Doi:10.11840/j.issn.1001-6392.2019.04.007

# 海洋地质地球物理数据分类与组织

刘志杰，孔敏，舒雨婷，王风帆，韩璐遥，田先德
（国家海洋信息中心，天津 300171）

**摘 要**：信息技术的快速变革，有效驱动海洋信息化的发展与创新。实现海洋信息资源的高效存储和管理，需要对海洋数据管理体系进行统筹规划。文章根据海洋地质地球物理数据特点，进行了数据层次分类研究，分析了数据标准化过程，确立了基于文件和数据库的海洋地质地球物理数据存储策略，提出结构化事务型数据库、结构化分析型数据库（NewSQL）和非结构化计算型数据库（NoSQL）相结合的数据组织架构。该数据分类组织方案为实现海洋地质地球物理数据规范化管理、集约利用和共享服务奠定了良好基础，也为其他海洋数据分类与组织研究提供了参考。

**关键词**：海洋地质地球物理；数据组织；数据分类；数据库
**中图分类号**：P628+.4 **文献标识码**：A **文章编号**：1001-6932(2019)04-0415-07

# Classification and organization of marine geological and geophysical data

LIU Zhi-jie, KONG Min, SHU Yu-ting, WANG Feng-fan, HAN Lu-yao, TIAN Xian-de
(National Marine Data and Information Service, Tianjin 300171, China)

**Abstract**: The fast development of information technology is an effective drive for the development and innovation of marine informatization. In order to realize efficient storage and management of marine information resources, marine data management system needs to be designed and establised in an integrated way. Based on the features of marine geological and geophysical data, their hierarchical classification methods were studied. The organization and management of marine geological and geophysical data were analyzed. On the basis of file and database, marine geological and geophysical data archiving strategies were established. And the data organization structure which is combined by structural transactional database, structural analytical database (NewSQL) and non-structured computational database (NoSQL) was developed. This research lays a good foundation for the standardization of management and service of marine geological and geophysical data, and could provide references for studies on other marine data classification and management.

**Keywords**: marine geology and geophysics ; data organization ; data classification ; database

进入 21 世纪，以信息技术为代表的新一轮科技革命方兴未艾，世界各国均以信息化为抓手，积极谋求掌握发展主动权（国家信息化发展战略纲要，2016）。数据强国、数据治国已成为我国大数据国家战略。海洋信息事业发展迎来机遇期，同时也面临诸多挑战。数据是海洋信息化实施的重要基础，数据资源整合和高效利用已成为海洋信息化发展必然趋势。近年来，海洋高新技术的应用极大促进了海洋地质科学的发展（王蛟 等，2015）。通过一系列国家重大专项及科研计划的实施，海洋调查频度增加，我国已在海岛海岸带、中国近海、大陆架边缘海和深海大洋开展了不同程度的海洋地质地球物理调查与勘测，取得了大量调查成果，积累了海量数据。随着海洋地质地球物理调查技术手段多样化，分析测试方法增加，形成的数据种类和要素也日渐繁多。这些数据来源广泛、覆盖范围广、时

收稿日期：2018-12-11；修订日期：2019-03-25
基金项目：全球变化与海气相互作用专项（GASI-01-01-09-04）。
作者简介：刘志杰（1977-），博士研究生，高级工程师，主要从事海洋地质地球物理数据管理和信息化工作。电子邮箱：kittylzj@163.com。

间跨度大、格式多样，数据快速增长与有效充分利用之间的矛盾日益凸显。如何解决数据多源异构、分类不统一、组织方式多样等问题，实现数据的有效管理和高效利用，是当前数据管理者亟须考虑的事情。海洋地质地球物理数据分类与组织是开展数据体系和数据资源建设的一项重要内容，其目的是使海量、异构分布的数据资源实现有序化（文峰，2013）。文章基于海洋地质地球物理数据特点，以数据库建设实践为基础，厘清数据逻辑结构关系，开展数据分类与组织研究，以期提高数据管理效能和服务水平。

