DOI: 10.13758/j.cnki.tr.2024.06.002
刘杰, 郑梦思, 王昌昆, 等. 土壤大数据体系构建及应用. 土壤, 2024, 56(6): 1156–1164.

土壤大数据体系构建及应用①

刘 杰[1], 郑梦思[1], 王昌昆[1,2], 郭志英[1,2], 孙维维[1], 马海艺[1,2], 袁自然[1,2], 潘贤章[1,2*]

(1 土壤与农业可持续发展重点实验室(中国科学院), 南京 211135; 2 中国科学院大学, 北京 100049)

摘 要: 中国已经积累了大量结构化、半结构化和非结构化土壤数据资源, 但它们往往来源多样、结构各异、组织无序且存储分散, 亟待进行体系化整合及高效管理, 以适应大数据人工智能时代对土壤数据的挖掘与应用。基于大数据、数据湖和数据仓库等理念和前沿技术, 本文提出了土壤大数据体系框架及其构建流程和方法, 重点论述了土壤数据采集预处理、非结构化数据识别、算法模型等关键技术和湖仓一体存储架构, 以及数据共享服务方案, 列举了其在土壤数据共享服务、污染场地智能化管控和土壤生态规律挖掘等方面的应用案例, 并探讨了其在第三次全国土壤普查土壤数据资源库建设方面的潜力, 以及数据驱动的土壤研究的应用前景。

关键词: 土壤大数据; 大数据体系; 湖仓一体化; 全国土壤三普; 土壤污染

中图分类号: S159.2; TP399 **文献标志码:** A

Soil Big Data Architecture: Construction and Practical Applications

LIU Jie[1], JIA Mengsi[1], WANG Changkun[1,2], GUO Zhiying[1,2], SUN Weiwei[1], MA Haiyi[1,2], YUAN Ziran[1,2], PAN Xianzhang[1,2*]

(1 *Key Laboratory of Soil and Sustainable Agriculture*, *Chinese Academy of Sciences*, *Nanjing* 211135, *China*; 2 *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

Abstract: Plenty of structured, unstructured and semi-structured soil-related data have been accumulated in China, but they are from multiple sources with varied structures and stored dispersedly and disorderly. In the age of big data and artificial intelligence, in order to satisfy with the requirements of soil data mining and application, it is urgent to construct soil big data architecture. In this study, based on concepts and cutting-edge technologies of big data, data lake and data warehouse, the approaches to construct soil big data architecture were proposed, including technologies on soil-related data acquisition and preprocessing, unstructured data processing, algorithms and models. Then, soil data storage scheme was built through data lake and warehouse, and soil data sharing and serving schemes were realized with the support of data catalog management. Meanwhile, soil big data have been used in soil data service, intelligent control of contaminated sites and mining soil ecological patterns. Finally, the potentials of soil big data were discussed in the usage of construction of the database of third national soil census and prospects for the application in data-driven soil researches.

Key words: Soil big data; Big data architecture; Lakehouse integration; The third national soil condition census; Soil pollution

我国将“数据”与土地、劳动力、资本、技术并列为生产的五种要素[1], 2020 年发布的《中共中央国务院关于新时代加快完善社会主义市场经济体制的意见》[2], 提出要进一步加快培育发展数据要素市场。土壤数据作为数据要素的重要部分,是土壤研究的基础要素和支撑保障,对土壤及其相关数据的深入分析和挖掘, 将有力推动土壤新规律的发现、新认知的形成, 进而在耕地地力提升、土壤污染监测和风险防控、耕地资源可持续管理、智慧农业等方面发挥重要作用[3-6]。

通过全国土壤普查、全国土壤环境背景值调查、全国土壤污染物调查、测土配方施肥、中国土系调查等国家级调查工作, 以及众多科学研究过程中进行的土壤调查和相关试验, 我国已积累了大量土壤数据。同时, 通过中国生态系统研究网络(Chinese ecosystem research network, CERN), 还积累了大量农田、森林、草地、荒漠等典型生态系统土壤长期监测数据,当前,

①基金项目: 国家重点研发计划项目(2020YFC1807401)、国家科技基础资源调查专项项目(2021FY100703)和中国科学院网信专项项目(CAS-WX2022SF-0201)资助。

* 通讯作者(panxz@issas.ac.cn)

作者简介: 刘杰(1986—), 男, 河南新乡人, 博士, 主要从事土壤大数据挖掘及应用研究。E-mail: liujie@issas.ac.cn

通过传感器和物联网，仍在不断产生土壤温度、水分、pH、盐分等动态数据。此外，利用“3S”、数字土壤制图等技术，还生产了大量土壤空间专题图，以及土壤质量、土壤健康、土壤生态多功能性等评价数据。基于以上数据，各方机构不仅建立了“中国 1∶100 万土壤数据库”、“中国 1∶400 万土壤数据库”、“中国土种数据库”、“中国土壤信息系统”等国家级土壤数据库和数据平台[7-12]，还建立了“河南省 1∶20 万土壤数据库”、“浙江省 1∶5 万大比例尺土壤数据库”等省县级区域土壤数据库[13-18]。

