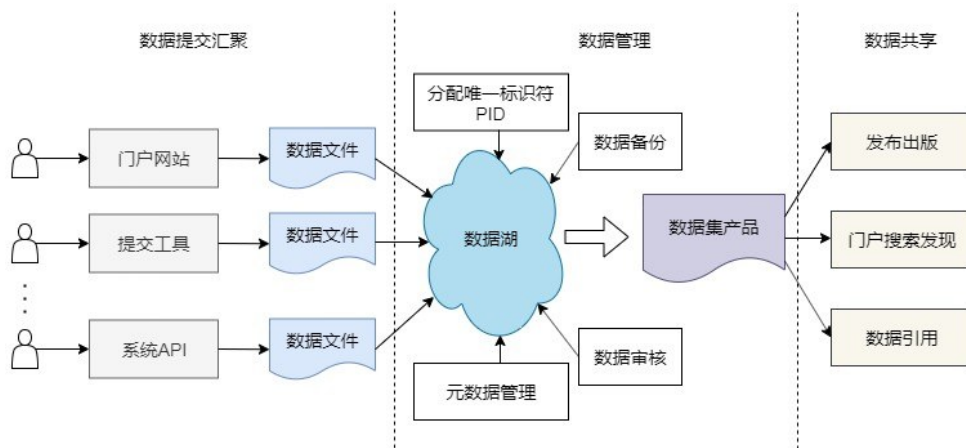


图 1 仓储型系统技术框架

Fig.1 Technical framework of repository system

### (2) 典型系统应用

仓储型系统通常可接收任意领域、任意格式的数据文件进行共享、发布和出版。当前代表性的系统为

Figshare<sup>[10-11]</sup>、Dryad<sup>[12-13]</sup> 及 ScienceDB<sup>[14]</sup> 等,如表 1 所示。

表 1 代表性仓储型系统

Table 1 Information of typical repository system

系统	所属机构	服务对象	提交方式	标识符	许可协议	质量控制
Figshare	英国 Digital Science 公司	科研人员、机构、出版商	门户网站上传、桌面工具、API	DOI	CC0	不提供标准的数据审核机制,由提交者自行核查
Dryad	美国国家科学基金会	科研人员、机构、出版商	门户网站上传、API	DOI	CC0	系统管理员进行数据质量审核
ScienceDB	中国科学院	科研人员、出版商、科研项目、机构高校	门户网站上传、FTP 工具、线下拷贝	DOI、CSTR	CC0、CC BY 4.0	提供数据质量审核

仓储型系统应用广泛,一个典型的服务场景如下:用户 A 向系统提交数据文件至个人空间,并录入

元数据,随后可对空间内大量数据文件进行组织、排序、管理等操作,经系统管理员审核通过后,以数据集

形式进行共享。一般而言,数据集可共享至项目组,在组内成员间完全共享;也可按照许可协议向公众共享。数据集共享发布后,另一用户B即可通过门户网站搜索并发现该数据资源,按照标准格式进行引用,并给予反馈。而数据上传者,即用户A也可通过系统查看相关数据的反馈评价。

## 2.3 联邦服务型系统

### (1) 技术内容与特征

联邦服务型系统,即在逻辑上集成多源数据,对外作为一个整体提供共享服务,支持数据资源的统一发现和访问。系统将分散在各地的组织机构、数据中心、科学数据库等联合起来,形成一个庞大的科学数据网络。其中,数据资源由各分布式节点自行管理和维护,系统则提供所有资源的统一目录,建立全局索引,对外进行共享和发布。系统往往整合了较大范围甚至是全球范围内的海量数据资源,使得用户能对来自不同节点的数据进行统一查找,提升了数据的可发现性和可重用性。

联邦服务型系统的一般技术框架如图2所示,通常包含数据节点、管理节点和服务接口三部分。系统对物理上分散的各数据节点仅进行逻辑上的集成,数据资源的具体存储、数据质量等由各数据节点自行负责。管理节点则对接入的数据节点中的资源进行注册,分配全局唯一标识符,维护所有资源的统一目录。此外,管理节点往往还能进行总体的监控统计,协调完成跨节点的数据复制。用户一般可通过门户网站、软件工具包、API等方式与系统交互,进行数据组织、搜索、访问、分析等操作。得益于松耦合的联邦式架构,系统能支持新数据节点的快速加入,具有弹性易扩展特性。

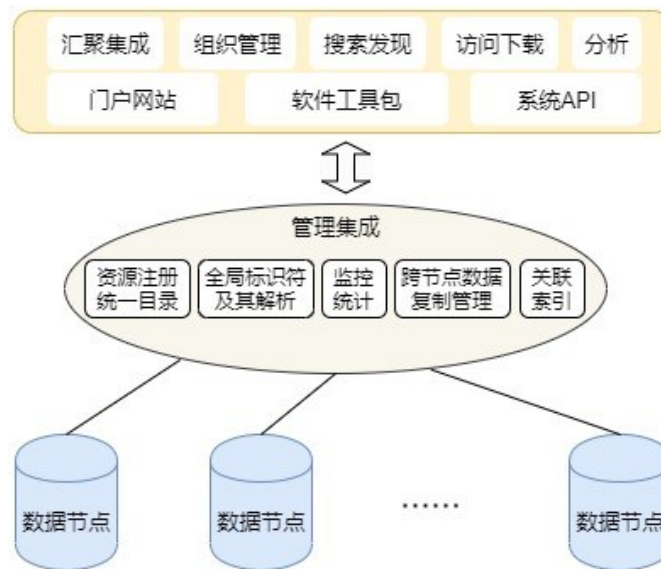


图2 联邦服务型系统技术框架

Fig.2 Technical framework of federal service system

### (2) 典型系统与应用

联邦服务型系统是当前主流的科学数据系统,国内外数据共享平台大多依此构建,典型代表为Data-ONE<sup>[15,17]</sup>、GEOSS<sup>[18,20]</sup>和中国科学院数据云<sup>[21-22]</sup>等,如表2所示。