## 1 数据分类

数据分类是根据数据自身属性和特征，依据一定的原则和方法进行归类，并建立起一定的分类体系，是数据处理和应用的重要环节和前提（邹国良等，2016）。海洋地质地球物理数据涉及现场采集、室内测试分析、处理解释和研究等环节。数据分类一般先根据流转和加工程度划分，然后再根据数据类型进行细分。以调查航次或区块为单元，海洋地质地球物理数据根据数据流转和加工程度，分为任务文档类、原始数据集、整编数据集、标准数据集、报告专著类、图件图集类、图像摄像类、软件类和其他类（表 1），其中既有结构化数据又有半结构化和非结构化数据，构成了整个海洋地质地球物理数据体系。

任务文档类是指调查航次前有关文档材料，主要包括：任务合同书、实施方案、航次报告、航次计划以及其他相关材料，可为数据审核处理提供参考。

原始数据集是指是由仪器自动采集获取的、未经过加工处理的数据和各种原始记录文件集合，主要包括：沉积物现场采样描述记录、导航数据、现场原位测试数据、海洋地球物理原始数据、悬浮体现场测试数据和仪器配置参数等。

整编数据集是根据《海底底质资料整编技术规程》（国家海洋局海洋科学技术司，2012）和《海洋地球物理资料整编技术规程》（国家海洋局海洋科学技术司，2012）有关要求，对海底底质室内样品分析测试数据和地球物理后处理成果数据进行整理分析形成。一般样品室内分析测试整编数据以 Excel 形式存储，仪器获取处理成果数据以文本文件形式存储。海底底质数据按样品类型又可分为沉积物、悬浮体和岩石，根据分析测试项目不同，沉积物数据集又分为沉积物粒度、沉积物化学、沉积物矿物、古生物、工程物理力学性质等近 10 多种小类；悬浮体数据主要测试项目包含浊度、浓度、现场激光粒度和颗粒有机碳氮等类型；岩石测试数据集主要包括化学测试和矿物测试等数据。海洋地球物理整编数据集根据调查手段分为海洋重力、海洋磁力、海洋地震、浅地层剖面、海洋电磁和海底热流等，每一类根据观测仪器或方式不同可以细分为不同小类（图 1）。

标准数据集是指整编数据集经代码转换、记录

**表 1 海洋地质球物理数据种类划分**

| 资料种类 | 英文名称 | 说明 | 格式要求 |
|---|---|---|---|
| 任务文档类 | Documents | 任务合同<br>任务实施方案<br>航次报告<br>航次计划等 | Word 或 PDF |
| 原始数据集 | Raw dataset | 现场原位测试及调查仪器自记观测数据的集合，含参数和格式说明 | 仪器导出格式，不固定 |
| 整编数据集 | Complicating dataset | 根据整编技术规范要求进行整理的分析测试及后处理数据的集合，含数据处理与质量评价报告 | Excel 或 TXT 文本 |
| 标准数据集 | Standard dataset | 经代码和记录格式转换、质量控制、排重合并等处理后形成的标准数据文件的集合 | TXT 文本 |
| 报告专著类 | Report & article | 研究报告、专著、发表论文等 | Word 或 PDF |
| 图件图集类 | Atlas | 编制的图件或图集（矢量文件） | 矢量文件或 PDF |
| 图像摄像类 | Image & camera | 由专业设备获取的摄像视像及图片等 | Jpg，Img，Mp4，Avi 等 |
| 软件类 | Software | 配套软件及说明 | |
| 其他类 | Others | 除上述范围内的其他数据 | |

