收稿日期：2017-11-26
改回日期：2017-12-21

基金项目：中国地质调查局地质调查项目“全国地质资料汇聚与数据整理”（121201004000150018）资助。

doi: 10.12029/gc2018Z109

论文引用格式 : 高学正 , 李晨阳 , 吴轩 , 孔昭煜 , 商云涛 , 齐钒宇 , 贾丽琼 , 李晓蕾 , 郭慧 . 2018. 全国地质资料馆数字地质资料馆主站日志数据集 [J]. 中国地质 , 45(S1):107-114.
数据集引用格式 : 高学正 ; 李晨阳 ; 吴轩 ; 孔昭煜 ; 商云涛 ; 齐钒宇 ; 贾丽琼 ; 李晓蕾 ; 郭慧 . 全国地质资料馆数字地质资料馆主站日志数据集 (V1). 中国地质调查局发展研究中心 ; 全国地质资料馆 [创建机构], 2014. 全国地质资料馆 [传播机构], 2018-09-10. 10.23650/data.H.2018.P4; http://dcc.ngac.org.cn/geologicalData/rest/geologicalData/geologicalDataDetail/8adaeff963f39bf80163f6c88a4b002c

全国地质资料馆数字地质资料馆主站日志数据集

高学正 [1,2]　李晨阳 [1,2]　吴轩 [1,2]　孔昭煜 [1,2]　商云涛 [1,2]
齐钒宇 [1,2]　贾丽琼 [1,2]　李晓蕾 [1,2]　郭慧 [1,2]

（1. 中国地质调查局发展研究中心，北京 100037 ；2. 全国地质资料馆，北京 100037）

摘要：为更加准确的摸清地质资料用户的需求，消除资料使用者与资料管理者之间的信息鸿沟，全国地质资料馆开展了数字地质资料馆主站访问日志的数据收集工作。数据集采用计算机自动记录的方法，将访问者的所在地、关键词、IP 地址等信息进行完整记录。为了更好的利用这些数据，采用了规范的数据处理方法和质量控制体系。数据集提供了 2014—2017 年数字地质资料馆主站的访问记录，有效地反映了访问者在进行地质资料获取过程中的行为习惯，可为日后地质资料网站建设、地质资料开发利用、地质资料的管理与服务提供依据。

关键词：数字地质资料馆；网站；日志

数据服务系统网址：http://dcc.cgs.gov.cn

1　引言

全国地质资料馆是国家地质行业中馆藏最多、最全的专业技术档案馆。目前已收藏了 14 万余种地质资料（王新春等，2016），包括区域地质调查资料，矿产勘查勘探资料，海洋地质报告资料，地球物理、地球化学和遥感地质勘查资料，水文地质、工程地质和环境地质勘查资料，地质科学研究成果报告，技术方法研究资料等（高学正等，2016）。全国地质资料馆数字地质资料馆，是以数字资源为主的地质资料馆，是在全国地质资料馆传统工作方式与环境的基础上，以现代信息技术改造传统工作流程，向社会提供地质信息服务的国家级地质信息资源基础设施（孔昭煜等，2016）。此项工作自 2012 年开始筹划并启动建设，历经 6 年的发展与完善，目前可以向社会提供多种服务，包括地质资料目录服务、元数据服务、地质图服务、专题服务、互联互查服务等内容。

在数字地质资料馆运行过程中，记录了大量的用户访问日志信息，利用这些信息可以更加充分了解用户的需求，捕捉用户的访问习惯，及时了解用户对网站的认知度、满意度，保障系统安全稳定运行（赵国洪，2010），使管理人员能够准确掌握

第一作者简介：高学正，男，1987 年生，工程师，主要从事地质资料管理及服务研究工作；E-mail：gxuezheng@mail.cgs.gov.cn。

网站的发展状况，为未来提出更为具体的发展途径（谢晓萍，2016），改善地质资料服务质量，加快其向高认知度和高满意度的方向发展，解决全国地质资料馆藏机构目前存在的资料服务管理方式单一、服务效率较低、地质资料获取成本高、电子地质资料未充分利用等制约地质资料发展的一系列问题，使网站建设向智慧化、精准化、主动化的方向发展（于施洋等，2013）。

数字地质资料馆主站日志数据集元数据简表见表 1。

表 1　数据库（集）元数据简表

条目	描述
数据库（集）名称	全国地质资料馆数字地质资料馆主站日志数据集
数据库（集）作者	高学正，中国地质调查局发展研究中心（全国地质资料馆） 李晨阳，中国地质调查局发展研究中心（全国地质资料馆） 吴　轩，中国地质调查局发展研究中心（全国地质资料馆） 孔昭煜，中国地质调查局发展研究中心（全国地质资料馆） 商云涛，中国地质调查局发展研究中心（全国地质资料馆） 齐钒宇，中国地质调查局发展研究中心（全国地质资料馆） 贾丽琼，中国地质调查局发展研究中心（全国地质资料馆） 李晓蕾，中国地质调查局发展研究中心（全国地质资料馆） 郭　慧，中国地质调查局发展研究中心（全国地质资料馆）
数据时间范围	2014—2017 年
地理区域	全国范围
数据格式	*.accdb
数据量	1.70 GB
数据服务系统网址	http://dcc.cgs.gov.cn
基金项目	中国地质调查局地质调查项目“ 全国地质资料汇聚与数据整理 ”（编号：121201004000150018）
语种	中文
数据库（集）组成	本数据集两部分组成，分别为：数字地质资料馆网站访问 IP 地址记录数据表 .accdb、数字地质资料馆网站搜索关键词记录数据表 .accdb

