大数据背景下地质云的构建与应用

陈建平[1,2], 李 婧[1,2], 崔 宁[1,2], 于萍萍[1,2]
CHEN Jianping[1,2], LI Jing[1,2], CUI Ning[1,2], YU Pingping[1,2]

1.中国地质大学(北京) 地球科学与资源学院,北京 100083;
2.北京市国土资源信息研究开发重点实验室,北京 100083
1. *School of Earth Sciences and Resources, China University of Geosciences (Beijing), Beijing* 100083, *China;*
2. *Land Resources Information Development and Research Key Laboratory of Beijing, Beijing* 100083, *China*

摘要:地质学属于数据密集型科学,并与地球科学面临的问题息息相关。已经收集的和将要收集的大量数字国土相关数据,由于科学研究的需要,正在不断加以检验和扩充。大数据时代背景下,中国国土资源数字化、信息化的战略行动具有深远意义,大数据的相关技术应用为实现地质工作的现代化和信息化提供了有效的支撑。重点介绍大数据背景下的"地质云"构建理念与方法,以及大数据在地学领域的应用。大数据为非结构化、半结构化的地质数据带来了新的处理方法与理念。地质云的构建旨在探索以需求带动的地质核心数据的应用,挖掘非结构化数据的新数据信息,以支撑国土资源管理决策。

关键词:大数据;地质云;非结构化数据处理
中图分类号:P628 **文献标志码:**A **文章编号:**1671-2552(2015)07-1260-06

Chen J P, Li J, Cui N, Yu P P. The construction and application of geological cloud under the big data background. *Geological Bulletin of China*, 2015, 34(7):1260-1265

Abstract: Geology belongs to the data-intensive scientific. A lot of digital data related to the geological and terrain have been collected and will be update, due to the need of scientific research, to better applied to serve the human activity, the data are more and more examined and expanded constantly, the data´s number is increasing. The informatization strategy of the land resources in China has profound significance. This paper focuses on the establishment of the "geological cloud" under the back ground of the big data and the application of the big data to geological field. Big data technology brings new methods and ideas for unstructured, semi-structured geological data. The establishment of geological cloud purpose to explore the application of the core data of geological, find the new information by data mining from the unstructured data, for decision support in land and resources administration.

Key words: big data; geological cloud; unstructured data processing

信息与网络技术的迅猛发展,不仅改变了人们的日常生活,而且改变了科学家的思维模式[1]。新近国际上诞生了科学研究的第四范式,即数据密集型知识的发现[2]。《易经》讲,太极生两仪,两仪生四象,四象生八卦,八卦演万物,是讲自然之数,逐渐演化到万物之象的道理。《易经的智慧》[3]指出,任何科学的发展都应该遵循自然规律循环,即万物皆有规律,万物皆要遵从自然规律。世界已经进入了科学研究的第四范式时代,即数据挖掘时代,那么也可以理解为任何自然规律都能通过数据来表达。由此可知,认识数据利用的重要性,充分挖掘现有数据,找到其表达的信息和知识,即可以解释其代表的自然规律。

数字地球作为一直以来地质学界乃至整个科学界的热点概念,首先要解决的问题就是信息数字化系统,利用信息技术,将地球表面每一点上的固

收稿日期:2014-04-15;**修订日期:**2015-05-28
资助项目:国土资源部公益性行业科研专项项目(编号:201511079—02)
作者简介:陈建平(1959-),男,教授,博士生导师,从事矿产资源定量评价研究工作。E-mail: 3s@cugb.edu.cn

有信息数字化。这些信息主要是与空间位置直接相关的相对固定信息,例如地形、地貌、植被、建筑、水文、地球内部构造等[4]。

维克托·迈尔-舍恩伯格[5]指出,世界的本质就是数据。从古至今,时代更迭,科技进步,都逃不开本质数据,用数据说话、挖掘有用的数据一直是科学研究的核心。

大数据已经作为现今最热门的技术话题出现在人们的视野中,其理念已在智慧政府、智慧城市的建设中发挥出了巨大潜能,将开启一次重大的时代转型。著名未来学家阿尔文·托勒夫[6]将大数据热情地赞颂为第三次浪潮的华彩乐章。在这样的时代背景下,地球科学研究和地质工作及国家地质资料管理也不能缺少大数据这一关键要素,其对于国土资源研究规划及管理具有极为重要的战略意义。