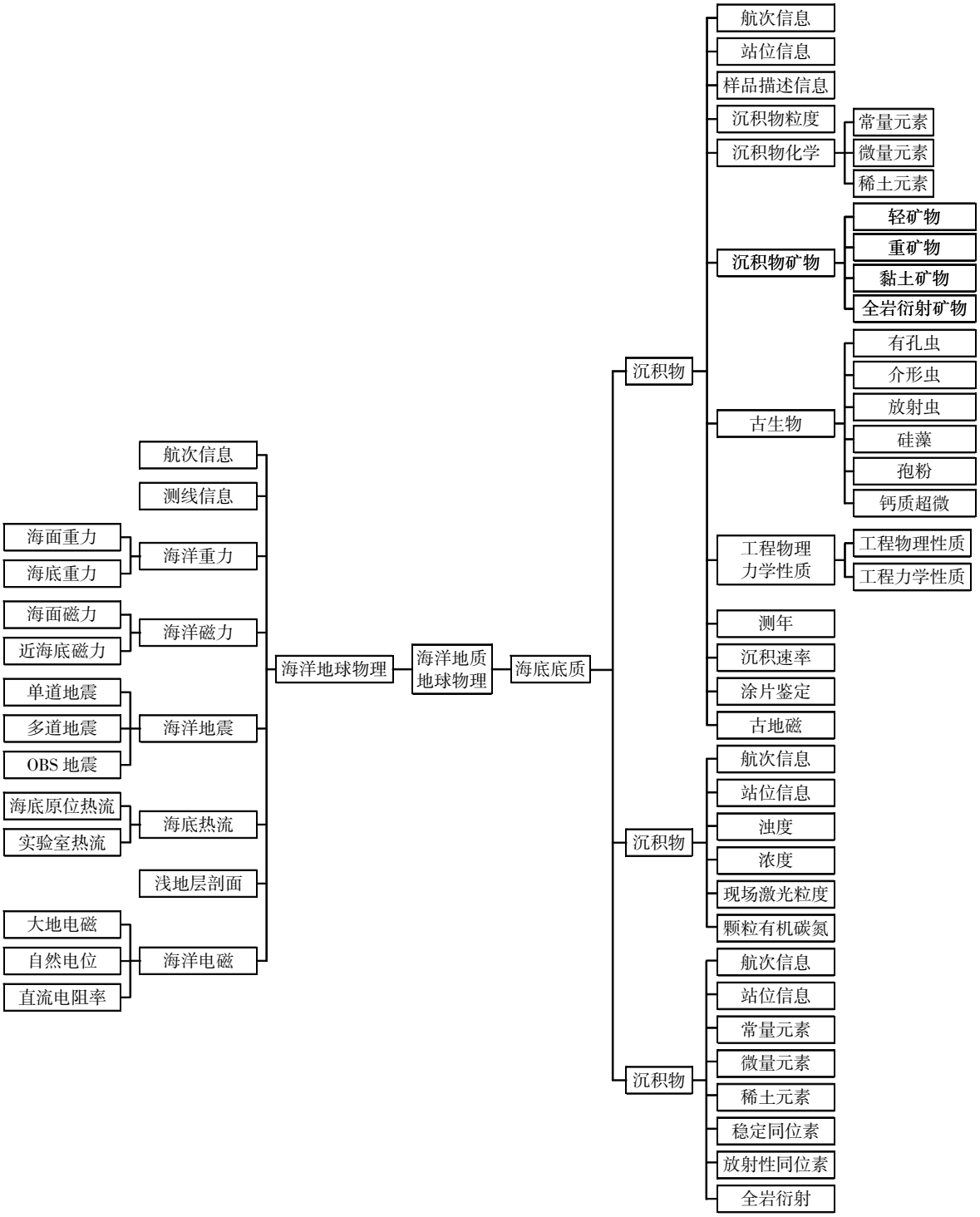


**图 1　海洋地质地球物理主要数据类型划分**

格式转换、质量控制、排重合并等处理后形成的标准数据文件的集合，因格式相对固定和统一，亦可作为交换数据集或入库数据集。标准数据集的制作是数据标准化的一项重要内容，所包含的数据类型同整编数据集，也是海洋地质地球物理基础数据库建设的重要内容。除此之外，还应包括相关元数据信息。

项目成果除数据集之外，还有研究报告专著类、图件图集类、图像摄像类和软件类等等。其中研究报告专著类是项目基于数据分析的成果的总结；图

件图集类是海洋地质地球物理成果的直观表达，包括沉积物类型图、各底质要素分布图、重力异常图、磁力异常图以及沉积物厚度图等；图像摄像类主要在海洋调查过程中产生的，以视频摄像、图片等形式存储的数据；软件类成果主要产生于系统研发项目，为海洋地质地球物理数据处理、管理和服务提供技术支撑。

