地质数据共享交换机制研究

马 岩 汪永新 雷 震 牛亚慧 荆 莹
河南省地质局矿产资源勘查中心 河南 郑州 450006

摘 要：根据河南地矿业务数据现状，对地质数据共享交换机制进行研究，开展全局数据归集、存储、治理与安全等方面的研究，建立数据标准规范体系，通过规范的业务流程和标准化的数据定义，为地矿数据资源的统一管理提供保障，推动地矿数据共享和综合应用。因此，本文旨在探讨大数据环境下地质学的发展趋势，以及数据共享开放对地质学研究的重要性，并在此基础上分析如何对地质数据进行管理。

关键词：地质；标准；规范；治理

在21世纪的信息技术革命中，大数据已经成为了引领时代潮流的关键力量，它不仅在人类生活中产生了深远的影响，而且在全球范围内重塑了研究世界的思维方式。作为一种战略资源，大数据的研究和应用已经引起了世界各国的广泛关注，并成为推动国家科技进步的重要动力。特别是在地质学这一数据驱动的学科领域，随着地球信息探测技术的飞速进步，数据获取能力得到了显著提升，地球观测数据的积累更是呈现指数级的增长趋势。在这样的背景下，地质勘查和地学研究迫切需要实现数据的共享与开放，以促进科学研究的深度和广度，提高数据利用效率，推动地质科学的创新与发展。

1 研究背景

在全球范围内，大数据的兴起已经促使各国政府认识到数据共享的重要性，并在政策层面推动了相关措施的制定和实施。美国《能源政策法案》的出台，便是其中一个典型的例子。该法案明确指出，美国地质调查局（USGS）和内政部门所收集的地学资料和数据应当在全国范围内实现通用共享。为了具体实施这一政策，美国发布了一项名为国家地质与地球物理数据保存计划（NGGDPP）的行动计划，旨在由USGS负责实施、管理和监督，内政部下属的各联邦机构作为成员单位，共同负责数据的收集和保存工作。该计划的实施取得了显著成效，USGS和地质调查机构已经公开了超过200万条地质样品元数据信息，建立了一套全国统一的地学数据和标本目录。这一举措不仅极大地提高了美国地质资料的保管和服务质量，也为全球地学领域的研究提供了宝贵的数据资源。

近年来，各省份都在积极推动地质数据共享服务，我国的地质数据共享实践呈现“多点开花”的局面。

地质调查局作为国家地质调查工作的主要负责机构，于2021年6月，推出“地质云3.0”版本[1]，包含基础地质、能源矿产、水资源、土地资源、海洋地质、地下空间等11大类核心数据库，近百个专业数据库，为全社会提供了更为丰富和权威的地球科学数据信息服务。

2017年，云南省地矿局着手探索地质工作与现代新技术的融合，并提出建设云南地质大数据服务平台[2]。到了2019年，该平台建成并整合了包括基础地质数据、城市地质数据、基础地理信息数据、自然资源现状数据、国土空间规划数据、行业专项管理数据等多种数据资源，成功构建了云南地质大数据支撑体系。

2021年12月，湖北省地质大数据平台正式上线[3]，汇聚了湖北省地质局在多个地质领域的丰富数据和成果，通过一系列云服务应用，如数字地质资料中心、地质一张图等，为地质工作者和社会公众提供了便捷的地质信息查询和资源共享服务。

2 技术路线和工作方法

2.1 技术路线

为了丰富地质数据资源和服务产品，提升地质数据信息共享服务的能力和水平，采取以提升数据采集能力为前提，构建有效的地质数据汇聚体系；以建立地质数据与信息服务体系为核心，确保数据的高效利用；以大数据基础设施平台为坚实基础，保障数据处理的规模和效率；以制度标准建设和机制形成为关键保障，确保地质数据管理的规范性和共享服务的可持续性。通过这些措施，可以全面提升地质数据信息共享服务的能力和水平，为地质科学研究、资源勘查、环境保护和灾害防治等领域提供强有力的数据支持。