人类正处在大数据时代的浪潮下,如何应用大数据理念和技术,在地学领域有效地组织和使用庞大的地质大数据,对其进行科学的挖掘,产生更高的价值,以实现相应的服务,具有十分重要的意义。由此,笔者提出了大数据时代背景下地质云的构建与应用。

1　地质云

1.1　构建背景及意义

2011 年 5 月,美国麦肯锡公司全球研究院发布《大数据:创新、竞争和生产力的下一个新领域》[7],将大数据从技术圈带入市场乃至整个社会。此后,各发达国家陆续发布大数据相关计划。

美国将大数据看作是“未来的石油”,投资 2 亿美元启动大数据研究和发展计划,并将这一计划上升到了国家高度。

在欧盟,相关报告[8]指出,欧盟公共机构的产生、收集或承担的地理信息、统计数据、气象数据、公共资金自助研究项目、数字图书馆等数据资源的全面开放,预计每年将会给欧盟带来 400 亿欧元的经济增长,欧盟认为大数据是促进经济增长的重要力量。

英国 CEBR(经济与商业研究中心)[9]2012 年通过研究进一步证实了大数据的经济价值,预测 2017 年大数据经济价值将达到 407 万英镑。

韩国提出“智慧首尔 2015”计划[10],“首尔开放数据广场”是开放性的数据中心,已有 33 个数据库,880 个数据集,为用户提供十大类的公共数据信息,包括公共交通路线、巴士到站时间、停车位、各地区天气预报等。韩国认为公共数据已成为具有社会和经济价值的重要国家资产。

当代是一个大规模生产、分享和应用数据的时代,人类已经处在一场数据革命中。而 IBM、亚马逊、谷歌、微软等大型企业抓住了这场数据革命的机遇,已经成功将大数据技术应用到各种层面,继 2012 年 Google 应用大数据成功预测了流感后,发展到今年对世界杯往届数据采用 Big Query 进行分析,成功预测了本届世界杯八强名单。马云的阿里巴巴于今年 7 月 8 日正式开放了商用 ODPS(Open Data Processing Service),它是一项 Web 服务,不用花大价钱建立数据中心,就能分析海量数据。100PB 的数据任务可以在 6 个小时内完成处理[11]。企业已成为了这场变革的推动力量。

2012 年底中国科学院院长白春礼[12]在一次论坛上提出,中国制定国家大数据战略的主要内容包括:构建大数据研究平台,即国家顶层规划,整合创新资源,实施专项计划,突破关键技术;构建大数据良性生态环境,制定支持政策、形成行业联盟、制定行业标准;构建大数据产业链,促进创新链与产业链有效嫁接。

地质大数据应用研究是国家大数据战略的组成部分。国土资源“十二五”科学和技术发展规划指出,加强 3S 技术、网络技术、云计算、物联网、数字地球等技术的跟踪和应用研究,加强地质资料信息的开发利用技术研究,开展地质资料分级分类的服务和互联共享机制研究,研发面向重点成矿区带、重点经济区、生态环境脆弱区、重大工程建设区和重大地质问题区的地质资料信息整合、深度加工、服务产品开发技术。

信息网络领域的发展使大数据存储取得突破,云计算、物联网、工业互联网等技术的兴起,使信息技术渗透方式、处理方法和应用模式发生变革,使地质研究中多系统联合与结合成为可能,实现由数字地质向智慧地质的转变。

地质云架构与应用就是充分利用地质、矿产、地球物理、地球化学、遥感、地形、地貌、植被、建筑、水文、灾害等地表每一点上的数字化地质数据(结构化和非结构化数据),以大数据背景下的地质数据应用与服务为主线,以大数据技术和产业发展为指

导,以挖掘地质信息资源为目的,在数据采集、资源整合、数据传输、信息提取、知识挖掘等相关技术研发与集成基础上,建设地质大数据环境和地质大数据平台,构建地质云,实现从数据到信息,从信息到知识,从知识再到智慧的地质大数据转换,为国土资源科学化管理、找矿突破战略行动和社会化服务提供数据分析、挖掘、组织、管理等服务,开展适用于政府决策、科学研究、企业服务,以及社会公众用户对地质数据的多层次、多角度、多目标的示范应用。