2　数据采集和处理方法

2.1　数据来源

数字地质资料馆主站日志数据均来自记录于数字地质资料馆网站的后台数据库，全部为全天实时自动记录。内容包括所有访问用户从登录网站、查询检索、浏览资料、专题查询、地质图调阅等一系列过程。数据集中详细记录了访问用户的 IP 地址、搜索的关键词、访问时间、用户类型、用户所在地等信息，并且可以区分出用户查阅的是地图信息还是文档信息[1]。数字地质资料馆服务模式及数据获取流程如图 1。

2.2　数据处理和应用

数字地质资料馆主站日志数据集经过数据记录、数据提取、入库等系统处理流程。首先，记录下网站访问用户的信息及操作记录，通过结构化的数据语句对所记录的数据开展数据项的提取工作，随后将其按照统一的数据格式存入数据库中。

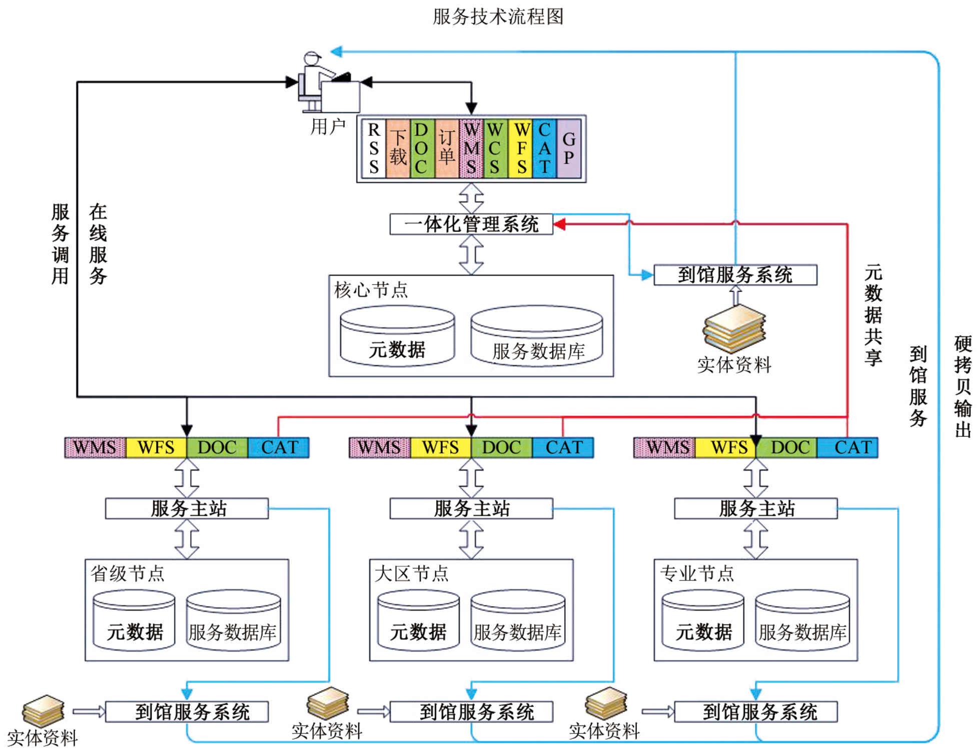

图 1　数字地质资料馆服务模式及数据获取流程

由于文本格式的数据并不能直观展现数据集内各字段中内容的变化情况，所以通过绘制国内访问用户区域分布图（图 2）、国外访问用户区域分布图（图 3）、用户检索关键词统计图（图 4）对数据样本进行展示。

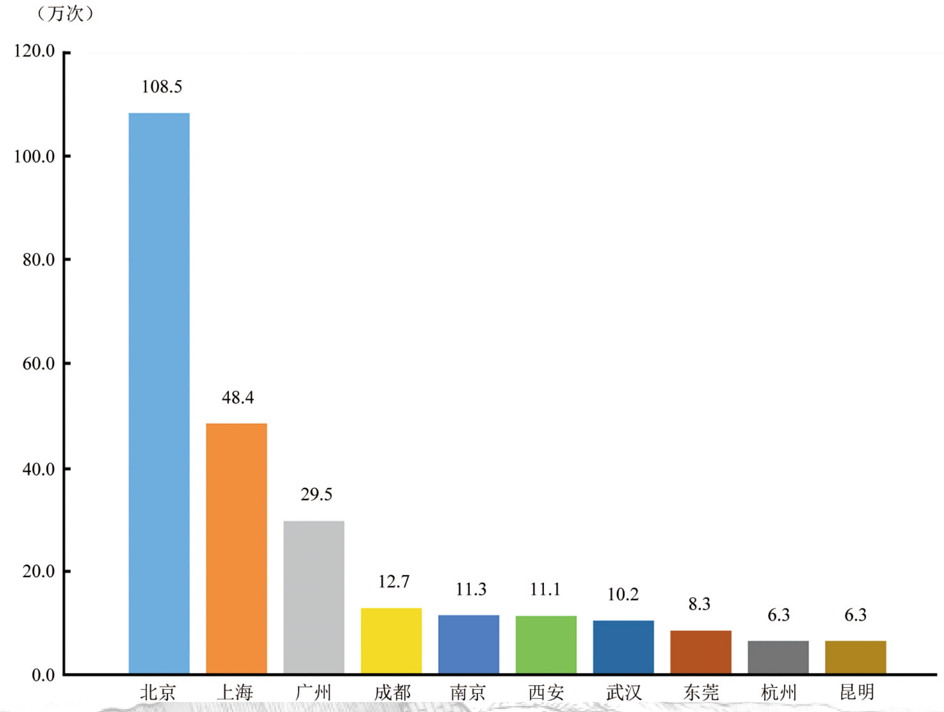

图 2　国内访问用户区域分布图

对本数据集中国内外访问用户的地理位置的分布进行统计，总访问次数达到529.5万次（5 294 535次），除去内部局域网的访问，总访问次数达到448.7万次（4 486 616次），其中，北京、上海、广州等一线城市分别位于访问前三位，总访问量达到186.4万次（1 864 312次）。由图中可知，长三角、珠三角地区访问用户较为集中，西部地区访问量较少。

