地质大数据技术研究与应用试点

促进大数据健康发展已于2015年成为国家战略。地质数据是地质工作过程和成果的记录，是国家大数据的重要组成部分。为研究地质大数据的基本问题，用大数据技术解决地质工作中遇到的各种难题，推进地质工作现代化，地质调查局发展研究中心联合国内高校和研究所等8家单位，在国土资源公益性行业科研专项“地质大数据技术研究与应用试点”项目、相关地调专项和自然科学基金项目支持下，2015年初开始研究，经4年努力，取得丰硕成果，于2019年8月通过评审验收。

成果与创新点突出：

1. 系统归纳概括了地质大数据基本内涵与特征，建立了地质大数据技术框架体系，奠定了地质大数据理论技术基础。

2. 提出了地质大数据分布式管理及计算新模式，构建了分布式计算资源池与实验平台，研发了智能检索系统，开展了地质并行计算模式适用性验证，实现了地质大数据一体化存储与管理。

3. 实现了非结构化地质大数据智能处理技术新突破，研发了包括图件、文本、遥感影像等地质数据结构化处理算法，开发了面向地质单文档的关键词提取和自动摘要系统，实现了非结构地质大数据快速智能处理。

4. 突破了大数据环境下地质实体抽取，关系识别及可视化表达等技术难点，开发了地学文献知识发现系统，形成了中文地质知识库技

术体系。

5. 攻克了面向基础地质、物探、化探、遥感等信息的关联分析技术，研发了多源地质信息综合分析机器学习算法，实现了示范区矿产资源评价的综合分析。

6. 针对矿产资源评价，提出了数据驱动下的评价新模式，搭建了智能发现与云服务平台，研发了基于决策树、SVM、卷积神经网络等算法，围绕多个示范区域，验证了模式的创新性。

学术价值与社会效益显著：

基于该成果，发表论文 50 余篇、出版科普书籍 1 部、获软件著作权 20 项、申请发明专利 15 项（授权 8 项），奠定了地质大数据理论技术基础。

提出的地质大数据概念内涵和基本特征，已被业内采纳；提出设立全球地质大数据工作组建议，得到 2019 年第九届联合国全球地学空间信息专家委员会认可，并写入大会决议。

提出地质大数据技术框架体系有效指导了地质大数据与信息服务工程的实施；研发的地质非结构化数据转化及应用，地质知识发现及关联分析等技术已形成软件在业界得到推广应用；数据驱动下的矿产资源评价新模式已在多个示范区得到推广。整体成果大大促进了国内外地质信息化发展及应用。

〃 〃

（一）研究背景及实施

大数据于 2013 年前后引起地质人的广泛关注，为了研究地质数据是否为大数据、地质大数据的特点、如何用大数据技术解决地质数据采集、管理和处理中的问题，探索形成地质大数据技术体系，为地质信息化建设提供框架性的指导和引领，2014 年初中国地质调查局发展研究中心依托自然资源部地质信息技术重点实验室，联合中国地质大学（北京）、中国地质图书馆、中国地质科学院矿产资源研究所、中国地质大学（武汉）、北京师范大学、成都理工大学等高校和研究所，开始了地质大数据技术研究方面的立项论证工作，经过近一年的文献研究、问题归纳、关键技术论证、主要技术路线试验等， 研究团队向原国土资源部科技部门提交了项目申报建议书，2015 年 5 月部在国土资源公益性行业科研专项中批准设立了行业基金重大项目*地质大数据技术研究与应用试点，项目周期三年(2015*2017)，经费共 788 万元。同时，各单位申报了相关的自然科学基金项目和地质调查项目，研究工作总经费 1500 多万元。

四年多来，研究团队通过该项目四个课题、9 个专题的组织实施，在地质大数据技术体系及平台、地质大数据处理技术、地质文献数据挖掘、矿产资源评价数据分析挖掘等方面开展攻关研究，总项目定期组织交流，并及时协调和解决各课题实施过程中遇到的问题，研究工作总体按照计划进行，并顺利通过年度检查和部组织的中期评估，经过 50 多位研究人员的艰苦努力，研究工作于 2017 年底基本完成，随后开始了示范应用工作，到 2018 年 6 月，研究工作全面完成，各项考核指标达到了要求，实现了预期目标。由于国家机构改革自然资源