## 2 数据标准化

不同单位、不同时期、不同项目来源数据格式、数据语义表达、组织结构都不统一，甚至项目间采用的数据处理参数、处理标准和方法也不完全一致（刘志杰 等，2013），这就导致同一区域不同来源数据难以综合使用，数据价值得不到充分发挥。因此，数据标准化是进行数据规范化和有效组织的重要前提。

海洋地质地球物理数据标准化是将调查或分析得到的分析整编数据，按照一定格式要求进行转换处理和质量控制，前提是不能改变数据原始属性信息，须遵循完整性、一致性和真实性原则。根据数据获取方式，分为两类：一类是取得样品后经室内分析测试得到的数据，如沉积物粒度、碎屑矿物、常微量元素、黏土矿物等；另一类是调查仪器自动获取得到的数据，如海洋重力、海洋磁力、海洋地震、浅地层剖面和悬浮体浊度数据等。数据标准化处理需要采用相应专业数据处理系统按照规范的处理流程开展。

### 2.1 文件命名规范化

数据文件规范化命名须采用统一的命名规则。海底底质数据文件名称由“项目编码”+“区块编码”+“数据类型”三部分构成。当一个区块由多个执行航次时，不同航次数据文件按照区块进行合并后再进行规范化命名。例如：123CJ01表层沉积物粒度.xls，其中“123”代表是项目编码，“CJ01”为区块编码，“表层沉积物粒度”代表数据类型。海洋地球物理数据则通常以测线进行文件划分，即文件名称反映测线名称，如“line001.txt”，代表一个航次或区块的某一条测线，而项目编码、区块编码和数据类型以文件夹名称体现。

### 2.2 数据完整性检查

数据完整性是指数据实体字段属性的存在或缺失程度。根据《海底底质资料整编技术规程》（国家海洋局海洋科学技术司，2012）和《海洋地球物理资料整编技术规程》（国家海洋局海洋科学技术司，2012）中格式要求，通过处理系统字段匹配功能，检查数据项内容是否完整。对于数据项内容不完整的文件，通过查找原始数据和资料处理报告等原始记录将缺失信息补充完善。数据文件中的关键字段信息如经纬度信息不能缺失，关键信息的缺失会导致数据无效。

### 2.3 数据代码与计量单位统一

标准化数据集中通常会涉及项目、区块、航次、取样方式和分析单位等公共字段的编码。不同项目产生的数据可能存在编码不一致的情况，为了便于数据的统一和规范化管理，在数据标准化处理时需按照统一编码标准进行代码转换和修正。此外，对同一类型数据同一要素的计量单位要转换成国家统一要求的标准计量单位，以确保不同来源数据计量标准的一致性。

### 2.4 数据格式转换

海洋地质地球物理数据来源不同，格式多样，实现异构数据格式转换，可使数据利用更加高效、便捷，有效促进数据交换共享。通过数据处理系统，对源文件进行解析，将原来不规则的源文件格式转成统一的、规范的文件。数据转换流程首先选择所要处理的数据类型，然后进行表头字段配置和转换规则设置，最后输出具有固定格式的标准数据文件。海底底质以及重力、磁力处理后标准数据为文本格式（国家海洋局，2008），字段具有相对固定位置，便于数据库的读取和调用，也为应用系统平台提供统一标准数据接口。海洋地震和浅地层剖面等海底声学标准数据为经后处理形成特定格式的剖面数据，通过提取测线进行文件管理。

### 2.5 数据质量控制

为保证数据的可靠性，海洋地质地球物理数据一般采用人工审查和计算机质量控制相结合的方式。在数据汇交、预处理阶段以及数据处理后的抽检一般采用人工审查方式。对于格式转换后的标准化数据集，一般采用专业质控软件进行处理。计算机质量控制方法常包括以下几种：①站位一致性检验：数据中的站位信息须与站位表中信息相对应；②数值范围检验：根据经验值，定义要素的取值变化范围，对数据进行检验；③着陆点检验：查验数