然而，土壤数据多源异构、组织无序、存储分散，很难满足研究和应用需求，亟待进行体系化高效管理。近年来，大数据相关理念及技术飞速发展[19]，其技术链囊括数据采集、处理、存储、治理、分析挖掘、共享应用等多个环节[20]，为土壤相关数据资源有效管理及挖掘与应用提供了有效途径。构建土壤大数据体系，实现土壤数据资源的统一存储管理和挖掘应用，可以多方位支持土壤污染状况调查与管控、第三次全国土壤普查等国家大型土壤相关项目，并为农业生产布局优化、生态环境保护、农业可持续发展等提供数据和技术保障。

1　土壤大数据体系框架

1.1　大数据概念

大数据指数量庞大且复杂的数据集，具有容量大(volume)、速度快(velocity)、多样化(variety)、低价值密度(value)、真实性(veracity)等“5V”特性。国内许多政府部门、科研机构、行业建立了自己的大数据体系，如广东省建成了“开放广东”全省政府数据统一开放平台，中国地质调查局建设了“地质云”3.0 版，中国科学院建立了科学数据中心体系，百度、阿里、腾讯、新浪等都建立了不同应用方向的大数据。

传统观点认为土壤数据体量小、高度结构化、数据质量高，具有典型的科学数据特征。但近年来，随着土壤信息获取技术的迅速发展，与土壤相关的传感器数据、遥感数据、土壤基因组和代谢组数据等新来源数据不断涌现，土壤数据越来越呈现出明显的大数据特征。面对来源多样、格式繁杂、内容迥异且存储分散的土壤数据资源，如何更好地管理和高效利用成为一个难题。

近年来，数据湖概念的提出为土壤大数据的组织管理提供了一种新的思路[21]。数据湖能够存储结构化和非结构化数据，是一种面向大规模、多来源、高度多样化数据的组织方法[22]，且具备完善的数据管理能力。然而，数据湖一般不直接面向数据应用，需要通过数据治理、处理、集成等，将数据湖中的数据对接到面向应用的数据仓库[23]，形成湖仓一体的数据管理应用架构。湖仓一体非常适用于土壤大数据的组织管理，再结合土壤专业分析算法和模型，可以实现多场景数据服务和挖掘应用。

1.2　土壤数据资源组成

土壤数据包括结构化、半结构化和非结构化数据。一般来讲，结构化数据是指可以由二维表结构表达的数据，是高度组织和整齐格式化的数据，具有易于检索、分析、存储等特点，如土壤理化属性数据。对于具有预定义字段或结构的矢量和栅格空间数据，也将其纳入结构化数据范畴。非结构化和半结构化数据是土壤相关数据中数量最丰富的数据，但是难以直接利用。

1.2.1　结构化数据　　1)二维表结构数据。土壤自身数据多以二维表结构形式存储，是指在土壤发生、发展、分类、评价及应用中产生的数据，包括土壤采样调查、理化及生物性状、土壤分类等相关数据，也包括土壤形成演变、土壤与其他界面物质和能量交换等过程中产生的科学数据，如野外调查及实验室化验分析获得的 pH、土壤养分、土壤元素等数据，水温盐传感器数据，以及温室气体排放数据等。但是这些高质量数据仅占土壤相关数据资源的 20% 以下[24]。

2)空间数据。例如中国土壤类型、土壤理化属性等空间分布图，每个图斑代表不同类型或者属性范围，是典型的空间数据格式。此外，大量反映土壤成土环境与过程的点位及面状数据，如地形地貌、气象气候、水文地质、植被、土地利用等数据。遥感数据也在一定程度上反映部分土壤信息，是土壤的一个重要空间数据源。土壤相关空间数据可以分为土壤特性和关联环境等空间数据，这些数据多以矢量或栅格形式存储，空间参照系可能不同且数据量庞大。

1.2.2　非结构化数据　　在土壤相关数据资源中，非结构化数据类型多、数量大，多以论文、报告、统计年鉴、图像、音视频等形式存在，需要首先进行技术处理，抽取关键信息，形成结构化数据，才能方便使用。以土壤科学论文为例，它们通常包括文本、表格和图形等，因此，需要首先进行文本实体数据抽取、表格数据识别，以及柱状图、散点图等图形数据自动提取等处理，以获取气候、母质、地形等环境背景信息，以及试验方法、统计数据、分析结果、结论等关键信息，以形成便于利用的结构化数据。

1.2.3　半结构化数据　　土壤相关数据资源中同样存在大量半结构化数据。例如，土壤测试分析仪器设

备产生的大量日志数据，各类土壤数据库和平台的系统日志，新闻、微博等网络舆情数据，多以便于传输交换的 JSON、XML、GML 等数据形式表达。以 GML 格式的土壤采样数据为例，介绍对半结构化数据进行结构化转换的过程。首先，归纳已有土壤采样 GML 数据，梳理出所包含的信息类别(如采样点空间位置、地形信息、天气信息等)，并加入其他关注的信息，建立相关信息子表；然后，设计开发相应工具或中间件，实现 GML 数据到数据库表的自动加载；最终，实现半结构化数据到数据库表的转换，以便于提升数据检索、统计分析效率。