由统计数据可知，国内的上海、东莞、杭州等地质工作开展频度较低的城市，却对数字地质资料馆网站的访问量和检索量较多。相反，传统的地质工作大省如山东省，访问最多的青岛市的排名仅仅名列国内第15位。主要原因在于，一是数字地质资料馆网络服务宣传力度不够，很多从事实际地质工作的人员对于数字地质资料馆网站及其发布的资料内容不了解，导致传统的地质大省所使用数据不多。二是我国一线城市及长三角、珠三角地区的信息技术较为领先，对于网络的敏感程度高，导致了其访问量较高。三是由于数字地质资料馆主要提供的是地质资料的检索与浏览，目前还没有提供资料的下载功能，用户不能获取相关资料，还不太利于资料的利用。此外，在实际的地质工作中尤其是野外地质工作中，若想利用数字地质资料进行地质工作，需要有网络的配合才能所发挥作用，但目前我国野外地质工作地区大多不具备连接互联网的条件。

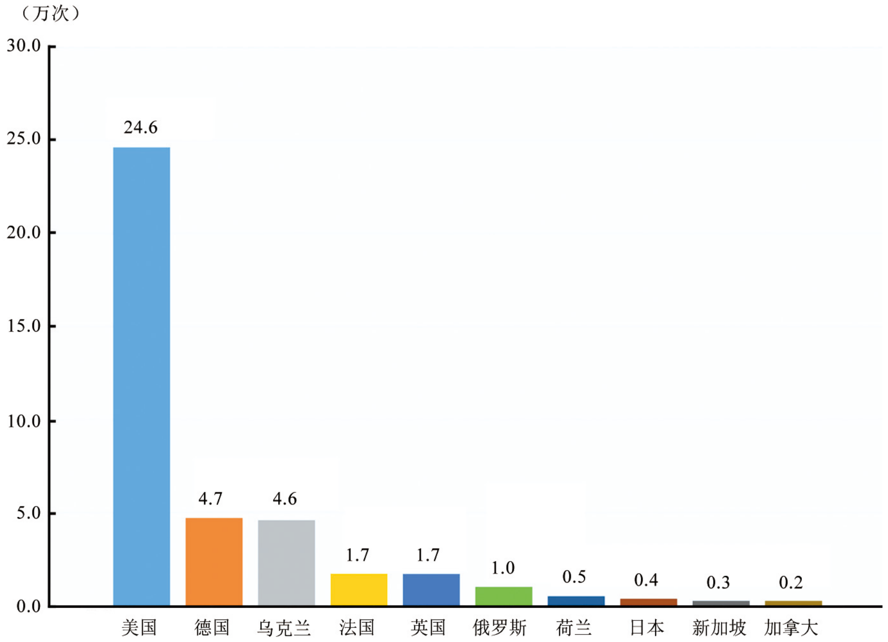

图3 国外访问用户区域分布图

国外访问国家中，美国的访问量达到24.6万次（246 333次），排在总访问量第四位。德国访问量4.7万次（47 091次），排在总访问量第15位。乌克兰访问量4.6万次（45 927次），排在总访问量第18位，其他的国家像英国、法国、俄罗斯、荷兰、日本、新加坡、加拿大、意大利、阿联酋、瑞典、巴西、芬兰、澳大利亚等国的访问量也都在千次以上。

表 2　国外访问国家关键词排序

排名	国家	关键词 1	关键词 2	关键词 3	关键词 4	关键词 5
1	美国	1:20 万地质图	1:25 万地质图	1:50 万地质图	1:20 万水文地质图	H45C003004
2	德国	1:25 万地质图	1:20 万水文地质图	1:20 万地质图	涞水	中国地下水资源分布图
3	乌克兰	无	无	无	无	无
4	法国	1:20 万地质图	仙居	江永	齐齐哈尔	全国 1:20 万重力资料分布图
5	英国	1:20 万地质图	oilfields	地质图	大庆	水文
6	俄罗斯	1:25 万地质图	鄂尔多斯盆地	鄂尔多斯	河北	山西
7	荷兰	1:25 万地质图	H4820	1:20 万地质图	1:20 万水文地质图	中国
8	日本	1:20 万水文地质图	1:20 万地质图	钨矿	1:50 万地质图	1:25 万地质图
9	新加坡	铝土矿	1:20 万地质图	广西	1:50 万地质图	铝矿
10	加拿大	西安	山西	地下水	夹皮沟	1:25 万地质图

由表 2 可知，在国外的 IP 地址访问中，“ 1∶20 万地质图 ”、“ 1∶25 万地质图 ”成为了最集中的检索关键词。国外用户访问还较为集中的关键词包括：1∶20 万水文地质图、1∶50 万地质图、山西等。其中，IP 地址为乌克兰的访问存在异常，并没有产生搜索关键词，但存在集中不断访问的情况。