系统的一个典型应用场景如图3所示。用户A和机构B分别将其数据资源及元数据存储于数据节点1和数据节点2,各数据节点对其所属数据进行自治管理,并向管理节点提交注册。管理节点为数据分

配全局标识符,并更新到资源统一目录。随后,用户C可通过门户网站搜索并发现相关数据资源。若用户C需使用A的数据,则可通过系统提供的数据位置,直接连接到数据节点1以访问数据。同样地,用户C也可通过连接到数据节点2,获取机构B提供的的数据资源。



表2 代表性联邦服务型系统  
Table 2 Information of typical federal service system

系统	所属机构	数据来源	建设目标	主要设计思想
DataONE	美国国家科学基金会 提供支持	全球范围内的地球科学、生物 及环境数据	支持分布于全球各地数据中心的快速数 据发现和访问	1. 最好能建立在已有数据中心的基础上; 2. 适应通用软件、标准及其发展
GEOSS	国际地球观测组织	全球范围内对地观测数据与 产品	建立一个综合、协调、可持续的全球地球 综合观测系统	1. 系统的系统型; 2. 强调基于标准接口的 互操作性; 3. 用户需求驱动; 4. 强调数据 及元数据的格式与标准
中国科学院 数据云	中国科学院	中国科学院内各学科领域数 据资源	推动中科院全院的数据整合、归档、汇聚 和发布共享服务	1. 联合各领域的科学数据库; 2. 多源异构数据集成, 形成统一的资源服 务目录

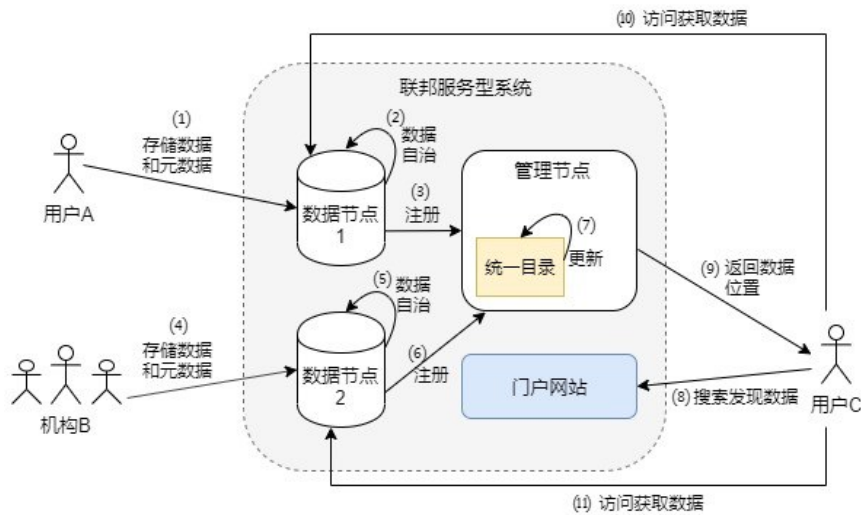


图 3 联邦服务型系统典型应用场景

Fig.3 Typical application scenario of federal service system

2.4 数据分发型系统

(1) 技术内容与特征

数据分发型系统主要依托大型科学装置、重大科学实验和科学监测站点, 汇聚大量专业规范的数据资源, 面向科研人员和公众进行分发。系统主要提供数据检索和数据获取两种核心服务。通过数据检索服务, 用户能快速浏览数据信息, 根据时空信息、所属系列等条件筛选出所需数据资源。通过数据获取服务, 如在线下载、电子邮件等方式, 系统可将相应的资源分发至用户。大多数分发型系统还能根据用户对数据进行深加工。

数据分发型系统的一般技术框架如图 4 所示。大量来源于某个或某几个大型科学装置与观测设备等的科学数据, 经规范化处理后汇聚到综合管理中心进行集中质量控制, 按照其公开的数据共享策略对外

提供访问、下载服务。此外, 系统还能汇聚其它已有的数据集产品。综合管理中心对汇聚的数据资源进行管理, 一般包括元数据维护、数据存档、发布、形成数据目录等, 以支持对外共享和分发服务。用户可检索数据、浏览元数据信息、下载数据到本地, 也可从系统提供的云环境中获取相应数据资源。

(2) 典型系统应用

根据数据来源是否单一, 一般可将数据分发型系统大致分为两类。一类依托某个大型科学装置, 向用户分发该装置产生的数据资源, 如以美国 Landsat 陆地卫星系统<sup>[23-24]</sup>和欧洲航天局 Sentinel 哨兵卫星系统<sup>[25-26]</sup>为代表的遥感卫星数据资源, 和以郭守敬望远镜 LAMOST<sup>[27,29]</sup>和斯隆数字巡天 SDSS<sup>[30-31]</sup>为代表的天文望远镜数据资源。另一类数据来源更为广泛, 提供综合性服务, 如地理空间数据云 (GSCloud)<sup>[32]</sup> 提供包括 Landsat 系列、MODIS、DEM、高分系列等在内的