部重新组建等原因，该成果于2019年8月通过自然资源部科技部门组织的完成评审验收。

（二）总体思路

地质大数据技术研究与应用试点的总体思路与技术路线是：以文件级地质资料、地学文献及全国矿产资源潜力评价成果数据为研究对象，展开地质大数据特征规律、组织管理、分析处理等系列技术研究和应用试点，在此基础上，归纳提出大数据技术地质应用技术体系，为构建大数据时代地质调查信息化建设总体框架奠定基础。

（三）主要成果与创新点

经过4年多集体攻关研究，取得以下成果和创新：

1.系统归纳概括了地质大数据内涵和特征，建立了地质大数据技术架构体系，奠定了地质大数据研究与应用理论基础

支持材料：《话说大数据与地质大数据》（地质出版社）专著；《地质数据的大数据特性研究》、《地质大数据体系建设的总体框架研究》等论文。

（1）梳理归纳整理了地质大数据的基本特质，提出了地质大数据内涵，实现了地质大数据的科学描述与自画像，达到了对地质大数据的基本认识，奠定了地质大数据理论基础。

（2）从地质大数据的采集、处理、分析挖掘、应用、平台及安全等基本处理技术入手，结合当前通用的大数据平台及计算模式，归纳

总结了地质大数处理的基本问题，评估了大数据技术在地学方面的应用需求及存在的问题，为地质大数据的发展提供了科学依据。

（3）在深入调研和分析大数据相关技术应用趋势基础上，系统提出了地质大数据技术框架体系和技术要点，为地质大数据的研究及应用提供了支撑。

2.提出了地质大数据分布式存储管理及计算新模式，搭建了地质大数据分布式存储计算资源池与实验平台，研发了智能检索系统，开展了地质并行计算模式适用性验证，实现了地质大数据一体化存储与管理。

支持材料：《Big Data Management for Cloud*Enabled Geological Information Services》、《Caber*physical*social*thinking modeling and computing for geological information service system》等论文；“一种互联网地质专题大数据检索与获取的方法及其装置”等发明专利；基于分布式多节点查询系统等软著。

（1）设计并搭建了地质大数据分布式计算并行计算框架，构建了地质大数据分布式计算资源池，实现了地质大数据的一体化存储与管理。

（2）提出了地质大数据多节点查询检索机制，研发了地质大数据分布式索引优化算法，提高了地质大数据分布式检索的准确率和效率。

（3）基于搭建的地质大数据实验平台，开展了多个典型地质计算并行化算法改进和实验，为地质大数据并行计算模式的研究和发展提

供了依据。

（4）提出了针对矿产潜力评价成果大数据的存储优化方法，设计了多级索引表格，实现了矿产潜力评价成果数据的分布式存储和高效访问。

（5）提出了地质大数据语义扩展模型，研发了面向地质大数据的智能检索系统，实现了地质大数据的智能检索。

3.实现了非结构化地质大数据快速智能处理技术的新突破，研发了包括图件、文本、遥感影像等地质数据的智能化处理算法，开发了面向地质单文档信息的关键词提取和自动摘要系统，实现了非结构地质大数据快速智能处理。

支持材料：《地质非结构化数据研究战略—以 JPG 图件为例》、《地质报告文本自动标引技术方法分析》等论文；“一种基于地质大数据的标引关键词提取方法和系统”等发明专利；标准彩色地质图专题信息自动提取系统 V1.0 等软著。

（1）针对地质图件各个要素基本特征，研发了针对点、线及面的地质要素提取算法，为地质图件到结构化数据的一键式操作及高效利用提供了基础。

（2）针对地质文本数据特征，基于通用的机器学习和条件随机场技术，研发了通用的地质分词算法，实现了高准确度的地质分词效果（目前可达 95%左右），这为地质文本大数据的智能处理提供了基础。

（3）首次提出了地质术语库自动和半自动更新方法，并基于目前

通用的地质大词典、地质叙词表等构建了面向地质文本信息的地质术语库。

（4）研发了面向地质资料单文档的自动标引和地质专题信息自动提取技术，构建了面向地质中文文本信息提取的语料库，实现了地质文本的自动/半自动标引及面向专题的信息自动提取。