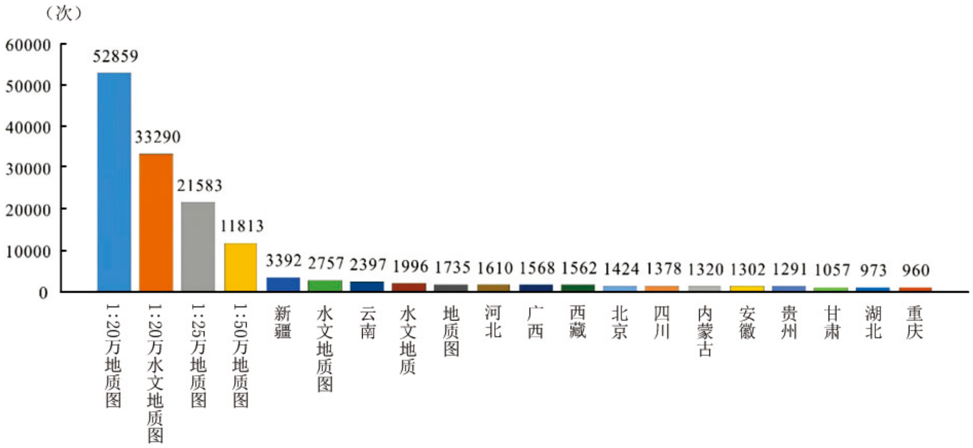

图 4　用户检索关键词统计图

值得注意的是，既不是地质大省也不是强省的河北成为了第三被检索的地位，分析认为，这与 2017 年中共中央、国务院决定设立河北雄安新区有重要关系。新区设立后，国家各部门随即开始编制新区规划，中国地质调查局也实地进行了地质调查，这些都加大了用户对于河北省的关注，同时提高了用户对于河北省的关键词检索。由此可见，用户的检索行为与需求和国家的政策方针紧密挂钩，与时事热点紧密联系，在今后的工作中，要更加针对用户的需求推出高质量的服务产品。

3　数据样本描述

全国地质资料馆数字地质资料馆主站日志数据集，为 Access 数据库型数据，共包括两个 Access 数据库文件，分别为数字地质资料馆网站访问 IP 地址记录数据表（表 3）、数字地质资料馆网站搜索关键词记录数据表（表 4）。其中，数字地质资料馆网站访问 IP 地址记录数据表记录了 2014—2017 年数字地质资料馆访问者的 IP 地址、访问的资料内容、是否为实名注册用户、用户访问所在地以及访问时间等数据项内容。数字地质资料馆网站搜索关键词记录数据表记录了 2014—2017 年数字地质资料馆访问者的 IP 地址、检索的关键词、检索时间等数据项内容。

表 3　数字地质资料馆网站访问 IP 地址记录数据表

序号	字段名称	数据类型	实例
1	唯一字符串	短文本	2293B1F0579C7019E05341015A0A617B
2	IP 地址	短文本	14.215.222.217
3	访问路径	长文本	/Data/FileList.aspx?MetaId=E928A0F55D2F7A73E0430100007F3D67&type=zw
4	用户账户	长文本	631795983@qq.com
5	用户所在地	长文本	中国 广东 佛山
6	访问时间	日期 / 时间	2015-10-21 9:21

表 4　数字地质资料馆网站搜索关键词记录数据表

序号	字段名称	数据类型	实例
1	IP 地址	短文本	59.71.224.2
2	关键词	长文本	浙江湖州
3	访问时间	日期 / 时间	2015-10-17 19:16
4	唯一字符串	短文本	224C06EB72A00920E05341015A0A506A
5	用户账户	短文本	匿名用户
6	用户所在地	长文本	中国 湖北 武汉 中国地质大学

4　数据质量控制和评估

本次全国地质资料馆数字地质资料馆网站日志数据集包含了数字资料馆开始进行数据日志记录以来的所有数据，数据集中的两个数据表格共记录数据 603 万条（6 034 025 条）。其中，由于计算机机房进行改造等原因停电两次，造成主站日志记录不完整，时间分别为 2017 年 3 月 17 日 18:40 至 3 月 18 日 20:10，2017 年 10 月 5 日 14:37 至 10 月 7 日 12:06，但不完整的记录只占总体记录的极小比例，不影响数据集整体的完整性、可靠性、适用性、准确性等性能指标。数据采集的具体工作过程为：从用户进入网站开始，数据日志即开始记录用户的 IP 地址、访问地、访问内容等信息，将记录的数据按照指令进行预处理，查询是否存在记录异常、非预期执行等行为，按时间顺序进行入库处理，保证入库数据的准确性与有效性。为保障数据库安全稳定运行，制定了数据库定期

更新与维护策略并严格执行，定期对数据库内容进行评估。在数据提取分析过程中，对数据项进行格式的统一，确保数据内容的整体一致。针对数据质量问题，还需继续完善数据集的更新与维护机制，增强现有日志的信息量，加强日志数据的版本标注，切实提高数据质量。

5　访客分析

在检索关键词方面，综合国内和国外用户，经过对数据集中用户关键词进行统计发现，用户检索前五位分别为1∶20万地质图（52 859次）、1∶20万水文地质图（33 290次）、1∶25万地质图（21 583次）、1∶50万地质图（11 813次）、新疆（3 392次）。“地质图”成为了用户检索的主要内容，在前五位中，“地质图”的检索比例达到97.2%，成为绝对主力，而1∶20万地质图的关键词检索量占到了前五位中的43.0%，所有关键词检索量的7.2%，成为了用户检索最多的关键词。