1.3 土壤大数据框架

土壤大数据体系框架主要由数据采集预处理、存储集成、非结构化数据处理、算法模型及应用服务等部分组成。首先，针对不同公共源土壤相关数据的特点，利用网络爬虫等技术，进行相关数据的采集和预处理。其次，利用数据湖和数据仓库理念，构建土壤相关结构化、半结构化及非结构化数据存储架构，结合数据治理、处理、集成等技术，形成面向特定专题的土壤数据仓库。再次，构建空间分析、统计分析、关联分析、机器学习等分析挖掘方法，以及土壤侵蚀、酸化、碳氮耦合、空间预测等专业模型，形成土壤算法模型库。最终，开展土壤数据共享服务、支撑全国土壤普查、生态环境保护、土壤知识挖掘等方面的应用。土壤大数据体系框架如图 1 所示，下面分别介绍各部分组成及功能。

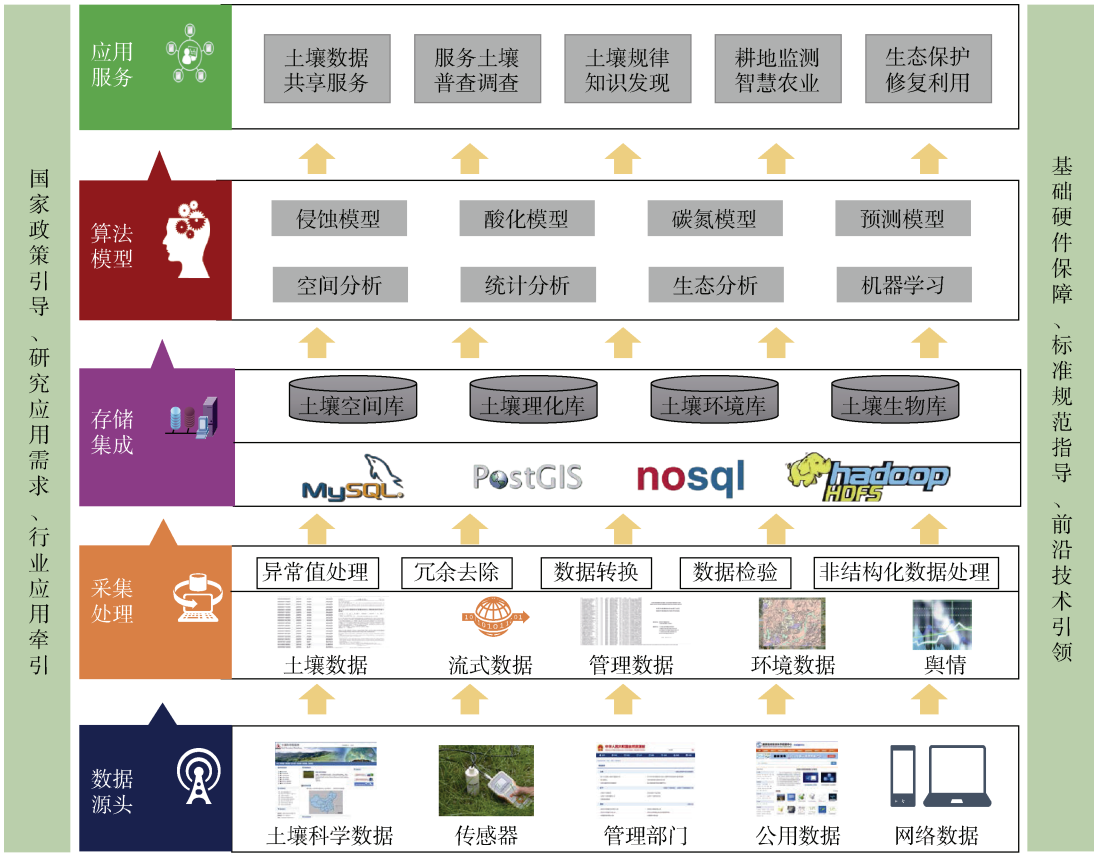

图 1 土壤大数据体系框架

Fig. 1 The framework of soil big data architecture

2 数据采集处理与分析挖掘

2.1 土壤公共源数据采集

在遵循国家相关法律和行业规范的前提下，采用网络爬虫、网页缓存、API 接口、数据库同步、批量接入等技术，构建公共源土壤数据采集系列技术(图 2)。对于采集获取的土壤相关数据资源，进行数据预处理，如冗余去除、缺失值处理、格式转换等，并记录数据来源、获取时间、特征等描述信息。

2.1.1 遥感数据采集　近年来遥感和计算机技术的飞速发展，为地上植被、土地利用、土壤关键属性等监测提供了快速、便捷、宏观、无损的方法[25]，遥感数据已经成为土壤研究的重要数据源之一。对于 MODIS、Landsat、Sentinel 等公开遥感数据，利用

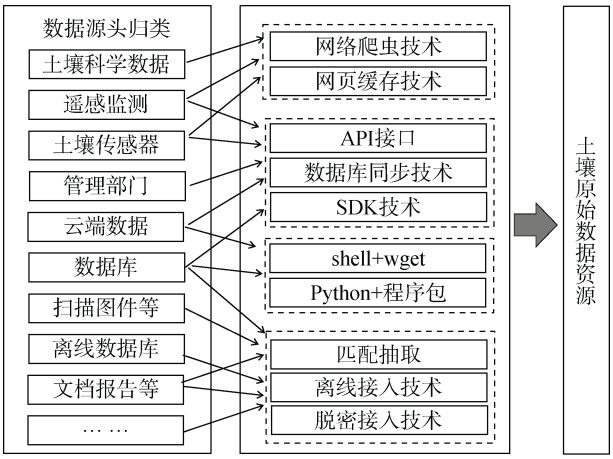

图 2　土壤数据采集技术框架

Fig. 2　The technical framework for collecting soil-related data

Python、R、Linux Shell 等开发语言，可以实现指定时空范围、云量、影像级别等条件的遥感数据批量采集处理。

2.1.2　部门公开数据采集　　在我国管理部门的公开数据中，土壤环境相关数据较丰富，如环境影响评价报告、水质监测信息、排污单位自行监测信息等。针对不同数据的特点，利用 Python 及其 urllib、requests、grab、pycurl 等程序包，通过模拟浏览器操作等方法，可以实现多种部门公开数据的采集。