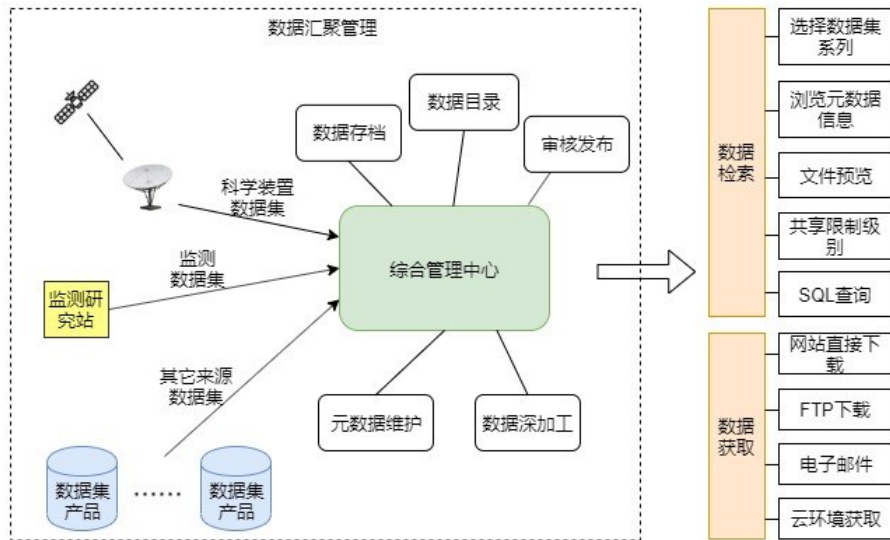


图4 数据分发系统技术框架

Fig.4 Technical framework of data distribution system

多种遥感数据资源。国家生态系统观测研究网络(CNERN)<sup>[33]</sup>则建立在中国生态系统网络(CERN)<sup>[34]</sup>的基础上,汇聚了来自综合中心、51个台站、2个子网的生态系统数据资源,面向用户进行分发<sup>[35]</sup>。

数据分发型系统在各领域得到了广泛应用,一个典型的服务场景如下:用户通过门户网站浏览检索数据,并筛选出所需数据产品。一般而言,若为完全公开共享级别数据,可直接进行下载;若为限制共享级别数据,用户需先提交数据申请订单,经系统管理员审核通过后,系统将通知用户并为用户分发相应数据。

## 2.5 按需计算与分析的云服务型系统

### (1) 技术内容与特征

按需计算与分析的云服务型系统,一般在数据共享的基础上,提供一个稳定高效的云端计算环境,为用户提供数据分析服务,主要表现为两种形式。一种是用户向系统提交个人数据,通过系统提供的多种分析模式在云端进行计算。另一种是系统本身发布和公开了一系列数据集,用户无需下载数据就可直接与系统交互,进行在线计算。系统通常能提供多种交互分析方式,主要包括利用相对固定的工作流等进行经典算法模型分析,使用Notebook等进行交互编程分析,通过软件工具开发复杂代码工程进行分析。

按需计算与分析的云服务型系统的一般技术框架如图5所示。系统的核心组成部分是云计算平台,

主要包括数据管理、算法模型管理、集群计算等三部分,支持通过多种交互分析方式为用户提供计算服务。用户可通过浏览器网页、命令行工具、API等方式直接与系统进行交互,其中以浏览器网页方式为主。对于一般公众或需使用经典算法模型的场景,可通过工作流服务选择算法模型直接进行分析。对于专业科研人员,可选择Notebook进行在线交互编程或使用领域内常用的软件工具包进行分析。对于科研项目团队及复杂分析场景,可选择开发代码工程项目进行分析。系统一般能通过大规模集群节点进行快速计算,并将计算结果反馈给用户。

### (2) 典型系统应用

按需计算与分析的云服务型系统主要解决了用户个人计算资源不足和计算手段单一等问题,实现了数据、算法和算力的有机融合。澳大利亚地球科学“数据立方体”(Data Cube)是一个云端数据处理方案,将多源对地观测数据整合为一个网格化数据分析环境,包括数据准备、软件环境、高性能计算环境三个基础核心组成部分<sup>[36]</sup>。Data Cube将原始的地球观测数据从经度、纬度、时间维度处理为可直接用于分析和计算的cube格式<sup>[37]</sup>,由澳大利亚国家超算平台提供计算支持,用户可通过多种交互方式实现对相关数据资源的处理和分析。此外,地理空间数据云(GS-Cloud)除数据分发外,也提供在线计算服务,用户可通过多种模式在线切割全球数字高程数据。GS-Cloud还基于容器技术构建科学数据云端交互编程