在地质访客方面，我国一线城市及长三角、珠三角等发达城市的访问量较高，西部地区访问量较少。综合来看，我国地质用户访问较为正常，而从很多国外用户的访问存在异常行为，从检索关键词、注册信息、检索语种以及检索频度来分析，这些检索行为更多的被认为是国内的代理机构为了掩饰真实地址而将服务器到国外进行跳转。因此，在网络访问的国外用户中，不排除具有恶意索取资料、窃取资料、抓取数据的嫌疑。

6　结论

（1）数字地质资料馆使用信息系统对传统的地质资料馆进行了全面改造与升级，其工作范式从文献处理范式转向数字技术处理范式，从以纸为主兼顾电子转向纸电并举。以电为主，在我国地质资料馆藏机构的资料管理与服务领域中实现了创新，业务模式和组织形式也发生了重大变革。

（2）全国地质资料馆数字地质资料馆主站日志数据集是一种记录馆藏数字资源访问的数据集，其记录的相关内容可以实现对地质资料网络访问用户的全流程跟踪。通过统计和关联的分析可以找出访问用户、访问时间以及访问内容等多要素之间的内在联系，为进一步追踪用户需求，了解用户习惯提供很好的数据支持。

（3）通过数据集内容分析，针对用户的兴趣点，通过制作资料服务专题，改善网站内容排序等方式，使数字地质资料馆的服务水平更进一步，达到精准、高效、快捷的服务目标。

（4）此数据集的记录内容、工作方法以及数据分析样本可以为各馆藏机构的服务网站提供借鉴，可为我国地质资料服务部门社会化服务改进，提高服务质量提供有效参考。

致谢：全国地质资料馆数字地质资料馆主站日志数据集的取得离不开项目组所有人员的共同努力，在数字地质资料馆系统设计与建设过程中，得到了中科软科技股份有限公司、图锐（北京）信息技术有限公司、Esri中国信息技术有限公司的大力支持，在此深表谢意。

注释：

[1] 中国地质调查局发展研究中心 . 2013. 全国地质资料馆核心数字系统研发成果报告 .

参考文献

高学正 , 孔昭煜 , 齐钒宇 , 商云涛 , 李晓蕾 , 贾丽琼 . 2016. 全国地质资料馆馆藏资料现状及服务趋势研究 [J]. 中国矿业 , 25(2): 73−76.

孔昭煜 , 商云涛 , 高学正 , 齐钒宇 , 李晓蕾 , 贾丽琼 . 2016. 国家地质资料数据中心构建研究 [J]. 中国矿业 , 25(2): 92−96.

王新春 , 齐钒宇 , 李晓蕾 , 高学正 . 2016. 资料数据集成与服务研究——以整装勘查区地质工作为例 [J]. 中国地质 , 43(2): 691−697.

谢晓萍 . 2016. 网站访问统计在档案用户研究中的应用分析 [J]. 兰台世界 , 18: 31−33.

于施洋 , 王建冬 . 2013. 政府网站分析进入大数据时代 [J]. 电子政务 , 8: 79−85.

赵国洪 . 2010. 政府门户网站访问量影响因素分析——基于中日政府网站的比较 [J]. 电子政务 , 5: 62−68.

Received: 26−11−2017
Accepted: 21−12−2017

Fund project:
Geological Survey of China Geological Survey "National geological data collection and data consolidation" (121201004000150018) funded.

doi: 10.12029/gc2018Z109

Article Citation: Gao Xuezheng, Li Chenyang, Wu Xuan, Kong Zhaoyu, Qi Fanyu, Shang Yuntao, Jia Liqiong, Li Xiaolei, Guo Hui. 2018. The dataset of user's accessing log to Digital Geological Library of National Geological Archives of China [J].Geology in China,45(S1):151−159.

Dataset Citation: Gao Xuezheng; Li Chenyang; Wu Xuan; Kong Zhaoyu; Qi Fanyu; Shang Yuntao; Jia Liqiong; Li Xiaolei; Guo Hui. The dataset of user's accessing log to Digital Geological Library of National Geological Archives of China (V1). Development and Research Center of China Geological Survey; National Geological Archives of China[producer], 2014. National Geological Archives of China[distributor], 2018−09−10. 10.23650/data.H.2018.P4; http://dcc.ngac.org.cn/geologicalData/rest/geologicalData/geologicalDataDetail/8adaeff963f39bf80163f8570b2b0069

The Dataset of User's Accessing Log to Digital Geological Library of National Geological Archives of China

GAO Xuezheng[1, 2] LI Chenyang[1, 2] WU Xuan[1, 2] KONG Zhaoyu[1, 2]
QI Fanyu[1, 2] SHANG Yuntao[1, 2] JIA Liqiong[1, 2] LI Xiaolei[1, 2] GUO Hui[1, 2]

(1.*Development and Research Center of China Geological Survey*, *Beijing* 100037, *China*; 2.*National Geological Archives of China*, *Beijing* 100037, *China*)

Abstract: In order to accurately find out the needs of users of geologic data and eliminate the information gap between data users and data managers, the National Geological Archives of China (NGAC) has conducted the data collection of user's accessing logs to the master station of the Digital Geological Library of NGAC. The automatic recording method is applied to register the user's location, their searching keywords, IP address and other related information for constructing the accessing log dataset. In order to make better use of these accessing data, a standardized processing method and quality control system are adopted. The dataset provides the accessing records to the Digital Geological Library of NGAC's master station from the year 2014 to 2017, which realistically reflects user's behaving habits while obtaining geological data. It also provides a sound basis for the further constructions of geological data service website, the development and utilization of geological data, and the geological data management and services.