据空间分布是否在所调查区域内；④逻辑一致性检验：根据数据项之间的逻辑关系进行判别，如沉积物柱状样层位关系，下层层位不能小于上层层位；⑤数据统计检验：根据数据统计特性进行数值检验，如沉积物粒度各粒级百分含量之和应为100%；⑥散点图检验：通过绘制散点图，发现离群值，进一步消除噪点数据或发现潜在的、有价值的海洋环境现象（向先全 等，2015）。

## 3　数据组织

数据组织就是按照一定方式和规则对数据进行分类、处理和存储的过程。良好的数据组织策略可使数据的管理和应用起到事半功倍的效果。数据的组织架构需要充分考虑数据本身特点，根据管理和应用需求分层设计。海洋地质地球物理不仅数据量大、时空尺度广、资料种类繁多，而且存储格式多样，包括诸如文本文档、表格数据、图形图像、二进制文件和矢量数据等。通常针对不同种类数据，采用不同组织方式。比如原始数据集一般需要专业的处理系统才能解析，通常以文件方式进行存储管理，保证文件能够通过元数据导航索引到即可。而对于结构化的标准数据集，需要建立相应的库表结构，采用关系型数据库形式进行存储和管理。

### 3.1　基于文件的数据组织

基于文件的数据组织方式是最常用的一种，通常以文件名和文件夹名称做标识，适用于各类数据的组织管理和备份存档。根据管理需求或目的不同，可以遵循不同逻辑关系，以文件为管理对象，采用层次法对文件进行组织。由于海洋地质地球物理调查或研究常以项目为主线开展，不同项目来源是数据文件组织的首要考虑因素。海洋地质地球物理数据主要来自国家组织实施的海洋专项调查、双多边合作调查、科技部重大专项、自然基金项目调查和国际共享交换数据等。因此，在数据资源汇集管理阶段遵循项目层次的管理方式，对数据对象进行抽象和分类。一个项目首先按照学科进行区分，每个学科由 n 个不同的航次调查任务组成，任务所产生的数据实体之间是有联系的（郭明航 等，2009），因此，一个任务产生的多个数据实体可先以航次或区块作为依据进行划分。同一航次前提下，再依据数据种类进行划分，比如一个航次任务获取的文件类通常包括：任务文档类、原始数据集、整编数据集、成果数据集、报告专著类和图集图件类等。文件组织可以根据任务取得成果种类进行扩充，每一文件类包含了若干属性相似的文件（图 2），一般存放于文件服务器中，并通过索引表与数据文件进行关联。

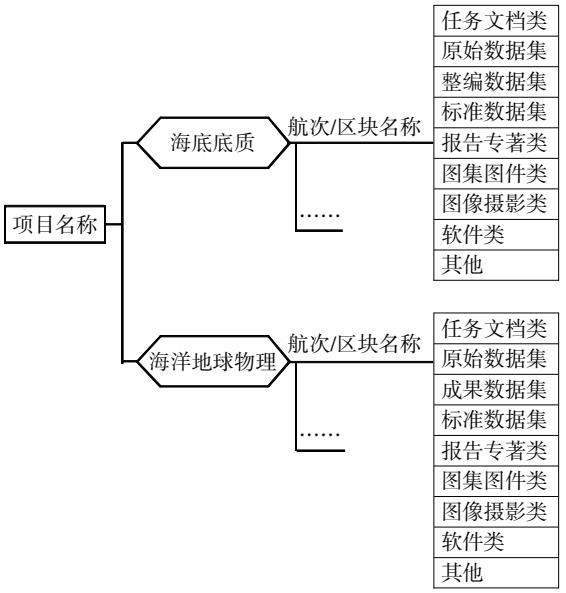


**图 2　数据文件目录组织结构**

文件的组织管理较为便捷，可随着调查仪器更新，调查技术的进步，动态进行资料类型的扩充，但前提是一定要有规范的文件命名规则、正确的分类体系和文件组织结构，否则数据文件的查询检索和应用会很烦琐，达不到高效管理的目的。