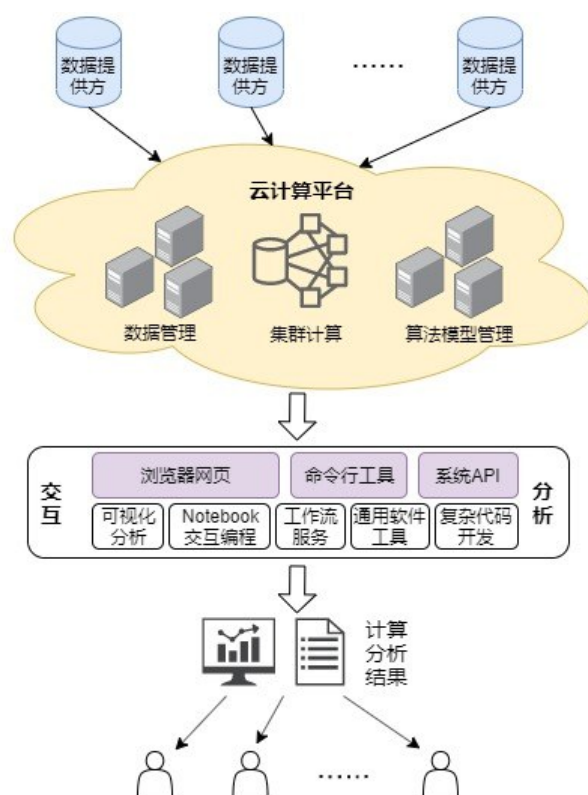


图5 按需计算与分析的云服务型系统技术框架

Fig.5 Technical framework of computing and analysis cloud system

分析服务,用户可通过 Notebook 与系统进行交互,方便快捷地实现数据分析和处理。

### 3 四类科学数据共享系统的比较分析

通过以上分析可以发现,各类型系统在服务内容、技术特点、使用场景等方面均有一定差别,表3对上述四种数据系统进行了比较分析。

表3 四类科学数据共享系统的比较分析

Table 3 Comparative analysis of four types of scientific data sharing systems

系统类型	数据来源	主要目标	核心服务	适用场景	缺陷不足
仓储型系统	来源于用户提交	收集科研人员的最新数据文件进行共享发布,形成科学数据新型出版模式	数据共享、数据出版	使用者希望发布数据资源或搜索最新科研数据	数据来源无法保证,数据质量参差不齐
联邦服务型系统	分散在各地的数据节点	逻辑上汇聚多源数据,形成统一资源目录,方便对外共享	数据汇聚、数据共享	使用者对数据类型、范围、体量有较大需求	数据资源散落分布,易受数据所有者影响
数据分发型系统	大型科学装置、科学实验、监测站点	提供领域内专业规范的数据资源,促进数据共享和重用	数据检索、数据获取	使用者对数据质量有多维度、细粒度需求	数据来源和类型较单一,受到行业领域性质局限
按需计算与分析的云服务型系统	用户个人提交或公开数据集获取	提供稳定的云端计算环境,通过多种交互分析模式快速进行数据计算	数据分析	使用者本身缺乏计算工具和环境,但关注分析结果而对具体处理流程要求较低	对数据质量和用户专业性要求较高,对多源异构数据分析处理较困难



通过表3对比分析可知,各类型科学数据共享系统能满足某些管理需求,提供的服务也各具特色。由于科学数据的生命周期过程包含数据获取、汇集、保存、加工、发布、共享、应用等多个流程(如图6所示),这些流程又涉及众多数据科学的理念和方法,通过循环流动实现了价值转变<sup>[38]</sup>,因此以上系统在面向科学数据全生命周期管理时仍然存在着不足。按需计算与分析的云服务型系统对用户的专业性和数据质量要求较高,数据必须经过充分地处理和准备转换为专业通用格式后才能用于计算,且数据资源主要通过用户个人提交或从公开渠道获得,这些数据一般集中

于某一领域内,当面对多源异构数据、跨领域的数据资源时系统处理可能存在困难。仓储型系统侧重于存储数据而其他功能较弱,数据来源无法保证,数据质量参差不齐;联邦型服务型系统适用于协作型优势互补的分散型使用场景,易受数据所有者影响;数据分发型系统的数据质量较高,但数据来源单一,具有一定的行业领域局限性。总体而言,各系统功能侧重点不一,从实现科学数据全生命周期管理的视角看,一种具备存储、分发、计算、分析、服务等多功能的数据系统将成为未来发展的必然。

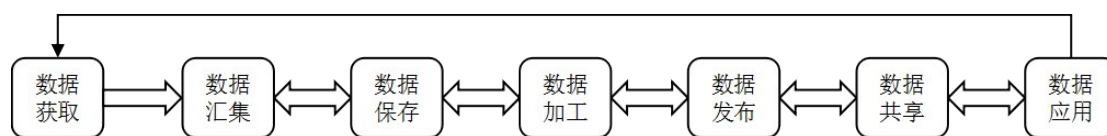


图6 科学数据全生命周期图

Fig.6 Life cycle diagram of scientific data

此外,科学数据共享还面临着数据边界扩张、数据结构异化、数据污染严重、隐私安全受侵等问题<sup>[39]</sup>,数据管理、数据存储、数据处理等方面的挑战依然存在<sup>[40]</sup>。尤其在云环境下,数据处理引擎需要专业化,数据处理平台需要多样化,数据计算分析需要实时化<sup>[41]</sup>,相当一部分数据是未能得到充分利用的,数据的准备、处理、存储和分析都存在很多问题<sup>[42]</sup>。以地球科学为例,其具有异构、多源、多时相、多尺度、高度复杂、非结构化等特点,地球科学研究已经由原本的收集经验数据、理论科学、计算仿真逐渐转向数据密集型科学发现,研究方法已经变成探索大量多学科跨学科数据集之间的相关性,联合国提出的SDG的17个目标中至少有8个能够不同程度、不同方式地通过地球大数据受益。在分析时面临的挑战有:不确定性、不完整性、时间变异性、空间自相关、空间异质性、多分辨率和多尺度等,这些挑战一方面和地球科学本身固有的时空特性有关,另一方面也与数据的应用目标相关<sup>[43]</sup>。解决这些问题我们需要一个平稳和高效的解决方案<sup>[44]</sup>,高性能平台、海量存储技术、综合自动化、高效计算、数据共享、服务系统等新技术成为现实所需<sup>[45]</sup>。Karim Fouad还提出了一个非常值得思考的问题<sup>[46]</sup>:谁有权利访问数据?谁又负责授予访问权限?数据的访问权限如何限制?因此,共享系统在数据质量、数据版权、作者分级、奖励机