Key words: Digital Geological Library; website; user's accessing log

Data service system URL: http://dcc.cgs.gov.cn

1 Introduction

The NGAC is the largest and the most complete professional archives in China's national geological industry. Currently, more than 140, 000 kinds of geological data have been collected (Wang Xinchun et al., 2016), covering such fields as regional geological survey, mineral prospecting and exploration, marine geological exploration, geophysical, geochemical and remote sensing geological survey, hydrogeology, engineering geology and environmental geological survey, theoretical geological researches, geotechnical researches (Gao Xuezheng et al., 2016). Stepping upon the on-site lending-borrowing working flow, the Digital Geological Library (DGL) of NGAC is a key national-level geological information infrastructure that features itself with digitalization. It reforms the traditional workflows with digital approaches, and provides the worldwide academic

About the first author: GAO Xuezheng, male, born in 1987, engineer, mainly engaged in geological data management and service research work, E-mail: gxuezheng@mail.cgs.gov.cn.

The corresponding author: QI Fanyu, male, born in 1986, engineer, mainly engaged in the research and management of geological data, E-mail: qfanyu@The mail.cgs.gov.cn.

society with the up-to-date on-line geological information services (Kong Zhaoyu et al., 2016). The construction had been launched since the year 2012. After 6 years of endeavors, the DGL can provide multiple services to the academic communities, including geological data index services, metadata services, geological map services, thematic services, and inter-links and reciprocal checking services.

During the operation of the DGL, a large amount of user's accessing log information is recorded, which can help better understand user's needs, accessing habits, and their awareness and acceptance about the website, and capture more information so as to guarantee safe and stable operation of website (Zhao Guohong, 2010). It also enables managers to accurately grasp the developing tendencies of the website, figure out more specific ways to develop the website in the future (Xie Xiaoping, 2016); meanwhile, it will also help improve the quality of geological data services, improve higher recognition and satisfaction worldwide, and further consequently better solve a series of problems such as dull and dry unitary data servicing way, low efficiency, high cost of acquisition, insufficient utilization of electronic geological data, and many other constraints towards the geological data development as well. Al these efforts will guide the website construction towards the directions of more intelligence, more precise, and more initiative (Yu Shi et al., 2013).

Table 1 lists the metadata of the accessing log to the master station of Digital Geological Library of NGAC.

Table 1 Metadata table of Database (Dataset)

Items	Description
Database (dataset) name	The dataset of user's accessing log to Digital Geological Library of National Geological Archives of China
Database (dataset) authors	Gao Xuezheng, Development and Research Center, China Geological Survey (NGAC) Li Chenyang, Development and Research Center, China Geological Survey (NGAC) Wu Xuan, Development and Research Center, China Geological Survey (NGAC) Kong Zhaoyu, Development and Research Center, China Geological Survey (NGAC) Shang Yuntao, Development and Research Center, China Geological Survey (NGAC) Qi Fanyu, Development and Research Center, China Geological Survey (NGAC) Jia Liqiong, Development and Research Center, China Geological Survey (NGAC) Li Xiaolei, Development and Research Center, China Geological Survey (NGAC) Guo Hui, Development and Research Center, China Geological Survey (NGAC)
Data acquision time	2014—2017
Geographic area	Nationwide
Data format	.accdb
Data size	1.70 GB
Data service system URL	http://dcc.cgs.gov.cn

Continued table 1

Items	Description
Fund project	Geological Exploration Project of China Geological Survey "National Geological Data Integration and Data Collation" (121201004000150018)
Language	Chinese
Database (dataset) Composition	The dataset is composed of two parts, namely: Digital Geological Library's website accessing IP address registration table .accdb; Digital Geological Library's website keywords searching record registration table.accdb

2 Data Acquisition and Processing Methods

2.1 Data Source

Accessing log data of master station of Digital Geological Library of NGAC is from the backend database in the Digital Geological Library website, recorded in real-time throughout the day. The log contents include all the visiting users' login sites, search queries, browsing information, thematic queries, geological map accessing and a series of other related processes. The dataset records in detail the IP addresses of the accessing user, the keywords searched, the time of accessing, the type of user, the location of the user and other related information; and it can also distinguish whether the user refers to the map information or the document information.

Fig. 1 shows Digital Geological Library of NGAC service model and data acquisition process.