### 3.2　基于数据库的数据组织

数据库是海洋数据资产化管理的有效手段（宋德瑞 等，2017），重点解决多元异构数据的一体化组织和管理（黄少芳，2016）。数据库技术自 20 世纪 80 年代就被应用于海洋信息化建设（何广顺，2008），至今已建立了一系列满足不同层面需求的海洋综合数据库和海洋专题数据库（张峰 等，2009；苏国辉 等，2003；崔爱菊 等，2015）。随着大数据和云计算时代的到来，数据库技术也经历了一系列变革。过去数据库建设主要采用“一种架构支持多类应用”的传统模式，不能很好地满足统计分析、共享服务等复杂应用的需求（胡伟忠，2005）。随着海洋地质地球物理资料种类日益增多，

数据规模增大，除传统的表格和文本数据外，还有大量的视频摄像以及各种影像文件，数据的多元化需要根据不同的应用需求来选择多技术架构下的数据管理模式（Wu et al，2006）。因此，设计“多种架构支持多类应用”的数据组织策略，满足海量结构化和非结构化数据的存储管理、复杂分析、关联查询和实时性处理等多方面的要求，是解决目前数据管理和服务瓶颈的有效途径（宋晓 等，2018）。根据海洋地质地球物理数据量大、数据类型多、数据结构复杂的特点，采用结构化事务型数据库、结构化分析型数据库（NewSQL）和非结构化计算型数据库（NoSQL）相结合的数据组织架构，实现优势互补，以满足面向海量海洋地质地球物理数据的存储管理和复杂应用需求。

3.2.1 海洋地质地球物理基础数据库

海洋地质地球物理基础数据库采用 Oracle 事务型数据库管理系统存储和管理经质量审核和标准化处理后形成的海洋地质地球物理数据及清单等事务性数据。按照调查专业和方法的不同，分为海底底质基础数据库和海洋地球物理基础数据库。通过数据对象属性和相互之间的关系，设定各表间的关联规则，确定数据库映射关系表。海底底质数据库根据数据类型又包含了若干数据库表，其中航次信息、站位信息与分析测试数据表之间通过关联字段一对多关联，从而保证每一类信息的完整性。海洋地球物理基础数据库用于存储经标准化处理后的海洋重力和海洋磁力成果数据，数据库表通过调查航次信息、测线信息与数据进行关联。结构化事务型数据库的建立主要面向日常数据管理、共享目录发布等基础性应用。

3.2.2 海洋地质地球物理综合数据库

海洋地质地球物理综合数据库依托 MPP 分布式并行数据库集群存储和管理以要素为主体的整合数据层，建设过程中选用南大通用数据技术股份有限公司自主研发的 Gbase 数据库管理系统，通过 ETL 方式进行数据抽取、清洗和再组织，构建形成以列为最小组成单元的海洋沉积物粒度数据库、沉积物化学数据库、沉积物碎屑矿物数据库、黏土矿物数据库、悬浮体浊度数据库、悬浮体浓度数据库以及海洋重力和磁力数据库等。结构化分析型数据库的建立为超大规模数据的统计分析和可视化展现等提供了架构基础，为实现各类海洋数据的分析研究和共享服务提供了可能。

3.2.3 海洋地质地球物理成果数据库

海洋地质地球物理成果数据库采用 Hadoop 框架实现非结构化海洋地质地球物理数据的存储和使用。Hadoop 是 Apache 开源组织的一个分布式计算框架，可以在大量廉价的硬件设备组成的集群上运行应用程序，构建一个大数据量、易扩展且具有高可靠性的灵活自由的并行分布式系统（崔杰 等，2012）。海洋地质地球物理成果数据库主要用于存储数据量大且繁杂的半结构化和非结构化数据文件，包括海底地震和浅剖数据以及报告、图集和摄像视像等，涉及的数据格式类型有 segy、doc、pdf、jpg、img、mp4 等 10 多种。非结构化计算型数据库具有易扩展、使用灵活的特性，在大数据量下的高性能读写以及数据模型构建方面表现出很大的优势，可以对各类海洋数据的关联分析、趋势预测和深度学习等大数据计算提供有力的结构基础，从而满足海量海洋地质地球物理数据知识挖掘等复杂应用需求。