制、付费机制等方面均面临挑战。

## 4 未来展望与初步探索

纵观科学数据管理与共享的发展态势,在需求牵引、政策驱动以及技术推动的三重力下,全方位政策体系日渐丰满、新兴技术应用持续助推、数据影响力全面计量<sup>[47]</sup>、数据的流动性逐步提高。但需要注意的是,数据价值仍然没有被充分挖掘,尚未实现从数据到信息、知识、智慧和决策的升华。因此,需要改变现有科学数据共享系统的发展瓶颈,亟待探索符合现代科研环境需求的系统迭代更新。

### 4.1 科学数据共享系统的发展趋势

结合上述分析,本文认为未来的科学数据共享系统将朝着以下方向发展:

#### (1) FAIR化

开放科学的主要理念是实现研究数据不仅被人重用,更要被机器重用,但现实情况是一些描述性文字、图形、表格等很难被机器所理解,针对这一困境,Mark D Wilkinson等提出了FAIR原则<sup>[48]</sup>,该原则的主要思想是为每一个数字对象分配一个全球统一且唯一的数据标识符(PID),通过数字对象体系结构、机器学习、通用对象操作等技术来扩展机器的能力,



使数据资源变得更易于被机器(不仅仅是人类)发现(Findable)、访问(Accessible)、互操作(Interoperable)和重用(Reusable)。FAIR原则将数据资源的范围从传统的结构化数据扩充至更广泛的数字化资源(如数据集、代码、工作流、研究对象等),为数据共享提供了指导<sup>[49]</sup>。在FAIR原则的指导下,未来的科学数据共享系统应该立足于使数据资源从机器可重用逐渐走向机器可操作,最终形成一个FAIR化的数据和服务网络<sup>[50]</sup>。

## (2) 融合化

未来的科学数据共享系统应该综合考虑科学数据全生命周期朝着多功能融合化趋势发展<sup>[51]</sup>,是一个集存储、计算、共享等多种功能于一体、满足纷繁复杂的技术标准、系统间实时交互的融合型系统<sup>[52]</sup>,将数据共享从单纯的共享数据转变成共享服务<sup>[53]</sup>。

## (3) 智能化

如果说开放数据能使我们访问到数据,那么关联数据就是让我们能访问到更多的相关数据,而且这个访问相关的过程是自动化的<sup>[54]</sup>。但自动访问数据和工作流有一定的先决条件<sup>[55]</sup>:首先,为每个数据对象赋予并嵌入PID;其次,针对单个数据对象提供的基本操作(包括创建、读取、更新、删除等),创建更复杂的操作;最后,通过PID访问元数据以建立输入数据、输出数据、处理脚本、工作流、执行用户之间的联系。区块链技术因封装了共识机制、智能合约、加密算法等多种数据库核心技术,被认为是解决数据共享的杀手级应用,已有学者构建了基于区块链技术的科学数据共享框架<sup>[56]</sup>;数据进入系统后,由智能合约自动检测数据体量、格式、内容等是否合规;数据共享完成后,由共识机制自动确认节点贡献程度并给予相应奖励。

## (4) 生态化

科学数据共享不仅流程复杂,而且涉及数据提供者、数据监管者、服务提供者、应用程序开发人员、应用用户、基础结构和工具提供者、数据代理等众多利益共同体<sup>[57]</sup>,需要构建一个具有正向反馈机制的、互惠互利的数据共享生态系统。该系统提供了一个创造、管理和维护数据共享倡议的环境,其由各种各样的数据自治者构成一组网络,这些参与者能直接或间接地产生、提供或消费数据集其他相关资源(如软件工具、服务、基础设施等)<sup>[58]</sup>,每个参与者在系统中扮演一个或多个角色,并通过某种特点关系与其他参与

者进行连接,参与者之间的协作与竞争最终促进了生态系统的自我“进化”<sup>[59]</sup>。未来的科学数据共享系统又该营造起“人人都是数据的提供者,人人都是数据的使用者”的共享环境,系统不仅要满足基本功能需求,还应是一个融数据、技术、人员、规则于一体的共享生态体系。

## 4.2 地球大数据云服务平台的探索

基于对上述“FAIR化、融合化、智能化和生态化”的趋势把握,笔者所在团队依托中国科学院战略性先导科技专项“地球大数据科学工程”(CASEarth),研发和建设地球大数据云服务平台(以下简称“平台”)。

FAIR化方面,平台为数据分配永久唯一标识DOI,且提供统一的元数据规范。针对不同数据资源类型,提供不同的汇聚和服务方式,例如对具有广泛共享需求的基础产品数据集进行集中存储,对其它数据及系统进行联邦式汇聚,整体对外提供统一的数据发现和访问服务,从而实现数据的可发现、可访问、互操作和可重用。