（5）设计并实现了基于大数据词频的关键词提取算法、大数据加权关键词算法，实现了基于单文档的关键词自动提取。

（6）面向地质资料报告，创新性的提出了基于位置权重、长度权重，基于篇章结构的摘要自动形成技术，开发了基于地质资料报告的自动摘要系统，实现了地质文本摘要的自动形成。

4. 突破了大数据环境下地质实体抽取，关系识别及可视化表达等技术难点，开发了地学文献知识发现系统，形成了中文地质知识库技术体系。

支持材料：《Intelligent Learning for Knowledge Graph towards Geological Data》、《基于文献的地质实体关系抽取方法研究》等论文；地学非相关文献知识发现辅助系统 V1.0（软著）。

（1）创新性的将非相关文献知识发现理论、关系抽取、机器学习等方法相结合并应用于地质大数据的知识发现与服务场景，实现了包括地质文献数据提取整理、自然语言处理(分词/词性标注、关系抽取)、可视化（图谱构建与应用）等知识发现全流程的集成整合，构建形成了中文地质知识库技术体系。

（2）基于地质文献大数据特征，设计并研发了基于统计语言模型和机器学习的关系抽取模型和基于 Bootstrapping 方法的关系扩展模型，通过专家及机器学习算法学习，构建了一定量的地质实体关系种子库，优化了地质实体抽取和关系识别效果；并基于图数据库、可视化综合等技术，形成基础地质学关联关系图谱和金矿领域知识图谱。

（3）搭建了以 Elasticsearch 为分析搜索引擎的文献大数据处理框架，实现了基于非相关文献间隐含的关联的智能识别，开发学非相关文献知识发现系统，提高了地质文献大数据的服务效率与水平。

5. 攻克了面向基础地质、物探、化探、遥感等信息的关联分析技术，研发了多源地质信息综合分析机器学习算法，实现了示范区域矿产资源评价的综合分析。

支持材料：《基于 Hadoop 的地质大数据融合与挖掘技术框架》、《Semi*supervised Hyperspectral Image Classification Based on Generative Adversarial Networks》等论文；“一种基于可信机制的地质大数据融合方法及系统”等发明专利；地球化学形态学相关系数分析系统 V1.0 等软著。

（1）以全国矿产地数据的热液型金矿为研究对象，提出了基于空间属性信息的基础地质关联性分析挖掘方法，以空间位置为基准，实现了不同类型数据之间关联分析，实现了矿产资源信息与基础地质信息的关联关系识别。

（2）提出了形态学相关系数、基于邻域约束聚类及局部相关系数

的地球化学异常提取方法，实现了化探异常的快速、准确、自动提取。并通过化探异常与地质数据关联分析，研究了元素组合特性与构造之间的关系，为矿产预测评价提供了理论依据。

（3）针对遥感影像数据处理，提出了基于直线和十字交叉点的遥感影像配准、基于 IK*HIS 遥感影像融合、基于对抗生成网络和卷积神经网络遥感影像分类等方法，实现了遥感影像数据从配准、融合到分类的全流程处理，得到了较高精度的遥感影像分类结果。

（4）基于提出的地质大数据综合分析框架，研发了能依据数据特性灵活调整融网格大小的自适应网格化多源信息融合及基于机器学习的综合分析算法，围绕应用示范区域，实现了面向矿产预测的地质大数据综合分析。

6. 针对矿产资源定量评价，提出大数据环境下评价新模式，搭建了智能发现与云服务平台，研发了基于决策树、SVM、卷积神经网络等模型算法，围绕多个示范区域，验证了模式的创新性。

支持材料：《大数据思维下的矿产资源评价》、《数据驱动下的矿产预测模型构建方法研究》等论文；“三维矿产资源预测评价中信息综合处理装置及其方法”等专利；面向成矿预测的地质大数据智能发现与服务平台 V1.0 等软著；甘肃、湖南及四川等应用示范。

（1）设计并搭建了地质大数据智能发现与云服务管理平台，研发了面向矿产资源评价的地质大数据智能发现服务系统，及软件算法云服务管理功能模块，实现了地质大数据的智能发现与应用软件算法模

型的有效集成、管理与云端服务。

（2）选取甘肃北山地区金矿、四川会理拉拉铜矿等作为应用试点，构建了基于决策树、SVM、卷积神经网络三种预测模型，综合地质、物探、化探、遥感多元综合信息，验证结果表明，基于机器学习的预测结果超过了传统的统计分析方法得到的结果，而且所选靶区面积更小，精度更高，部分结果与地质方面的相关理论高度一致。