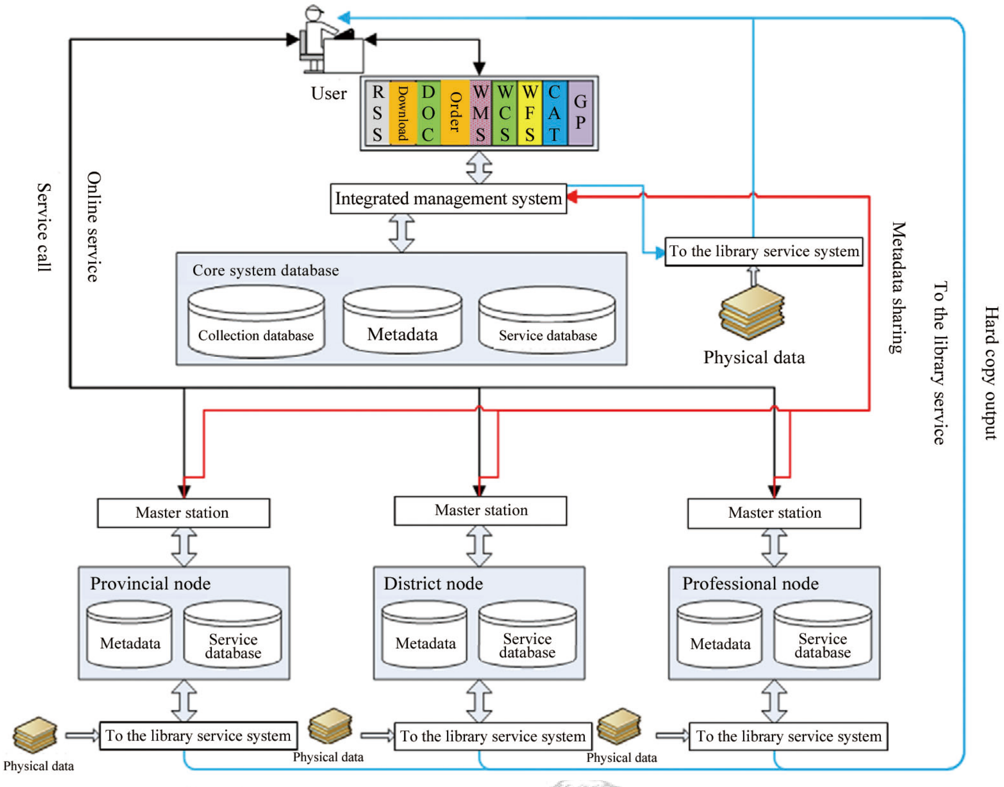

Fig. 1 Shows Digital Geological Library of NGAC service model and data acquisition process

2.2 Data Processing and Application

Data recording, data extraction, warehousing and other system processing is conducted for accessing log dataset of master station of Digital Geological Library of NGAC. Firstly, the website access user's information and operation are recorded; secondly, the data entries are extracted for the data items through the structured data statement; and thirdly, they are stored into the database according to a unified data format.

Since the data in text format can not directly show the changes of the content in each field of the dataset, the data samples are displayed through the domestic (Fig.2) and foreign (Fig.3) accessing user's distribution area, and user's keyword searching statistics (Fig.4).

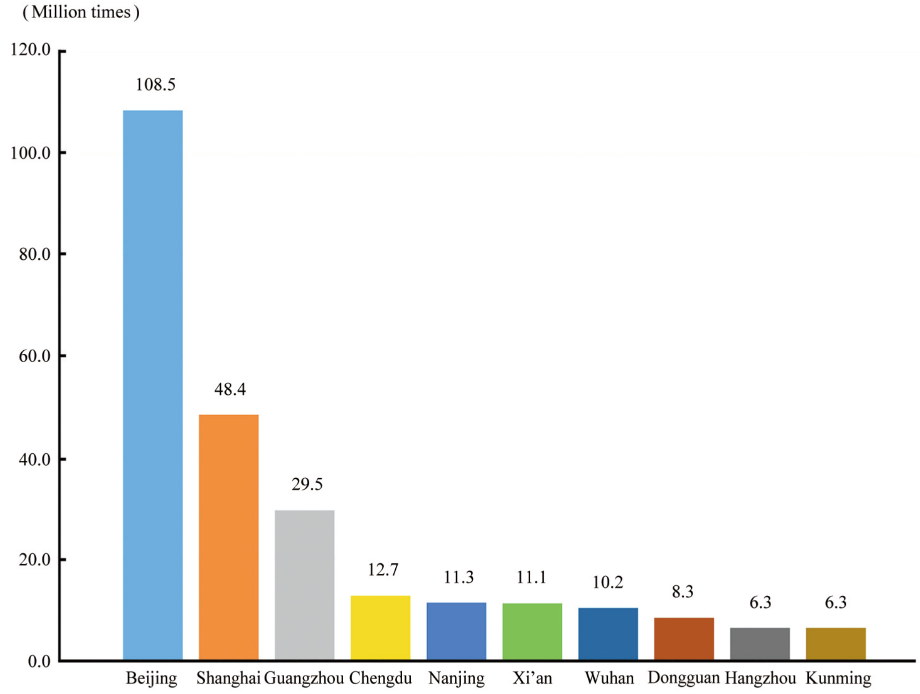

Fig. 2 Domestic accessing user's geographical locations

Based on the statistics on the geographical distribution of domestic accessing users in this dataset, the total times of user's accessing reaches 5.295 million times (5, 294, 535 times); if that from the internal LAN is omitted, it would be 4.487 million (4, 486, 616 times). Beijing, Shanghai, Guangzhou are the top three cities in terms of accessing times, with a total of 1, 864, 000 times (1, 864, 312 times). Fig. 2 shows that the users in the Yangtze River Delta and Pearl River Delta are more concentrated and that from the western China are much less.