2.2 工作方法

2.2.1 采用调查法进行数据摸底的研究

通过实地调研，可以全面了解数据的基本情况，包括数据来源、类型、形态、模式、总量和增量等。大数据中心的建设不仅需确保现有原始数据的全量保存，还应根

据数据所涉及的对象和业务领域进行分类式的深层次设计和处理，以实现数据资源的最大化利用和价值提升。

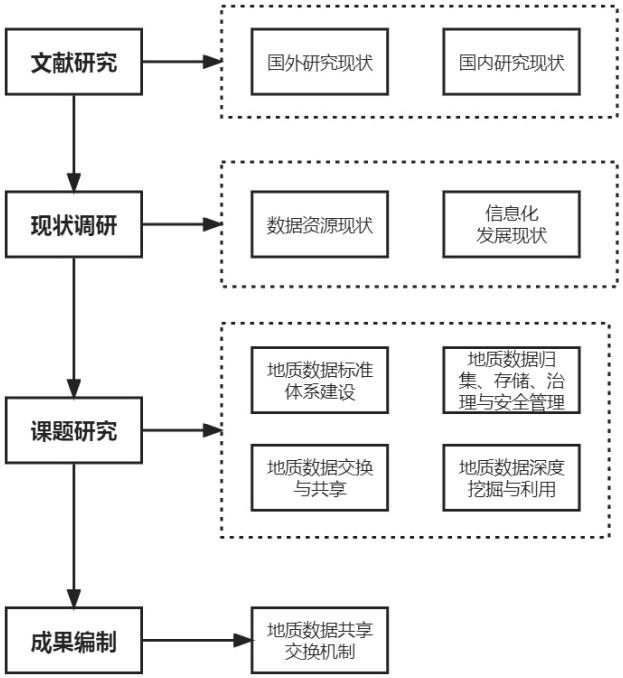

图1 技术路线图

2.2.2 采用经验法进行数据接入方式的研究

在确定接入的数据源后，研究历史数据的全量抽取与业务数据增量接入方式是确保数据有效整合的关键步骤。全局数据分散在各个业务系统中，因此在数据接入过程中，需要特别关注数据产权、数据模式冲突、异构数据兼容性等技术问题，以确保数据能够顺利集成和高效利用。

2.2.3 采用功能分析法进行数据应用、治理的研究

从技术视角来看，大数据中心为了支撑数据的全生命周期管理与应用，需要具备相对完善的数据管理、类目管理、流程编排、任务调度、数据溯源、数据治理、质量管理、权限管理等能力。在计算能力上，需要支持SQL和可编程的批处理两种模式（对机器学习的支持，可以采用Spark或者Flink的内置能力）；在处理范式上，可以采用基于有向无环图的工作流的模式，并提供集成开发环境。