2.1.3　网络动态数据采集　　在互联网公开数据中，工商企业黄页信息以及微博、新闻网站、微信公众号等网络动态数据均为土壤大数据的重要来源。通过解析不同类型网络动态数据载体的特征，基于 Python 语言和相关程序库，可以实现土壤相关网络动态数据的采集处理。

2.2　非结构化数据处理

土壤相关科研论文、调查报告、评价报告等文档中包含丰富信息，对其关键信息进行结构化识别处理，生成便于直接使用的数据形式，是大数据处理中的重要技术环节[26-27]。以建设用地土壤污染状况调查报告为例，介绍非结构文档识别流程和方法(图 3)：首先，构建“图–表–文”主题内容解构方法，对其中的“图、表、文”进行抽取，获取结构/半结构化的表格数据，以及非结构化的图片数据；然后，基于抽取的文本数据，采用自然语言处理方法(Natural language processing，NLP)进行文本要素抽取，获取(半)结构化的文本要素数据。

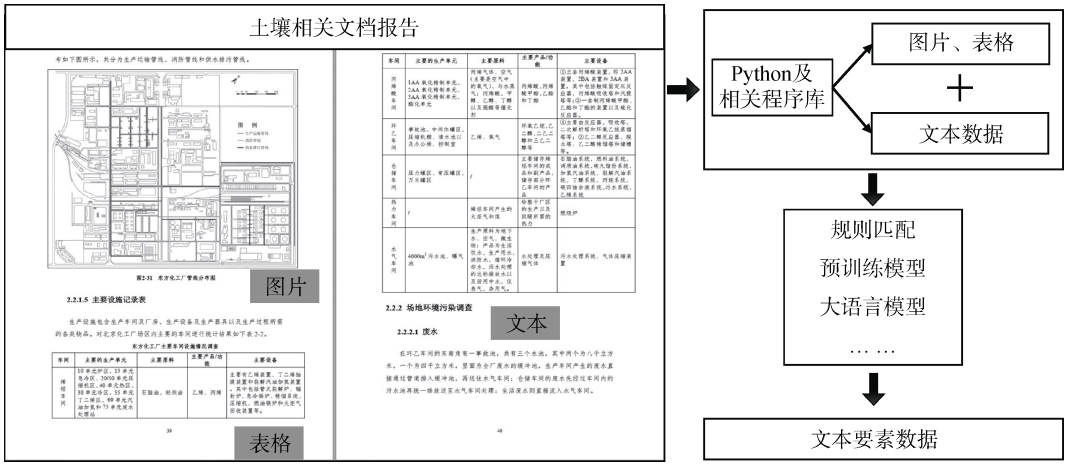

图 3　土壤相关非结构化文档识别处理技术流程

Fig. 3　The technical workflow for recognition processing of soil-related unstructured documents

2.2.1　非结构化文档解构　　在非结构化文档报告中，图注通常位于图的下方，表注大部分位于表的上方，且表格数据可能跨多页。结合图和表的不同特点，利用 Python 语言及 pdfplumber、pymupdf 等程序包，可以实现图片和表格及其说明文字的自动提取；根据文档报告目录，将文档转化成片段式的文本文件。

2.2.2　文本要素抽取　　文本要素抽取是指从自然语言文本中抽取预先定义好的要素标签对应的信息，如人名、地名、机构名等短语级要素，或事件的经过等句子级甚至段落级要素，从而将文本转化为计算机可处理的信息。可采用规则匹配、预训练模型和大语言模型等 NLP 方法，进行文本要素的自动抽取。

1)　基于规则匹配的方法。基于规则匹配的方法通过定义相应的匹配规则集合，对特定类型的文本要素进行识别[28]。例如，在污染场地调查报告中，包含“位于”、“占地面积”、“建成”、“停止运行”等关键词的句子，与要素标签“地块位置”、“地块面积”、“起始时间”、“结束时间”等要素标签有关。理论上，只要制定足够量的匹配规则及合适的优先级，便可取得较高的提取准确率，但该方法费时费力且需要丰富

的匹配规则构建经验。

2) 基于预训练语言模型的方法。预训练语言模型是 NLP 的重要模型，借助于预训练阶段从海量通用数据中学习到的词汇、结构、语义等知识，针对土壤污染状况调查报告的标注数据进行模型微调，可以实现文本要素智能抽取[29]。预训练语言模型法可以实现较高精度的文本要素抽取，但需要足够多的训练样本进行标注和模型多次微调，同样需要大量人力和时间投入。