1.2 核心主线

地质云的构建贯彻2条主线,即以需求为核心目标,以数据链技术流程和大数据技术方法为手段。

(1)数据需求

地质行业历史悠久,地质资料积累十分丰厚。随着信息化技术的发展,地质行业的需求和信息技术引导着地质工作的发展。2006年赵鹏大院士[13]在成都理工大学一次报告会上第一次使用了“数字地质”这个名词。并于2013年在新疆召开的第十二届全国数学地质大会上做了“大数据时代的数字地质”报告。大数据时代的来临,为地学领域提供了跨越式发展的重要机遇[14]。

地质资料经过长期积累,种类、数量不断增长,包括各类电子文件,结构化、半结构化、非结构化的数据,以及文档、图件、数据库(图件数据库、空间数据库、属性数据库等)、图片、表格、视频、音频等。如果没有“需求”为指引,大数据相当于淹没在数据中,先要明确“需求”,然后拥有大规模处理能力,接下来才是数据挖掘、算法和分析,最终数据才能产生价值。

地学领域的大数据,对不同层次的人员有着不同的需求,包括针对社会公众个人的地质资料社会化服务的公众需求,以及针对地质科研机构、相关国有企业及政府部门的专业数据需求。针对不同的需求,应用大数据分析方法,挖掘数据中的价值,是以需求带动、以大数据为手段开展各项研究。

(2)数据链技术流程方案

按照大数据分析的技术路线(图1),结合云计算、物联网等强大、先进的网络体系和大数据的基本处理流程,形成一个完整的从数据到信息,从信息到知识,从知识到服务的数据链。①数据化:数字化、定量化及行为化产生数据(核心数据脱密,非密集数据整合、公共域上获取的信息热点需求);②信息化:数据的解析化、集成化、综合化产生信息;③知识化:信息的模型化、智能化和专业化产生知识和产品;④服务:知识和产品的实用化、网络化和可视化产生财富和效益,并服务于公众和社会,满足国土资源管理决策服务社会大众的需求;⑤再循环:在服务公众和社会过程中,又产生大量新的数据。

以构建中国的地质云,紧密围绕需求分析实现数据链循环(图2)为技术路线。在确定数据类型、数据模型的基础上,采集和整合数据;组合关键技术实现数据驱动,以信息分析和数据挖掘的方法和技术实现模型驱动;根据应用需求选择概念模型或数字模型,筛选相关变量,进行综合信息分析;最终结合专业知识,数字分析的凝炼提升为知识驱动;最后将获取的数字知识,用于解决实际问题,经过

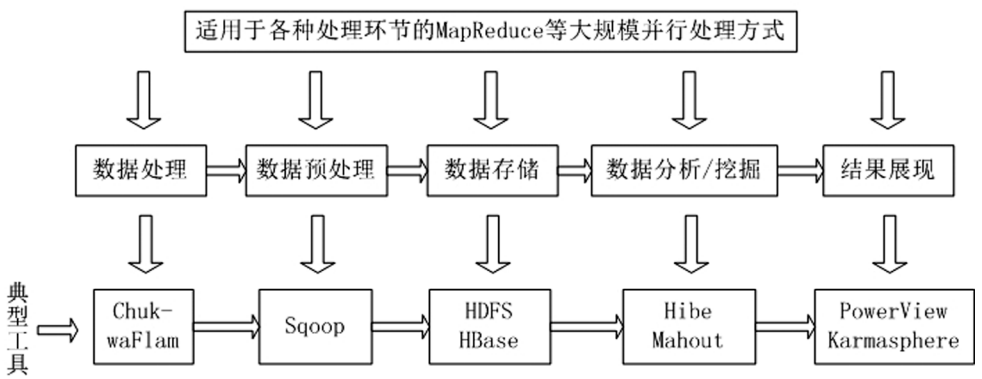

图1 大数据处理流程

Fig. 1 The processing flow of big data

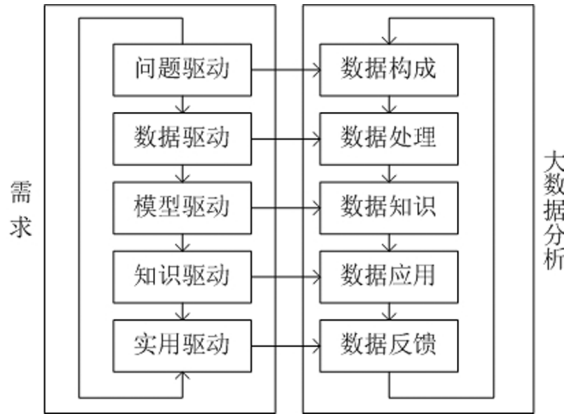

图2 “数据链”构成及循环

Fig. 2 The structure and circulation of "data link"