2.3 总体架构

（1）数据能力平台。从数据的全生命周期视角，构建数据资源共享共用的大数据中心，实现数据从采集、治理、资产管理、分析到提供数据服务的全生命周期管理和服务能力。

（2）数据资源中心。包括数据来源、贴源数据存储、数据资源库、数据仓库以及数据应用。

（3）数据运营平台。充分运用“互联网+”和大数据思维，通过数据资产运营监控、数据运营不断挖掘并提升数据资产的价值。

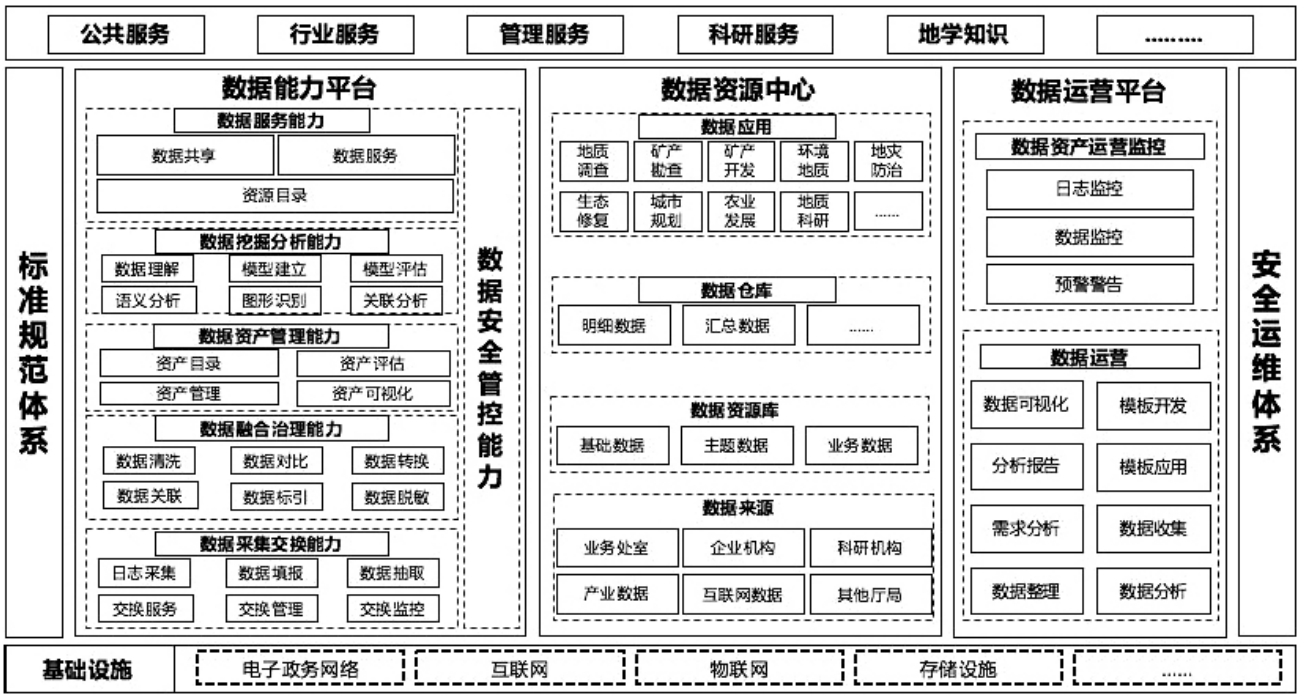

图2 总体架构图

（4）基础设施体系。充分利用现有的互联网及政务外网云平台基础环境，结合本项目对网络、系统软件、

系统资源等的需求，添置必要的软硬件资源及设备，整合构建大数据中心各组成部分运行、对外服务、监控、管理等所需基础设施。

（5）标准规范与安全运维体系。标准规范体系由总体标准、业务标准、应用标准、数据标准、安全标准、管理标准等六大类标准规范组成。通过规范的业务流程和标准化的数据定义，促进大数据中心的数据统一接入、统一管理，确保整个平台的规范性、通用性和可扩展性。安全运维体系依托云平台安全体系进行整体考虑，保障大数据中心的稳定、安全运行。

3 研究工作评述

3.1 地质数据标准体系建设

标准规范管理系统中的规范建设内容包括数据采集规范、数据治理规范、数据管理规范。

3.1.1 数据采集规范

数据采集规范和数据接入规范共同确保了来自不同渠道的数据能够有效地接入和处理。数据接入规范定义了各单位内部系统数据、互联网数据、感知数据、视频数据等多样化数据源的技术要求，以便顺畅地整合这些数据。而数据处理规范则对数据治理工程中涉及的数据处理过程的技术要求进行了规定，确保数据在处理过程中的一致性和准确性。这两套规范为数据的标准化采集和处理提供了重要的技术指导，是构建高效、可靠数据治理体系的关键组成部分。

3.1.2 数据治理规范

数据治理工作需要遵循一系列规范，以规范不同渠道来源数据的接入和处理。这包括数据资源目录管理标准，它指导数据资源的采集、汇聚、编目和挂接，推动数据治理工作的开展；数据质量管理标准，它规范数据质量管理的主要工作、方法、实施流程、质量评价指标和审核评估过程；数据安全管理标准，它指导数据资源的分级分类管理，数据脱敏的实施，并明确数据在采集、存储、加工、传输、应用等环节的安全要求，确保数据治理过程中的全生命周期安全。这些规范共同构成了一个全面的数据治理框架，旨在提高数据的可用性、准确性和安全性，支持有效的数据管理和决策制定。

3.1.3 数据管理规范

数据管理规范在数据治理工程中起着至关重要的作用，它从数据分级、数据运维、数据质量和数据标签等方面对数据管理的各个阶段进行贯穿和指导，适用于数据治理工程项目的规划设计者、开发者、建设者和使用者。数据分级分类规范明确了数据资源的分级分类原则、方法和使用原则，指导数据治理平台在开放和共享数据资源时进行分类，并为数据定级提供参考。数据运维规范规定了各类数据在采集和维护阶段的管理方式和技术要求，包括数据库管理员的主要职责、日常管理工作、安全管理规定等。数据质量规范确保了数据质量管理的主要工作、方法、实施流程、质量评价指标和审核评估过程的规范性。标签管理规范则对数据标签体系的构建原则、规则定义、分类原则、计算方式和服务接口等进行规定，以实现对数据的专业分类和作业阶段分类的管理。这些规范共同构成了一个全面的数据治理框架，旨在提高数据的可用性、准确性和安全性，支持有效的数据管理和决策制定。