（3）创新性地应用深度学习算法在湖南省香花岭地区进行矿产预测，通过对该示范区地质、地球化学数据等全样本数据学习训练模型，提取深层隐含特征，分析预测含矿地质单元。验证结果显示模型具有一定的可靠性，初步实现了预测评价过程的自动化与智能化，为进一步深入挖掘地质大数据中深层隐含信息提供技术支撑。

（四）保密方面

该项研究成果无保密内容。

（五）国际对比

地质数据是地质工作过程和成果的数字化记录，我国的地质工作对象、地质工作方法和地质工作体系具有鲜明的中国特色，所形成的地质数据来源多、类型多、粒级多，也就具有中国独特的烙印。在国内中国地质调查局发展研究中心是地质信息化研究建设的龙头单位，是中国地理信息产业协会原地质矿产信息工作委员会的依托单位，针对中国的地质数据归纳概括的地质大数据内涵、特点和地质大数据技

术体系等成果，属于原创性成果。

研究取得的基于彩色地质图件的地质信息定量快速智能提取、面向地质文本的自动标引等非结构化地质数据向结构化数据转换技术达到国际领先水平。研发的地质中文分词、语料库构建、关键词提取、自动摘要、知识图谱构建等技术达到国际先进水平。

研究取得的中文地质知识库技术体系、面向矿产资源评价的关联分析挖掘技术在地质大数据智能分析挖掘方面达到国际先进水平。

〞〞主要的应用推广：

1. 成果的理论转化与应用。依托该项目成果，共发表论文 50 多篇，其中在《中国矿业》2017、2018 年第 9 期组织专栏各发表论文 6 篇，出版《话说大数据与地质大数据》科普读本 1 部，现已发行 1000 多册。该书自出版以来，受到了各年龄段人士青睐和好评，中国工程院王家耀院士评论该书为：目前为止，第一本地质大数据方面的科普著作，具有科学性强、趣味性搞、传播效果佳等特点。此系列成果的发布与传播基本上实现了地质大数据技术研究与应用试点成果向理论的转化应用，丰富和发展了地质大数据理论和方法体系。

2. 提出的地质大数据概念内涵和基本特征，已经基本作为地质大数据理论的核心内容，广泛被业内研究所采纳。研发的非结构化地质数据向结构化数据转化、地质数据知识发现与服务等技术已经形成

软件系统得到了推广应用。提出的地质大数据云架构应用试点平台，地质大数据存储管理模式，地质大数据综合分析挖掘思路和框架，地质大数据并行算法实验以及数据驱动下的矿产预测评价新模式等目前已经通过各种方式提供给中国地质大学（北京）、中国地质大学（武汉）、中山大学、吉林大学、成都理工大学、煤炭地质总局勘查研究总院，陕西省地质环境监测总站、湖南地调院、四川地调院和吉林地调院、航空物探遥感中心等 20 余家单位，在指导各地单位地质大数据研究及应用等方面发挥了重要作用

3. 依托项目开发形成的首个中文地质分词算法模型，中文地质术语库、语料库目前均已逐步对外开放，依托该项成果提出的面向地球系统的知识库构建计划等已列入中国地质调查局十四五规划，同时也以国家十四五规划项目需求征集方式提交科技部。通过该研究构建形成的中文地质知识库技术体系，引起了国外地质学界的广泛关注，目前已经和联合国全球地理信息专家委员会、美国、英国地质调查局专家，进行了多次研讨，期望能通过中西方研究，实现中英文地质知识库的集成整合。

4. 依托该项目设计的地学文献知识发现技术体系架构将文献知识发现理论、关系抽取方法、机器学习方法相结合，具有一定的创新性。开发的地学文献数据加工处理工具，地学非相关文献知识发现辅助系统，基于单文档的关键词提取与摘要自动形成技术等，目前在广州海洋地质调查局，天津地质调查中心，物化探所　岩溶地质研究所，中国科学院大学、国家海洋信息中心，中国地震台网中心，中国林业