平台融合的探索不仅局限于数据,而是扩展至数据汇聚、存储、共享、计算与分析功能于一体<sup>[60-61]</sup>,具体包括支持数据资源汇聚、存储与共享的“地球大数据资源库”,海量时空数据统一管理的“格网数据引擎”,海量多源数据实时计算的“地球大数据计算引擎”,以及交互式数据分析引擎“EarthDataMiner”等。其中,“地球大数据资源库”的数据汇交子系统主要服务于专项内外的数据汇交,具备大文件切片快速上传、数据永久唯一标识管理、元数据管理、数据出版服务等核心工程;数据共享服务子系统是专项对外进行数据资源发布与共享的门户,提供数据发现与访问、关联与推荐、在线浏览与展示、下载与评价等服务。“格网数据引擎”则面向海量时空格网数据提取与计算分析需求,从时间、空间、频谱、专题等多个维度进行格网数据剖分存储管理,能实现PB级格网数据的高效组织和分散存储。而“EarthDataMiner”则在“格网数据引擎”和“地球大数据计算引擎”等的支持下,通过集成的各种数据分析模型和算法,为使用者提供所见即所得的交互式数据分析云服务。

智能化探索方面,平台强调机器自动发现和使用数据的能力,通过对数据资源和元数据的规范管理,提升了机器可操作性,不仅能精确查找和发现数据,且能推荐相关数据资源。同时,积极引入区块链、机

器学习与人工智能等技术,以提升数据在共享和计算等方面的智能性。

在此基础上,平台致力于构建完整的地球大数据生态系统:不仅为专项科学数据各环节和全流程提供数据管理,而且将数据提供者、数据使用者、服务提供者、应用开发者等通过数据资源紧密地联系起来。将平台建设运行贯穿数据全生命周期,以挖掘地球大数据的价值,实现数据生态系统的进化。

## 5 结语

科学数据共享系统是多种治理思维、共享模式和技术应用的复杂综合体,依政策导向、技术水平、学科领域、组织机构等要素而千差万别。多样的共享系统运行意味着不同的驱动机制、权责关系以及不同的管理方式与质量绩效等。本文结合系统所服务的科学数据共享模式以及系统本身的技术特色与服务内容,将科学数据共享系统归纳为仓储型系统、联邦服务型系统、数据分发型系统、按需计算与分析的云服务型系统四种类型。不可否认,此分类在一定程度上存在交叉重叠,但这并不妨碍我们沿着科学共同体的既有框架、主流趋势循序渐进式的温变革。万变不离其宗的科学数据全生命周期仍将是未来科学数据系统依附载体,在功能和性能的权衡天平之下,资源的FAIR化和云化、服务集成化与智能化、生态综合体系化仍将是未来的主流趋势。

## 参考文献

- [1] 国家科技基础条件平台中心. 国家科学数据资源发展报告(2017)[M]. 北京: 科学技术文献出版社, 2018: 34.  
National Science and Technology Infrastructure Center. National Scientific Data Resource Development Report(2017)[M]. Beijing: Scientific and Technology Documentation Press, 2018:34.
- [2] 黎建辉, 吴超, 张丽丽, 等. 科学数据出版调查与分析[J]. 中国科学数据, 2016, 1(1):64-74.  
Li J H, Wu C, Zhang L L, et al. Survey and Analysis of Scientific Data Publishing[J]. China Scientific Data, 2016, 1(1): 67-74.
- [3] Christine L, Borgman. The Conundrum of Sharing Research Data[J]. Journal of the American Society for Information Science and Technology, 2012, 63(6):1059-1078.
- [4] 黎建辉, 李跃鹏, 王华进, 等. 科学大数据管理技术与系统[J]. 中国科学院院刊, 2018, 33(8):796-803.  
Li J H, Li Y P, Wang H J, et al. Scientific Big Data Management Technique and System[J]. Bulletin of Chinese Academy of Sciences, 2018, 33(8):796-803.
- [5] 黎建辉, 沈志宏, 孟小峰. 科学大数据管理:概念、技术与系统[J]. 计算机研究与发展, 2017, 54(2):235-247.  
Li J H, Shen Z H, Meng X F. Scientific Big Data Management: Concepts, Technologies and System[J]. Journal of Computer Research and Development, 2017, 54(2):235-247.
- [6] Grossman R. Data Lakes, Clouds, and Commons: A Review of Platforms for Analyzing and Sharing Genomic Data[J]. Trends in Genetics, 2019, 35(3):223-234.
- [7] 张丽丽. 科学数据共享治理:模式选择与情景分析[J]. 中国图书馆学报, 2017(2):54-65.  
Zhang L L. Scientific Data Sharing Governance: Model Selection and Scenario Analysis[J]. Journal of Library Science in China, 2017(2):54-65.
- [8] 李成赞, 张丽丽, 侯艳飞, 等. 科学大数据开放共享:模式与机制[J]. 情报理论与实践, 2017, 40(11):45-51.  
Li C Z, Zhang L L, Hou Y F, et al. Scientific Big Data Opening and Sharing: Models and Mechanisms[J]. Information Studies: Theory & Application, 2017, 40(11):45-51.
- [9] 张丽丽, 黎建辉. 科研数据的开放:进展、模式与新探索[J]. 大数据, 2016(6):25-33.  
Zhang LL, Li J H. Research Data Openness: Development, Models and New Exploration[J]. Big Data Research, 2016(6): 25-33.
- [10] Figshare Features [EB/OL]. [2020-02-10]. <https://figshare.com/features>.
- [11] Hahnel M. Exclusive: figshare a new open data project that wants to change the future of scholarly publishing[EB/OL]. [2020-02-10]. <https://blogs.lse.ac.uk/impactofsocialsciences/2012/01/18/can-we-do-better-with-scholarly-publishing/>.
- [12] Dryad[EB/OL]. [2020-02-10]. <https://datadryad.org/>.
- [13] Vision T. The Dryad Digital Repository: Published evolutionary data as part of the greater data ecosystem[J]. Nature Precedings, 2010: 1-1.
- [14] ScienceDB[EB/OL]. [2020-02-10]. <http://www.scidb.cn/>.
- [15] DataONE[EB/OL]. [2020-02-10]. <https://www.dataone.org/>.
- [16] Michener W, Vieglais D, Vision T, et al. DataONE: Data Observation Network for Earth—Preserving Data and Enabling Innovation in the Biological and Environmental Sciences[J]. D-Lib Magazine, 2011, 17(1/2): 12.
- [17] Michener W K, Allard S, Budden A, et al. Participatory design of DataONE—enabling cyber infrastructure for the biological and environmental sciences[J]. Ecological Informatics, 2012, 11:5-15.
- [18] Christian E J. GEOSS Architecture Principles and the