3.2 地质数据归集、存储、治理

3.2.1 地质数据归集存储

依据全周期、全样本、全方位的采集原则，通过互联网等技术加快数据采集能力的提升，实现地矿数据的全领域归集，形成统一的大数据资源池，保障数据的实时更新和集中存储。按照“一数一源、一源多用”原则汇聚地质数据，并遵循“多源校核、动态更新”原则确保数据的准确性、完整性和时效性，保持数据一致性，提供统一高效的数据服务。

采集数据源管理通过配置和维护数据库连接，支持各种数据源，包括关系型数据库、大数据平台组件和分布式数据库。数据抽取支持结构化、半结构化、非结构化数据的批量抽取，提供丰富的抽取组件和界面化维护。数据转换包括文件和记录级别的转换，支持自定义数据处理规则。数据装载支持将处理后的数据写入存储数据库或服务器文件中，支持装载组件的封装和定制。流式数据采集支持准实时和实时的数据接入，包括小文件批次采集、消息中间件采集、日志数据集成采集等。数据校验包括文件级和记录级的校验，支持复杂场景下的校验规则，对异常情况进行全链路监控。

3.2.2 地质数据治理

数据治理系统的目标是实现统一、高质量、可用、好用、价值最大化的数据资源，涵盖数据标准管理、元数据管理、血缘管理、数据质量管理、数据资产管理、数据目录管理、数据生命周期管理、影响分析等功能。通过这一系统，能够解决数据定义不规范、数据价值未最大化、数据质量差、数据影响分析困难、数据资产分散等问题，实现数据全生命周期的管理，提高数据管控能力，为业务系统提供高标准、高质量、高可用性的数据资源。

3.3 地质数据交换与共享

地质数据共享交换面临诸多挑战，包括数据归属方

的不愿共享、共享方式的不统一、使用权限和边界的模糊，以及潜在的安全风险、争议和责任问题。这些难点导致了地质数据共享的低效和不稳定，限制了数据的充分利用和价值发挥。