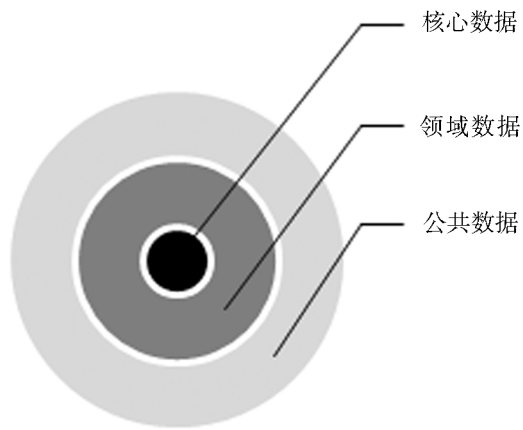

图3 “地质云”数据构成

Fig. 3 The data structure of "geological cloud"

实践检验,进入实用驱动,再获取新数据,如此形成数据的循环和深入发展。

1.3 地质云构建

构建中国的地质云与当前基于互联网、物联网的大数据分析不同。有学者[14−15]指出,不同领域,不同问题的数据性质和类型各不相同,要收集好,处理好,分析好,解释好不同领域的数据,必须深刻了解数据的特性。地质学属于数据密集型科学,地质数据具有多源、多元、异构、时空性、方向性、相关性、随机性、模糊性、非线性等特征。因此,地质云具有专业性和一定的保密性,针对的是长期积累的大量的地质数据。对于地质灾害、地质环境等问题还有大量的实时数据。地质云(图3)包括核心基础数据,即已建成的结构化数据库、邻域数据及公共数据3个部分。所以,要利用好已有的传统的结构化数据,同时利用大数据技术手段处理相关的非结构化数据,还要利用外围公共数据。

与长期基于地质数据库的信息分析不同,地质大数据具有多样性,存在多维的、结构化与非结构化数据并用等特点,大数据分析的技术方法也与以往专业数据库的应用需求大不相同。另一方面,地质云的构建与当前主流的物联网大数据分析也不尽相同,长期的地质调查与地学研究与多年地质信息化工作的积累,形成了内容丰富、专业性强的数据基础,是中国国土资源科学管理、地质大调查和地学信息公众服务的重要基础保证,这种“专业云”客观地需要专业局域网的构建、数据共享平台的搭建(图4)、地质大数据的可视化服务等技术研发又是物联网大数据分析所不能包含的。因此,地质云构建紧密围绕国土资源管理、部署决策和公众服务的应用需求,研发的关键技术包括:非结构化数据的信息提取与挖掘分析,结构化与非结构化数据混合存储与管理、大数据共享平台、数据传输、可视化等。

2 大数据在地学领域的应用

(1)大数据技术和理念在地质图件中的应用

传统的地质图件图像处理思路是矢量化,现在有大量的地质资料都是电子档(包括扫描件、已经矢量化的图、附图、附表、报告),一幅完整的图件图像,通常包括图例、比例尺、图名等要素,并用不同的颜色、不同的花纹来表示相应的地质构造或区域等。这些长期积累的地质图件、图像资料数量庞大。目前,全国地质资料馆馆藏资料约12万6千种。电子地质资料约10万档。包括结构化、半结构化、非结构化的文档、图件、数据库、图

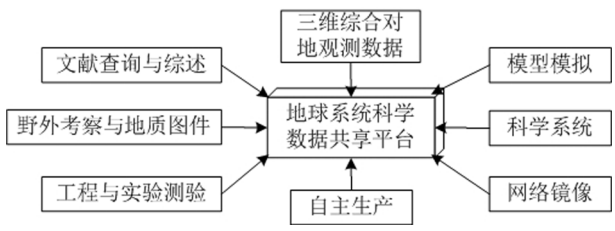

图4 地球科学数据共享平台示意图(据参考文献[16]修改)