科学研究院，中国地质大学（武汉）青藏高原研究中心、四川地调院及中国地质科学院地质研究所等 20 余家单位部门得到推广应用。此成果的推广大大在提高地学文献大数据的服务效率和水平基础上，也人工智能深层次应用，如：面向地质大数据的知识问答，找矿机器人研发等提供了重要的理论和技术基础。

5. 提出的地质大数据体系建设的总体框架和技术要点有效引导和指导了地质大数据与信息服务工程的实施，带动和促进了地质调查信息化建设行业技术进步。其中依托该项目成果，独立和协助举办地质大数据相关大型研讨会 20 余次，与会人员达 3000 多人次。其中包括项目内部研讨会，6 次，地质信息技术论坛 3 次，中国地质青年年会的地质大数据分会论坛 2 次，全国大数据与数学地球科学学术研讨会 3 次，大大的促进了整个地质信息化建设行业的发展。

6. 项目的实施，有效推进了地质信息技术重点实验室建设，形成了高层次的地质大数据业务团队；其中依托该项目直接、间接培训研究生、博士生 50 多名；成功申报国家重点专项课题 2 项，自然基金面上项目 1 项，有力的推进了地质科学研究进展。

7. 依托该项目成果，申请发明专利 15 项，获软著著作权 20 项，其中基于地质大数据的综合分析挖掘技术方法，矿产资源潜力评价新模式已经在 4 个应用示范区进行了矿产资源评价工作。有效的推动了数据驱动下的矿产预测评估新模式的应用，奠定了矿产预测评价新理论。

8. 项目沉淀形成的地质大数据研究系列成果得到国内外普遍认

可。在2019年召开的第九届联合国全球地学空间信息专家委员会上，依据该项研究提出的设立地质大数据工作组的建议，得到与会代表的广泛认可，并写入大会决议，这将大大促进全球地质大数据的研究与应用进展。

30余家推广应用单位列表：中国地质大学（北京）、中国地质大学（武汉）、成都理工大学、吉林大学、湖南科技大学、中国煤炭地质总局勘查研究总院，陕西省地质环境监测总站，数学地质与遥感地质研究所、国家海洋信息中心，中国地震台网中心，中国林业科学研究院林业科技信息研究所，中国地质大学（武汉）青藏高原研究中心，广州海洋地质调查局，天津地质调查中心，航空物探遥感中心，物化探所，岩溶地质研究所，地质科学院力学研究所，甘肃地质调查院，湖南地质调查院，数学地质四川省重点实验室等。

〞　该成果奠定了地质大数据研究的理论技术基础；提出的地质大数据内涵和特征等广泛被业内采纳，提出的设立全球地质大数据工作组建议，得到2019年第九届联合国全球地学信息专家委员会认可，并写入大会决议，扩大了国际影响力；提出的地质大数据体系建设框架有效指导了地质调查信息化建设；研发的非结构化地质数据转化、地质数据分析等技术已形成软件得到了应用；总体成果已在多个示范区得到应用，促进了地质行业信息化发展。〞

中国地质调查局发展研究中心是地质调查系统信息化建设的牵头部门和龙头单位，是部地质信息技术重点实验室依托单位，一致致

力于地质信息化的建设和发展研究工作。2014 年初就牵头联合业内相关单位开展地质大数据的研究和科技计划项目的论证申报工作，2015 年申报的国土资源公益行业科研专项重点项目“地质大数据技术研究与应用试点”获得国土资源部批准，该研究是比较早期专门开展大数据在地质行业应用的项目。

在该项目支持下，发展研究中心依托部重点实验室开发平台，联合地质调查系统、地质院校等多家单位，就大数据技术在地质行业应用发展的基础问题进行了联合攻关研究，取得了丰硕成果，实现了预期目标：系统界定了地质大数据的内涵与特征，搭建了地质大数据实验平台，研发了包括彩色地质图件、文本及遥感影像数据在内的非结构化信息自动处理算法模型，突破了地质文本自动标引、自动摘要技术难点，在地质数据知识发现与分析挖掘、地质大数据体系建立方面等有创新；提出了地质大数据体系建设的总体框架，绘制了地质大数据发展的技术蓝图；围绕示范区域，开展了多处矿产预测评价试点工作，示范效果良好。该项研究，奠定了地质大数据技术基础，推进了地质信息化行业技术进步。

经过我中心科学技术委员会审议，同意推荐该成果申报地理信息科技进步一等奖。

"