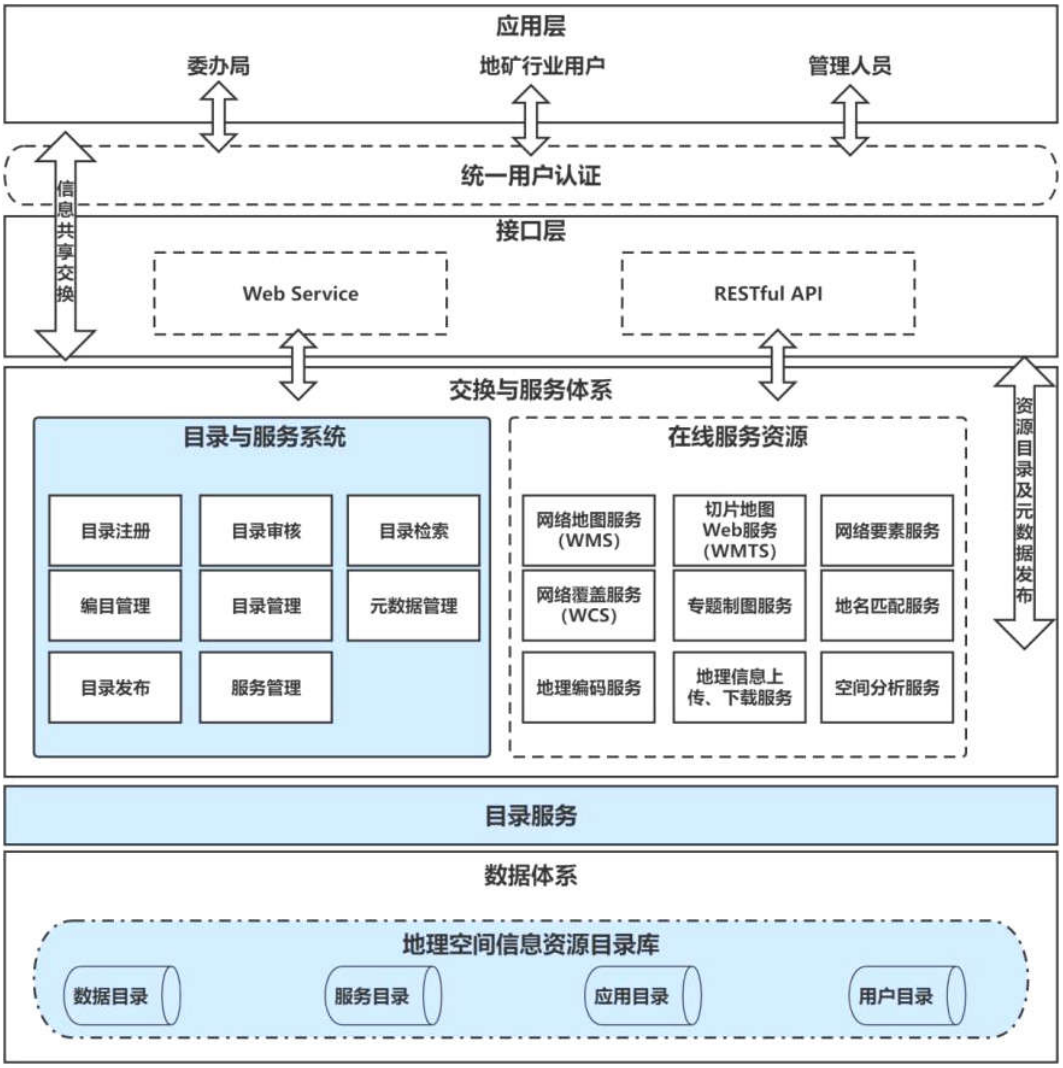

图3 地质数据共享交换平台架构图

地质数据共享交换的大前提是按照统一的标准开展基础数据库建设，通过统一数据共享交换机制实现数据联动，从共享机制模式角度可归纳为以下几类。

3.3.1 “中心化”模式

中心化模式是一种数据共享交换方式，它通过建立一个共享交换平台来实现数据的集中管理，无论是通过物理集中还是逻辑集中。在这种模式下，任何部门在需要使用其他部门的数据时，都必须通过这个平台来进行数据的获取。平台上存储的数据可能是物理集群中的数据，也可能是通过平台调取的其他部门的数据。这种模式有助于提高数据共享的效率和实时性，同时也有利于明确数据的使用权限和边界，减少数据共享过程中的安全风险和争议。

3.3.2 “三权分立”模式

在数据共享使用过程中，涉及到数据的拥有方、使用方和管理方三个对象。分别赋予这三方不同的权利和责任：数据拥有方拥有对数据的定义和解释权，负责数据的收集、整理、维护和更新，并对数据资源的完整性和质量负责；数据使用方有权在规定的范围内使用数据资源，按照最低限度的权限申请使用，并遵守相应的权限要求；数据管理方则有权决定数据的共享方式和程度，负责协调和解决数据共享过程中的争议和问题。这种三权分立的模式在数字政府建设中常见，通过设定权责边界和制度规则，以及利用技术手段如区块链和智能合约，建立一个安全、可信、可追溯的数据共享交换机制。在此机制中，数据目录是关键要素，将部门的职能

和数据目录相衔接，确保应用系统和数据资源的对应关系，从而提高数据共享交换的效率和安全性。

3.3.3 接口技术

该系统提供了丰富的接口管理功能，包括对API采集接口的增加、修改、删除操作配置，以及授权管理，确保只有授权的采集人员可以使用API接口。同时，系统支持基于Socket、WebService、SOAP和HTTP等多种协议的采集规则配置，以适应不同类型的数据源。数据采集后，系统将数据输出到消息队列中，需要配置相应的参数信息，以便后续的实时数据开发可以从消息中间件中获得数据源。这种灵活的数据采集和管理方式，确保了系统能够高效、安全地处理和传输各种格式的数据。

4 结束语

本项目通过梳理国内外研究资料、开展调查研究和对比分析，制定了用于地矿各单位地质数据交换共享的接口规范，对地质数据的共享交换机制进行了初步探讨，旨在提升地质数据的应用价值，避免重复投入，减少资源浪费，同时为地矿服务政府、服务企业、服务公众起到一定的推动作用。

参考文献

[1]谭永杰,文敏.地质信息化建设研究进展与展望[J].中国地质调查,2023,10(02):1-9.DOI:10.19388/j.zgdzdc.2023.02.01.

[2]李加明,云南地质大数据平台技术研究及应用.云南省,云南省地矿测绘院有限公司,2022-04-21.

[3]樊旭东,陈祺,于鑫,等.数字地质服务标准化体系关键技术研究与实践[J].资源环境与工程,2024,38(03):308-313.DOI:10.16536/j.cnki.issn.1671-1211.2024.03.009.