Fig. 4 Schematic diagram of sharing platform for earth science data

片、表格、视频、音频等。

这些地质资料采用中国科技档案通用的卷方式保存，案卷级目录无法详细地显示每件资料包含的信息。国土资源管理、研究等专业人员面临着获取有用信息的窘境，面对数量庞大的地质资料，获取和发现有用信息的难度越来越大。

大数据时代的来临，让笔者对这些资料的整理处理方法有了新的思路。对于非结构化的图件，可以建立基于元数据的计算机自动描述动态模型，通过对地质资料进行统一标识图号、图名等，对图件按图例颜色，或花纹符号进行匹配，并按照图例分图层。以地质资料中的图件、插图、附图及表格为研究对象，通过大数据相关技术，进行并行快速的处理。这样可以初步实现由非结构化数据转化成为半结构化数据、结构化数据，并可以将数据进行数据化；按照数据链思路，进一步进行统计分析，从中提取信息；接下来建立应用模型分析，形成知识体系；并最终实现地质资料的服务。

由此可知，不同的路径与技术流程最终得到相同的结果。传统的处理方式是针对结构化数据的，而大数据的引入对非结构化的数据处理有了新的方法。

(2)大数据技术和理念在地质文本类资料中的应用

地质云的核心数据包括地质资料馆的馆藏资料，有大量的非结构化的文档类资料。如何在这些资料中高效地挖掘和提取信息，是当前地质资料社会化服务的重点研究方向。围绕需求和数据链这2条主线，应用大数据技术方法和理念，可以展开对非结构化数据的信息挖掘与提取工作。例如用汉字识别系统提取，识别过程中针对不同段落内容标记属性，加锚点，依次匹配与属性相同的内容和段落，如构造、地貌等。可以将非结构化数据转化为半结构化的、结构化的数据进行处理，对自动标引和自动摘要的研究有了新的思路和方法。地质资料从专业角度可以分为基础地质、矿产、物探、化探、遥感等方面，针对不同类型的地质资料，从专业角度分析不同类型、类别文本资料的组成结构，以及不同组成部分与主旨内容的相关性，在此基础上建立不同结构的动态描述模型，并结合主旨内容赋以不同的权重，为下一步自动标识和摘要选择标引源和语义、语法的分析奠定基础。

3 结 语

本文旨在构建地质云，探索以需求带动的核心数据的应用，结合大数据理念与相关技术方法，挖掘非结构化数据的新数据信息，以满足国土资源管理决策的需求。

地质云的构建是一个长期的系统工程。本着“立足现实，着眼未来”和“从长远和全局着眼，从当前和局部入手”的基本原则，按照大数据分析的技术路线，逐步实现地质云公共数据及核心数据的分析与应用，最终完成地质云总体架构[17-19]①。

地球作为一个整体，无论是人类的行为、气候的变化、各种资源的开发与利用，还是自然灾害和环境污染及生态系统的循环，都是国土资源管理应该涵盖的。大数据的引入，可以将这些资源信息整合起来，提供统一调整整个地球信息资源的问题的功能，对于国土资源管理战略规划具有不容小觑的作用。

美国地质调查局(USGS)于2012年6月，在其官网正式发布了《美国地质调查局核心科学体系科学战略(2013—2023)》，即10年发展规划，提出了核心科学体系战略的要点，以促进对复合性地球系统的综合性描述与认识。描述出了一个概念模型和框架。方案集中于7个科学使命领域：气候变化、土地利用变化、核心科学体系、生态系统、能源和矿产、环境健康、自然灾害和水。

地质云是模块式核心科学体系构架的重要组成部分，是中国基于地质数据研究方面的核心实力。依托计算机和信息科学，以基础性和应用性调查为基础，通过跨学科的数据综合、学科合成及新的调查方法，形成合成产品，并建立一种自然而然的整合新数据、新应用和其他科学产品的工作流程，最终目的是更好地表述和认识复杂的地球系统和地质格架，为描述中国的陆地表面和生物多样性特征，提供科学依据，使得有关部门有能力处理亟待解决的复杂社会问题。