3) 基于大语言模型的方法。近期大语言模型(Large language models，LLMs)的出现，大大推动了文本要素抽取技术的进步[30]。国内外推出了 ChatGPT、文心一言、通义千问、盘古、星火等商用 LLMs，以及 ChatGLM-6B、LLama、Alpaca 等开源 LLMs。通过开源 LLMs 的本地化部署，利用 Python 程序设计语言和 Open AI API 接口库，并根据所提取数据的特征构建抽取提示词，可以实现文本要素信息的自动抽取。然而，LLMs 法可能会抽取出无关信息、错误信息甚至未出现信息，需要通过模型微调、提示词优化等方法改进。

2.3 数据分析挖掘

土壤大数据的深入挖掘是实现从“数据到知识”的关键节点，基于统计分析、生态分析、空间分析、机器学习等通用数据分析方法，以及侵蚀模型、酸化模型、碳氮模型等土壤专业模型，可构建土壤分析挖掘算法模型库。

1) 通用分析方法。统计分析可以对数据进行初步了解，常用分析方法有：相关性分析、方差分析、集中趋势分析、离中趋势分析、主成分分析等。空间分析是发现土壤空间规律的重要手段，如空间中心计算、空间自相关分析、栅格计算等方法。生态分析是发掘土壤生态规律的重要手段，如土壤生物多样性计算、降维分析、聚类分析、差异检验、驱动因子分析等方法。此外，还有回归分析、空间插值、地统计，以及支持向量机、随机森林、神经网络等方法。

2) 土壤专业模型。土壤专业模型可以对土壤中的物理、化学和生物相互作用，进行多尺度、多层次的定量描述，是土壤科学研究中的重要手段[31]，也是进行土壤大数据知识挖掘的重要途径，例如土壤侵蚀模型、土壤酸化模型、土壤有机碳过程模型(如 RothC 和 DNDC)、陆面过程模型等。

3 土壤数据资源库

土壤相关数据经预处理之后需进行统一存储，便于后续使用和管理。针对表格数据、数据库数据、文档数据、空间数据、图片数据、舆情数据等众多类型土壤数据资源，采用文件存储、(非)关系数据库、空间数据库等方式存储其原始数据，设计开发数据抽取、转换、装载等中间件，集成整合为面向特定专题的土壤数据仓库，形成符合湖仓一体化要求的土壤数据存储方案(图 4)。

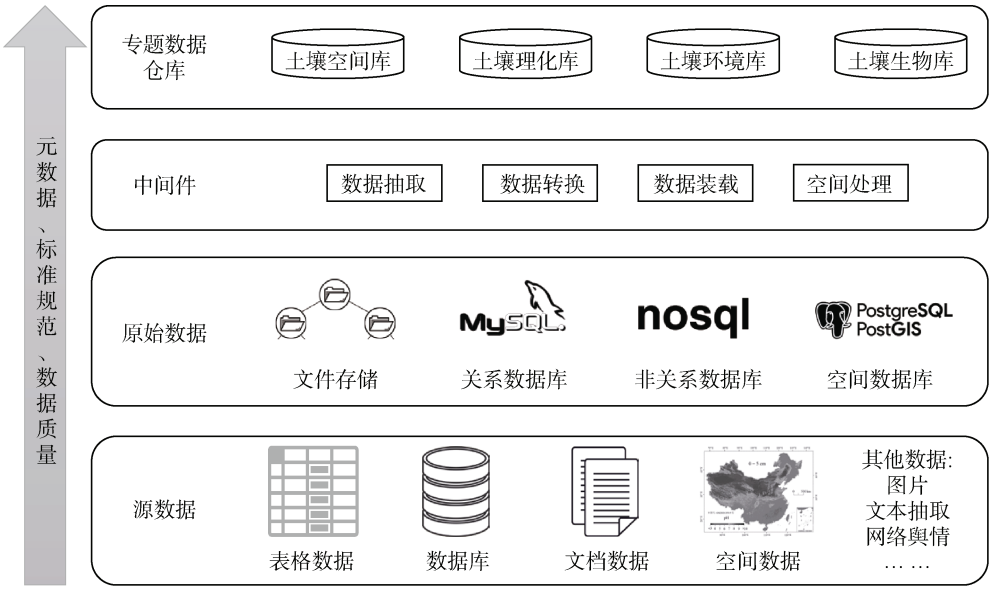

图 4 土壤相关数据存储架构

Fig. 4 The architecture for soil-related data storage

3.1 原始数据存储

预处理后的土壤数据依然格式多样，利用文件存储、关系型数据库、非关系型数据库、空间数据库、分布式文件系统等存储技术，对土壤相关表格数据

(如采样调查信息、土壤理化属性、土壤生物多样性、样点环境因子等)、数据库数据(如中国土壤数据库、中国土种数据库、农田生态系统土壤养分动态数据库、工商企业数据库等)、文档数据(如科研论文、环境影响评价报告、污染场地调查评估报告等)、空间数据(如土壤属性图、地形图、土地利用图、遥感数据等)，以及图片、文本抽取数据、舆情数据等，选取合适的技术存储管理原始数据。

3.2　集成整合中间件

统一存储的原始数据仍可能存在字段名称差异、度量单位不统一等现象，需进一步对相关数据进行抽取、转换、变换等处理，构建面向专题应用的标准统一的数据仓库。例如，对于来源不同内容相似的结构化数据，通过字段映射、单位转换、数据抽取等中间件，实现结构化数据的集成整合；对于大量格式、坐标系统存在差异的空间数据，通过格式转换、坐标变换等中间件，实现时空数据的标准化集成。