- GEOSS Clearinghouse[J]. IEEE Systems Journal, 2008, 2(3): 333-337.
- [19] Group on Earth Observations. GEOSS 10-Year Implementation Plan[EB/OL]. [2020-02-10]. <http://www.earthobservations.org/documents/10-Year%20Implementation%20Plan.pdf>.
- [20] Bai Y, Di L, Nebert D D, et al. GEOSS Component and Service Registry: Design, Implementation and Lessons Learned[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2012, 5(6): 1678-1686.
- [21] 中国科学院数据云[EB/OL]. [2020-02-10]. <http://www.csdb.cn/>.
- Data Cloud of CAS[EB/OL]. [2020-02-10]. <http://www.csdb.cn/>.
- [22] 黎建辉, 周园春, 胡良霖, 等. 中国科学院科学数据云建设与服务[J]. 大数据, 2016, 2(6):3-13.
- Li J H, Zhou Y C, Hu L L, et al. Construction and service of scientific data cloud of Chinese Academy of Sciences[J]. Big Data Research, 2016, 2(6):3-13.
- [23] Landsat Data Access[EB/OL]. [2020-02-10]. <https://earthexplorer.usgs.gov/>.
- [24] Irons J R, Dwyer J L, Barsi J A. The Next Landsat Satellite: The Landsat Data Continuity Mission[J]. Remote Sensing of Environment, 2012, 122: 11-21.
- [25] ESA. Sentinel Online Data Access[EB/OL]. [2020-02-10]. <https://sentinels.copernicus.eu/web/sentinel/sentinel-data-access>.
- [26] Torres R, Snoei P, Geudtner D, et al. GMES Sentinel-1 Mission[J]. Remote Sensing of Environment, 2012, 120: 9-24.
- [27] LAMOST Data Access[EB/OL]. [2020-02-10]. <http://dr.lamost.org/data-release.html>.
- [28] Zhao G, Zhao Y H, Chu Y Q, et al. LAMOST Spectral Survey—An Overview[J]. Research in Astronomy and Astrophysics, 2012, 12(7):723.
- [29] Luo A L, Zhang H T, Zhao Y H, et al. Data release of the LAMOST pilot survey[J]. Research in Astronomy and Astrophysics, 2012, 12(9):1243.
- [30] SDSS Data Access[EB/OL]. [2020-02-10]. <http://skyserver.sdss.org/>.
- [31] Eisenstein D J, Weinberg D H, Agol E, et al. SDSS-III: Massive Spectroscopic Surveys of the Distant Universe, the Milky Way, and Extra-solar Planetary Systems[J]. The Astronomical Journal, 2011, 142(3): 72.
- [32] 地理空间数据云[EB/OL]. [2020-02-10]. <http://www.gscloud.cn/>.
- Geospatial Data Cloud[EB/OL]. [2020-02-10]. <http://www.gscloud.cn/>.
- [33] 国家生态系统观测研究网络[EB/OL]. [2020-02-10]. <http://www.cnern.org/>.
- National Ecosystem Research Network of China[EB/OL]. [2020-02-10]. <http://www.cnern.org/>.
- [34] 中国生态系统网络[EB/OL]. [2020-02-10]. <http://www.cern.ac.cn/>.
- Chinese Ecosystem Research Network[EB/OL]. [2020-02-10]. <http://www.cern.ac.cn/>.
- [35] 郭学兵, 苏文, 唐新斋, 等. 云计算环境下CNERN资源管理与服务平台的构建[J]. 中国科技资源导刊, 2017, 49(1): 30-37.
- Guo X B, Su W, Tang X Z, et al. Construction of CNERN Resource Management and Service Platform in Cloud Computing Environment[J]. China Science & Technology Resources Review, 2017, 49(1):30-37.
- [36] Lewis A, Oliver S, Lymburner L, et al. The Australian geoscience data cube—foundations and lessons learned[J]. Remote Sensing of Environment, 2017, 202: 276-292.
- [37] Kopp S, Becker P, Doshi A, et al. Achieving the Full Vision of Earth Observation Data Cubes[C]. International Conference on Data Technologies and Applications, 2019, 4(3):94.
- [38] 柏永青, 杨雅萍, 孙九林. 国内外科学数据管理办法研究进展[J]. 农业大数据学报, 2019, 1(3):5-20.
- Bai Y Q, Yang Y P, Sun J L. Advance in the Study of Domestic and Foreign Data Management Methods[J]. Journal of Agricultural Big Data, 2019, 1(3):5-20.
- [39] 温亮明, 张丽丽, 黎建辉. 大数据时代科学数据共享伦理问题研究[J]. 情报资料工作, 2019, 40(2):38-44.
- Wen L M, Zhang L L, Li J H. Research on Ethical Issues of Scientific Data Sharing in the Big Data Era[J]. Information and Documentation Services, 2019, 40(2):38-44.
- [40] Bica M, Bacu V, Mihon D, et al. Architectural Solution for Virtualized Processing of Big Earth Data[C]. IEEE International Conference on Intelligent Computer Communication & Processing, 2014.DOI:10.1109/ICCP.2014.6937027.
- [41] 程学旗, 靳小龙, 王元卓, 等. 大数据系统和分析技术综述[J]. 软件学报, 2014, 25(9):1889-1908.
- Cheng X Q, Jin X L, Wang Y Z, et al. Survey on Big Data System and Analytic Technology[J]. Journal of Software, 2014, 25(9):1889-1908.
- [42] Killough B. Overview of the Open Data Cube Initiative[C]. 2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2018), 2018:8629-8632.
- [43] Karpatne A, Liess S. A Guide to Earth Science Data: Summary and Research Challenges[J]. Computing in Science and Engineering, 2015, 17(6):14-18.