## 4 结语

数据管理是基础，应用是根本，良好的数据组织方式至关重要。海洋地质地球物理数据分类与组织是海洋数据资源建设的一项基础性业务工作，其内容并非一成不变，会随着技术手段的提高不断完善和优化，以确保其先进性和适用性。本文基于海洋地质地球物理数据特点，研究了数据基础分类体系，介绍了数据标准化处理和质量控制过程，确立了基于文件和数据库的海洋地质地球物理数据存储策略，提出结构化事务型数据库、结构化分析型数据库（NewSQL）和非结构化计算型数据库（NoSQL）相结合的数据组织架构。该研究成果已应用到海洋地质地球物理业务化工作中，对于数据集中有效管理、数据库建设、信息资源整合和共享，具有十分重要的意义。

下一步应坚持需求导向，根据数据组织和分类特点，建立相关标准规范，进一步指导和规范海洋地质信息化建设。在此基础上，根据数据来源、精度和范围，开展数据分类定级研究，拓展数据共享范围，提高海洋地质地球物理信息资源利用率。充分利用大数据挖掘分析技术，形成高附加值的信息

产品，服务于海洋科学研究和经济国防建设。

## 参 考 文 献


Wu B, Kshemkalyani A D, 2006. Objective-optimal algorithms for long-term Web prefetching. IEEE Transactions on Computers. 55 (1): 2-17.

崔爱菊, 王建村, 苏天赟, 2015. 海洋地球物理数据库设计与实现. 海洋科学, 39(3): 116-121.

崔杰, 李陶深, 兰红星, 2012. 基于 Hadoop 的海量数据存储平台设计与开发. 计算机研究与发展, (49): 12-18.

郭明航, 田均良, 李军超, 2009. 地球科学研究数据的分类与组织研究. 水土保持研究, 16(4): 203-206.

国家海洋局, 2008. GB/T 12763.7-2007.海洋调查规范第 7 部分-海洋调查资料交换. 北京:中国标准出版社.

国家海洋局海洋科学技术司, 2012. 海底底质资料整编技术规程. 北京: 海洋出版社.

国家海洋局海洋科学技术司, 2012.海洋地球物理资料整编技术规程. 北京:海洋出版社.

国家信息化发展战略纲要, 2016-08-14. http://www. gov.cn/xinwen/2016-07/27/content_5095336.htm.

何广顺, 2008. 海洋信息化现状与主要任务. 海洋信息, (3): 1-4.

胡伟忠, 2005. 海量海洋数据一体化管理研究. 杭州: 浙江大学.

黄少芳. 基于地质大数据的地质资料信息化与标准化. 中国煤炭地质, 2016, 28(7): 74-78.

刘志杰, 公衍芬, 周松望, 等, 2013. 海洋沉积物粒度 3 种计算方法的对比研究. 海洋学报, 35(3): 179-188.

宋德瑞, 曹可, 张建丽, 等, 2017. 大数据视域下的海洋信息化建设构想. 海洋开发与管理, (9): 50-53.

宋晓, 梁建峰, 李维禄, 等, 2018. 基于多架构混搭模式的极地海洋数据库建模技术研究. 极地研究, 30(4): 412-420.

苏国辉, 戴勤奋, 魏合龙, 等, 2003. 海洋地质数据库数据的存储结构. 海洋地质动态, 19(6): 5-7.

王蛟, 莫杰, 2015. 高新技术促进海洋地学科技发展. 海洋技术学报, 34(2): 118-125.

文峰, 2013. 一种面向应用的多层次数据资源描述框架. 计算机应用与软件, 30(7): 221-223.

向先全, 路文海, 杨翼, 2015. 海洋环境监测数据集质量控制方法研究. 海洋开发与管理, (1): 88-91.

张峰, 石绥祥, 殷汝广, 等, 2009. 数字海洋中数据体系结构研究. 海洋通报, 28(4): 1-8.

邹国良, 韩金菊, 屠正飞, 等, 2016. 基于 BP 神经网络的海洋监测数据等级划分. 海洋通报, 35(2): 187-193.


（本文编辑：崔尚公）