3.3　专题数据仓库

面向土壤数据共享服务、普查调查、知识发现、智慧农业、生态环境保护等应用方向，基于关系数据库和空间数据库，利用中间件进行对原始数据的加工处理，建立专题数据仓库，提高土壤数据利用效率，以便于土壤知识挖掘。

4　数据管理与共享服务

以土壤原始数据资源和专题数据仓库为基础，形成土壤大数据资源目录，对土壤数据资源进行检索和浏览。利用数据库视图、Web 服务、地图服务、FTP 服务等方法，为获得权限的外部用户(单位、组织、个人等)提供数据共享服务(图 5)。

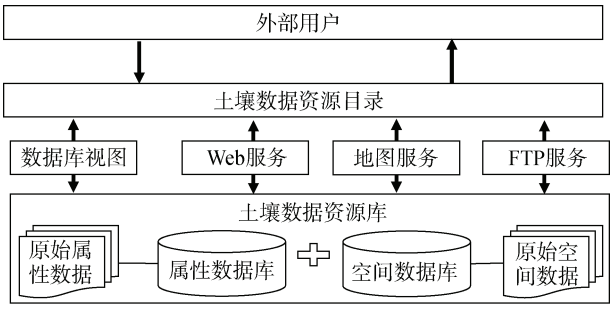

图 5　土壤相关数据共享服务总体架构

Fig. 5　The technology framework for soil-related data sharing services

4.1　数据目录管理

数据目录是土壤数据集中式元数据存储库，目录管理系统能让数据资源管理变得简单高效，一般包括目录分类创建、目录编制、目录审核及目录更新等功能。

1) 目录编制。包括目录资源基本信息、共享属性、开放条件、描述信息等的自定义设置，支持新增、查询、导入、导出，及目录模板化自动导入等操作，目录编制好后提交审核。

2) 目录审核。根据数据自身特征、用途、类型，以及相关法律、法规、政策、标准等，对数据目录进行审核并提出意见，由目录编制人员对驳回的目录根据审核意见进行修改。

3) 目录更新。对于已通过审核的目录进行上线发布，对于不再使用的目录进行下线处理，实时更新数据目录。

4.2　数据共享服务

1) 基于 FTP 的共享。针对以文件形式存储的土壤原始数据，按照一定的组织方式构建文件 FTP 服务器，依托数据资源目录管理，将相应数据资源链接到其目录，进行土壤原始数据资源的共享。

2) 基于数据库视图的共享。数据库视图是一种虚拟表，不在专题土壤数据库中实际存在，而在使用中动态生成。通过数据库视图技术，可对特定用户开放特定数据集，保障数据的安全性。

3) 基于 Web 地图服务的共享。地图服务是通过 Internet 或 Intranet 提供地图的方式，使地图、要素和属性数据可用于多种类型的应用程序。利用 GeoServer、ArcGIS Server 等工具，可实现土壤空间数据的在线服务。

5　土壤大数据应用案例

5.1　基于大数据的土壤信息服务平台

基于土壤大数据资源建立的国家级土壤信息可视化与分析平台，支持网页端(http://www.soilinfo.cn/)和移动端(http://www.soilinfo.cn:8080/WebSoil/APP.jsp)(图 6)，具有空间数据可视化、空间分析及私有数据管理等功能，提升了土壤数据空间可视化分析水平，以及对决策模型的支撑能力和用户获取土壤数据资源的便利性。

5.2　土壤大数据支持的污染场地数据高效管理

近年来，我国对土壤环境保护及污染修复特别重视，污染场地土壤相关数据资源爆炸式增长，亟需土壤大数据来支持场地数据资源的组织管理。土壤大数据体系，为污染场地管理提供了土壤数据资源，并有针对性优化形成适用于污染场地的数据采集、非结构化数据处理、集成整合等技术，大力支撑了污染场地数据资源的高效管理(图 7)。

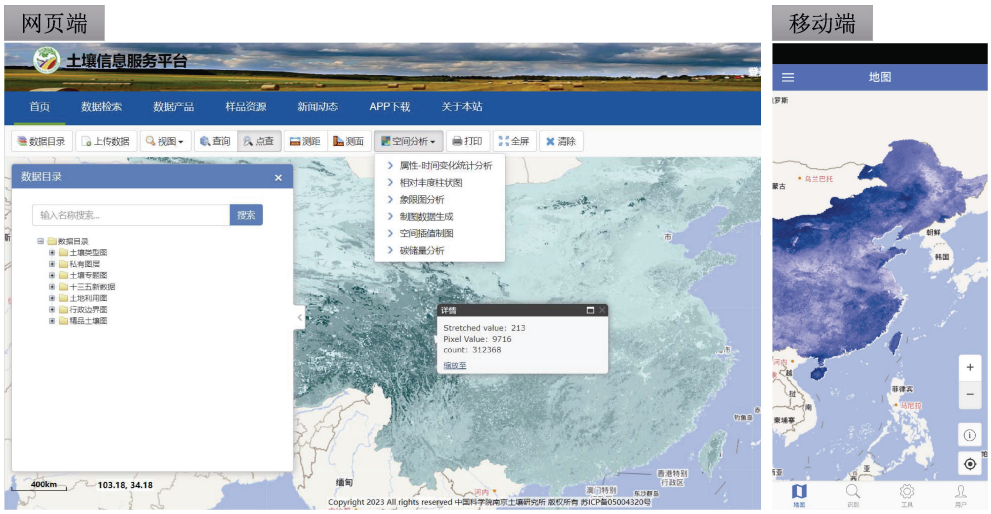

图 6　土壤信息服务平台

Fig. 6　Soil information service platform

图 7　土壤大数据支持的污染场地数据高效管理

Fig. 7　The application of soil big data in efficient management of data resources on contaminated sites