- [44] Camara G, De Assis L F, Ribeiro G, et al. Big Earth Observation Data Analytics: Matching Requirements to System Architectures[C]. International Workshop on Analytics for Big Geospatial Data, 2016:1-6.
- [45] Guo H D, Liu Z, Jiang H, et al. Big Earth Data:A New Challenge and Opportunity for Digital Earth's Development[J]. International Journal of Digital Earth, 2017, 10(1):1-12.
- [46] Fouad K, Bixby J L, Callahan A, et al. FAIR SCI Ahead: The Evolution of the Open Data Commons for Preclinical Spinal Cord Injury Research[J]. Journal of Neurotrauma, 2019, <https://doi.org/10.1089/neu.2019.6674>.
- [47] 张丽丽, 温亮明, 石蕾, 等. 国内外科学数据管理与开放共享的最新进展[J]. 中国科学院院刊, 2018, 33(8):774-782. Zhang L L, Wen L M, Shi L, et al. Progress in Scientific Data Management and sharing[J]. Bulletin of Chinese Academy of Sciences, 2018, 33(8):774-782.
- [48] 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):160018.
- [49] Wilkinson M D, Sansone S, Schultes E, et al. A Design Framework and Exemplar Metrics for FAIRness[J]. Scientific Data, 2018, 5(1):180118.
- [50] Mons B, Schultes E, Liu F H, et al. The FAIR Principles: First Generation Implementation Choices and Challenges [J]. Data Intelligence, 2019, 2(1/2):1-9.
- [51] Aiello G, Giovino I, Vallone M, et al. A Decision Support System Based on Multisensor Data Fusion for Sustainable Greenhouse Management[J]. Journal of Cleaner Production, 2018: 4057-4065.
- [52] Stadnikia K, Henderson K, Koppal S J, et al. Data Fusion for a Vision-aided Radiological Detection System: Correlation Methods for Single Source Tracking[J]. Nuclear Instruments & Methods in Physics Research Section A-accelerators Spectrometers Detectors and Associated Equipment, 2019. DOI:10.1016/j.nima.2019.02.040.
- [53] Khankalantary S, Rafatnia S, Mohammadkhani H, et al. An Adaptive Constrained Type-2 Fuzzy Hammerstein Neural Network Data Fusion Scheme for Low-cost SINS/GNSS Navigation System[J]. Applied Soft Computing, 2020. DOI: 10.1016/j.asoc.2019.105917.
- [54] Lnenicka M, Komarkova J. Big and Open Linked Data Analytics Ecosystem:Theoretical Background and Essential Elements[J]. Government Information Quarterly, 2019, 36(1): 129-144.
- [55] Weigel T, Schwarzmänn U, Klump J, et al. Making Data and Workflows Findable for Machines[J]. Data Intelligence, 2019, 2(1/2):30-39.
- [56] Wen L M, Zhang L L, Li J H. Application of Blockchain Technology in Data Management:Advantages and Solutions [J]. Lecture Notes in Computer Science, 2019(11473): 239-254.
- [57] Martin S, Turki S, Renault S. Open Data Ecosystems[C]. Electronic Government and the Information Systems Perspective: 6th International Conference, 2017: 49-63. DOI: 10.1007/978-3-319-64248-2\_5.
- [58] Zuiderwijk A, Janssen M, Davis C, et al. Innovation with Open Data:Essential Elements of Open Data Ecosystems[J]. Information polity, 2014: 17-33.
- [59] Marcelo Iury S Oliveira , Bernadette Farias Lóscio. What is a Data Ecosystem[C]. Proceedings of the 19th Annual International Conference on Digital Government Research, 2018: 1-9. <https://doi.org/10.1145/3209281.3209335>.
- [60] 郭华东. 地球大数据科学工程[J]. 中国科学院院刊, 2018, 33(8):818-824. Guo H D. A project on Big Earth Data Science Engineering [J]. Bulletin of Chinese Academy of Sciences, 2018, 33(8): 818-824.
- [61] Yang C W, Yu M Z, Li Y, et al. Big Earth Data Analytics: A Survey[J]. Big Earth Data, 2019, 3, (2):83-107.