2013年，越来越多的企业大数据项目走出概念验证阶段，进入了生产和实施阶段，数据分析将被运用到运营及决策步骤中，从2014年开始，大数据[20-22]将走向实用化阶段，更多的应用将被实现。笔者认为，地质云在此背景下构建具有实际意义，它为地质资料提供了有效的数据分析方法和理

念,通过挖掘地学领域的数据、提取信息、凝炼成知识,能够更好地为公众、专业人员、管理决策人员服务。

致谢:成文过程中得到了中国科学院院士赵鹏大先生的帮助,在此表示感谢。

参考文献

[1]中共中央办公厅, 国务院办公厅. 2006—2020年国家信息化发展战略[EB/OL](2006-12-14)[2014-04-13]http://www.itsec.gov.cn/export/sites/itsec/standard/relation/7a0a1b35-9e5b-11e0-82c2-755a4519629 1/. 2006.

[2]杨宗喜, 唐金荣, 周平, 等. 大数据时代美国地质调查局的科学新观[J]. 地质通报, 2013, 9: 1337-1343.

[3]曾仕强. 易经的智慧[M]. 西安: 陕西师范大学出版社, 2011.

[4]赵鹏大, 马连杰. 数字地球与全球战略——21世纪谁主沉浮[M]. 武汉: 中国地质大学出版社, 2000.

[5]维克托·迈尔-舍恩伯格, 肯尼思·库克耶. 大数据时代[M]. 杭州: 浙江人民出版社, 2013.

[6]阿尔文·托夫勒. 第三次浪潮[M]. 北京: 中信出版社, 2006.

[7]孙爱鸿. 大数据——下一个新领域[EB/OL](2011-06-11)[2014-04-13]http://www.infoq.com/cn/news/2011/06/BigData/. 2011.

[8]朱雪征, 李莉. 欧盟空间数据基础设施规划研究[J]. 测绘通报, 2010, 8: 7-10.

[9]新华社.英国"尝鲜"大数据时代[EB/OL](2013-05-19)[2014-04-13]http://www.farmer.com.cn/kjpd/dtxw/201305/t20130521_844136.htm. 2013.

[10]廖瑾. 来自"智慧首尔2015"的启示[J]. 上海信息化, 2012, 1: 21-23.

[11]财新网. 阿里大数据走向变现[EB/OL](2014-07-15)[2015-02-15]http://www.36dsj.com/archives/9805. 2014.

[12]白春礼. 制定国家大数据战略[EB/OL](2012-12-30)[2014-04-13]http://www.chinanews.com/gn/2012/12- 30/4449013. shtml. 2012.

[13]赵鹏大. 数学地质与矿产资源评价[J]. 地质学刊, 2012, 3: 225-228.

[14]赵鹏大. 大数据时代的数字地质[C]//中国数学地质大会, 2013.

[15]滕艳, 张地珂.大数据时代需要重视数学地质研究[N]. 国土资源报, 2013-03-15.

[16]沈慧. 数描地球,应用无限[N]. 经济日报, 2014-03-31, 15版.

[17]Manegold S, Kersten M. Big Data[J]. ERCIM News, 2012, 89: 33-36.

[18]李国杰. 大数据研究的科学价值[J]. 中国计算机学会通讯, 2012, 8(9): 8-15.

[19]中国国土资源报信息中心. 服务为本——我国地质资料社会化服务综述[EB/OL](2008-09-03)[2014-04-13]http://www.mlr.gov.cn/xwdt/jrxw/200809/t20080903_109768.htm. 2008.

[20]中国社会化服务营销提供商译制. 信息图: 大数据及2014数据分析趋势[EB/OL](2014-03-13) http://www.36dsj.com/archives/6701.

[21]曹建菊. 2014年我们可以期待大数据和云计算的发展[EB/OL](2014-03-27)http://www.36dsj.com/archives/6941. 2014.

[22]李国杰. 大数据的三件事[EB/OL](2013-12-05)[2014-04-13]http://tech.163.com/13/1205/09/9FARPA5K00094OB0.html.2013.

①中国地质调查局发展研究中心. 美国地质调调查局核心科学体系科学战略(第二版). 地调情报总第37期(内部出版). 2012.