5.3　基于土壤大数据的土壤生态研究

基于集成整合的土壤大数据资源，利用生态统计分析、空间分析、预测模型等方法(图 8)，Liu 等[32]探索了干旱–半干旱荒漠区土壤微生物与植被生产力间的复杂关系，发现资源可用性、物种的环境适应性及物种竞争等生态机制联合引起了土壤微生物多样性–生产力之间的耦合关系；同时，其通过大尺度联合对比分析，发现集约化农业措施特别是灌溉、施肥等，削弱了土壤微生物多样性对气候因子的响应[33]。

6　小结与展望

针对海量、异构、组织无序、存储分散的土壤大数据资源，利用大数据、“3S”、数据库、人工智能等前沿技术与方法，本文初步提出了土壤大数据体系框架及其构建的流程和方法，并简要介绍了土壤大数据体系面向土壤数据共享服务、污染场地智能化管控和土壤生态规律挖掘场景下的初步应用，但其仍需在以下方面进一步提升完善。

1) 进一步完善土壤大数据体系。土壤大数据研究和应用需求十分迫切，但仍有多个技术环节需进一步完善，尤其在非结构化数据处理、高效存储管理、集成整合等方面。此外，在文本可视化、网络图可视化、时空数据可视化等技术方面需要全面加强，以实现土壤大数据可视化展示和管理。

2) 构建全国土壤三普大数据资源库。目前第三次全国土壤普查已经全面展开，很快将形成国家级、省级和县级等不同规模的结构化的土壤三普数据库，同时还将产生大量调查图片、调查分析报告、专题图

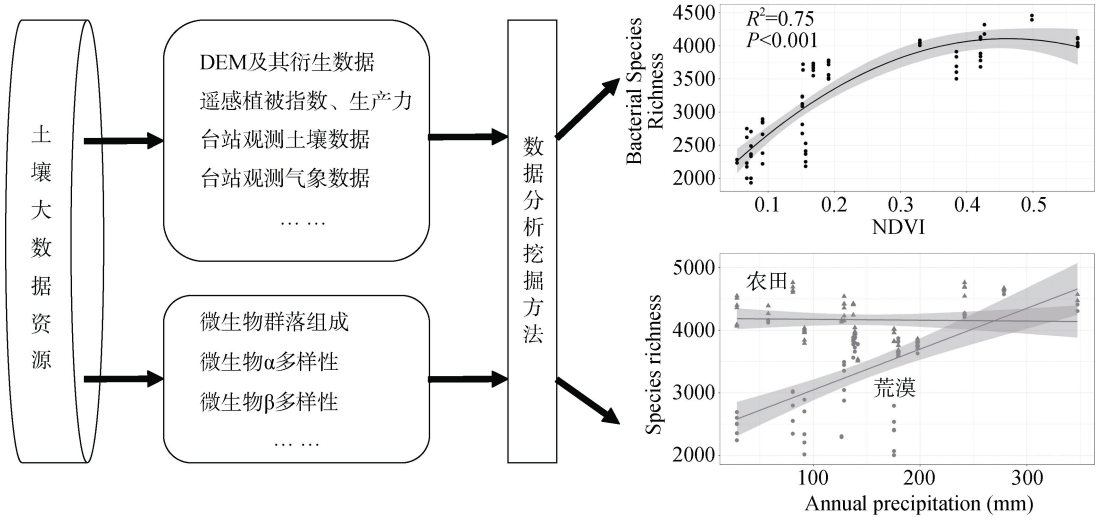

图 8　土壤大数据在生态规律挖掘中的应用

Fig. 8　The applications of soil big data in soil ecological researches

等非结构化数据库。基于土壤大数据体系框架，可以全面整合三普相关结构化和非结构化数据，构建全国土壤三普大数据资源库，为土壤数据深入挖掘及推动成果产出提供支撑。

3) 强化土壤大数据挖掘及应用。大数据和人工智能技术快速发展带来了科研范式变化，在充分融合环境背景、土壤物理、土壤化学、土壤生物、人为活动等土壤大数据的基础上，结合土壤专业知识和模型，利用人工智能等前沿技术，挖掘土壤领域的新规律和新知识，将为黑土地保护利用、退化耕地智慧监测、土壤污染防控等重大战略提供更有力的支撑。

参考文献：

[1] 中华人民共和国中央人民政府. 中共中央 国务院 关于构建更加完善的要素市场化配置体制机制的意见[OL]. 2020-04-09.https://www.gov.cn/zhengce/2020-04/09/content_5500622.htm.

[2] 中华人民共和国中央人民政府. 中共中央 国务院 关于新时代加快完善社会主义市场经济体制的意见[OL]. 2020-05-11.https://www.gov.cn/gongbao/content/2020/content_5515273.htm.

[3] 骆永明, 滕应. 中国土壤污染与修复科技研究进展和展望[J]. 土壤学报, 2020, 57(5): 1137–1142.

[4] 沈仁芳, 王超, 孙波. “藏粮于地、藏粮于技”战略实施中的土壤科学与技术问题[J]. 中国科学院院刊, 2018, 33(2): 135–144.

[5] 张佳宝, 孙波, 骆永明, 等. 开展健康耕地建设行动 夯实粮食安全基础[J]. 中国农村科技, 2023(1): 21–22.

[6] 赵春江. 智慧农业发展现状及战略目标研究[J]. 智慧农业, 2019, 1(1): 1–7.

[7] Shi X Z, Yu D S, Warner E D, et al. Soil database of 1∶1, 000, 000 digital soil survey and reference system of the Chinese genetic soil classification system[J]. Soil Horizons, 2004, 45(4): 129.

[8] 史学正, 于东升, 高鹏, 等. 中国土壤信息系统(SISChina)及其应用基础研究[J]. 土壤, 2007, 39(3): 329–333.

[9] 施建平, 宋歌. 基于 Web 的中国土种数据库[J]. 土壤, 2016, 48(6): 1246–1252.

[10] 张定祥, 潘贤章, 史学正, 等. 中国 1∶100 万土壤数据库建设中的几个问题[J]. 土壤通报, 2003, 34(2): 81–84.

[11] 张甘霖, 龚子同, 骆国保, 等. 国家土壤信息系统的结构、内容与应用[J]. 地理科学, 2001, 21(5): 401–406.

[12] 周慧珍. 中国土壤信息共享研究——1∶400 万中国土壤分布式查询数据库[J]. 土壤学报, 2002, 39(4): 483–489.

[13] 吴克宁, 杨锋, 吕巧灵, 等. 河南省 1∶20 万土壤数据库的构建及其应用[J]. 河南农业科学, 2007, 36(5): 77–80.

[14] 吴顺辉, 胡月明, 戴军, 等. 广东省土壤资源信息系统数据库的研制[J]. 华南农业大学学报, 2001, 22(4): 22–25.

[15] 张定祥, 于东升, 史学正. 苏南 SOTER 数据库的建立及其在水稻土生产力评价上的应用[J]. 安徽农业大学学报, 2001, 28(2): 119–124.

[16] 张学雷, 张甘霖, 龚子同. SOTER 数据库支持下的土壤质量综合评价——以海南岛为例[J]. 山地学报, 2001, 19(4): 377–380.

[17] 吴世蓉, 位佳, 邱龙霞, 等. 基于大比例尺数据库的福建省耕地土壤固碳速率和潜力研究[J]. 土壤学报, 2022, 59(5): 1293–1305.

[18] 吴嘉平, 胡义镰, 支俊俊, 等. 浙江省 1∶5 万大比例尺土壤数据库[J]. 土壤学报, 2013, 50(1): 30–40.

[19] 李国杰, 程学旗. 大数据研究: 未来科技及经济社会发展的重大战略领域——大数据的研究现状与科学思考[J]. 中国科学院院刊, 2012, 27(6): 647–657.

[20] 刘智慧, 张泉灵. 大数据技术研究综述[J]. 浙江大学学报(工学版), 2014, 48(6): 957–972.

[21] James, D. Pentaho, hadoop, and data lakes [OL]. James Dixons Blog, 2010. https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/.

[22] Oracle 中国. 数据湖[OL]. 2023-09-14. https://www.oracle.com/cn/big-data/data-lake/what-is-data-lake/.
[23] Janssen N E. The evolution of data storage architectures: Examining the value of the data lakehouse[D]. University of Twente, 2022.
[24] 中国电子技术标准化研究院. 非结构化数据管理解决方案白皮书(2020 版)[EB/OL]. 2020-09-21. https://www.cesi.cn/202009/6824.html.
[25] 陈仲新, 任建强, 唐华俊, 等. 农业遥感研究应用进展与展望[J]. 遥感学报, 2016, 20(5): 748–767.
[26] 包小源, 黄婉晶, 张凯, 等. 非结构化电子病历中信息抽取的定制化方法[J]. 北京大学学报(医学版), 2018, 50(2): 256–263.
[27] 郑梦悦, 秦春秀, 马续补. 面向中文科技文献非结构化摘要的知识元表示与抽取研究——基于知识元本体理论[J]. 情报理论与实践, 2020, 43(2): 157–163.
[28] 郭喜跃, 何婷婷. 信息抽取研究综述[J]. 计算机科学, 2015, 42(2): 14–17, 38.
[29] 王宇琪, 周庆山. 基于预训练语言模型的互联网开源信息抽取与情报分析应用研究——以“学术、讲座、论坛”等会议活动为例[J]. 情报理论与实践, 2024, 47(1): 154–163.
[30] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners. Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, BC, Canada: Curran Associates Inc., 2020: 1877–1901.
[31] Vereecken H, Schnepf A, Hopmans J W, et al. Modeling soil processes: Review, key challenges, and new perspectives[J]. Vadose Zone Journal, 2016, 15(5): 1–57.
[32] Liu J, Wang C K, Guo Z Y, et al. Linking soil bacterial diversity to satellite-derived vegetation productivity: A case study in arid and semi-arid desert areas[J]. Environmental Microbiology, 2021, 23(10): 6137–6147.
[33] Liu J, Wang C K, Guo Z Y, et al. The effects of climate on soil microbial diversity shift after intensive agriculture in arid and semiarid regions[J]. Science of the Total Environment, 2022, 821: 153075.