The statistical data show that there is a relatively higher times of access to and search in the website of Digital Geological Library of NGAC by users from domestic cities of China such as Shanghai, Dongguan and Hangzhou, where traditionally less geological activities have been conducted. On the contrary, cities in traditional geological major provinces rank low in access times, for example, Qingdao in Shandong Province only ranks the 15th in China. This is mainly due to the following reasons. First, propaganda for network service provided by the Digital Geological Library of NGAC is insufficient, so many people engaged in geological activities have no idea about the Digital Geological Library of NGAC and the data published by it, resulting little use of the geological data

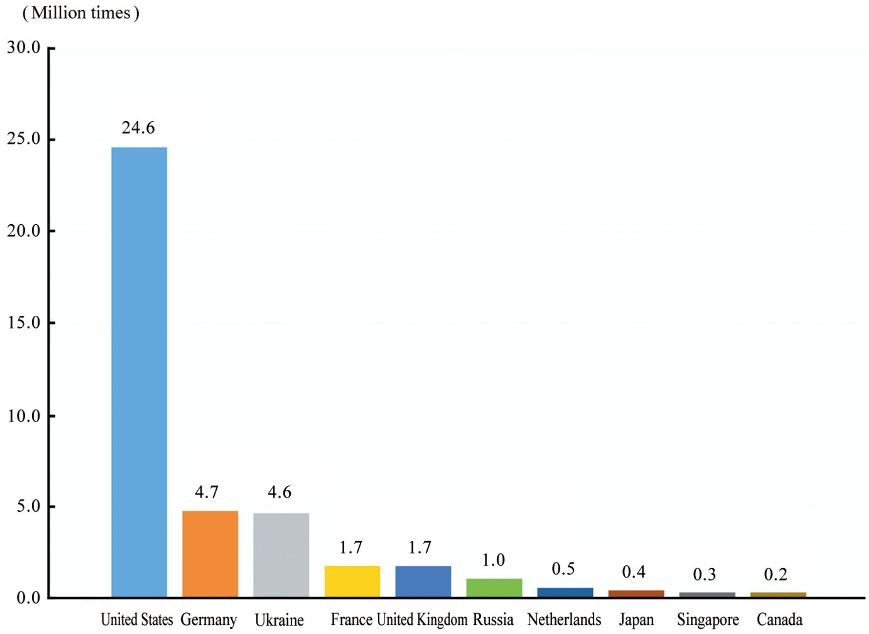

Fig. 3 Foreign accessing user’s geographical locations

by traditional major geological provinces. Second, the first tier cities of China, Yangtze River Delta region and Peal River Delta region are leading in information technology and more sensitive to network, resulting in a higher accessing times from these regions. Third, the Digital Geological Library of NGAC mainly provides search and page view of geological data, no download function is available at present, and users can not acquire the data they are interested in. This is unfavorable for utilization of the digital geological data. Furthermore, the need of geological activities, particularly field geological activities to utilize digital geological data relies on Internet accessibility. However, at present field geological activities in China do not have the condition to access the Internet in most cases.

Among the foreign countries, the accessing times to NGAC from the United States reaches 246, 000 (246, 333 times), ranking the 4th. That from Germany is 47, 000 (47, 091 times), ranking the 15th. Ukraine has 46, 000 visits (45, 927 times) to NGAC, ranking the 18th. That from other countries such as Britain, France, Russia, Netherlands, Japan, Singapore, Canada, Italy, UAE, Sweden, Brazil, Finland, Australia and other countries are also respectively more than one thousand times. Therefore, we can see that foreign countries are also very concerned about the content released by the website of Digital Geological Library of NGAC.

Table 2 Keyword sorting of foreign country

Ranking	Country	Key words 1	Key words 2	Key words 3	Key words 4	Key words 5
1	United States	1:200, 000 geological maps	1:250, 000 geological maps	1:500, 000 geological maps	1:200, 000 hydrogeological maps	H45C003004
2	Germany	1:250, 000 geological maps	1:200, 000 hydrogeological maps	1:200, 000 geological maps	Laishui	Distribution map of groundwater resources in China

Continued table 2

Ranking	Country	Key words 1	Key words 2	Key words 3	Key words 4	Key words 5
3	Ukraine	nothing	nothing	nothing	nothing	nothing
4	France	1:200, 000 geological maps	Xianju	Jiangyong	Qiqihar	1:200, 000 gravity data distribution map in China
5	United Kingdom	1:200, 000 geological maps	oilfields	geological maps	Daging	Hydrology
6	Russia	1:250, 000 geological maps	Ordos Basin	Ordos	Hebei	Shanxi
7	Netherlands	1:250, 000 geological maps	H4820	1:200, 000 geological maps	1:200, 000 hydrogeological maps	China
8	Japan	1:200, 000 hydrogeological maps	1:200, 000 geological maps	tungsten ore	1:500, 000 geological maps	1:250, 000 geological maps
9	Singapore	Bauxite	1:200, 000 geological maps	Guangxi	1:500, 000 geological maps	Aluminum mine
10	Canada	Xi'an	Shanxi	groundwater	Jiapigou	1:250, 000 geological maps

Table 2 shows that ‘1:200, 000 geological maps’ and ‘1:250, 000 geological maps’ have been the two most commonest keywords for search by access from foreign IP addresses. Other keywords concentrated by foreign access users include ‘1:200, 000 hydrogeological maps’, ‘1:500, 000 geology map’ and ‘Shanx’. Abnormalities exist in access by IP addresses in Ukraine, where no search keyword is generated but there is concentrated continuous accessing.

According to the statistics of keywords being searched in the dataset, the top 5 keywords are ‘1:200, 000 geological maps’ (52, 859 times), ‘1:200, 000 hydrogeological maps’ (33, 290 times), ‘1:250, 000 geological maps’ (21, 583 times), ‘1:500, 000 geology map’ (11, 813 times), Xinjiang (3, 392 times). ‘Geological map’ has become the main content of user’s searching targets. Among the top 5, the retrieval ratio of ‘geological map’ reaches 97.2%, being the absolutely main target keyword; while that of ‘1:200, 000 geological maps’ occupies 43% in the top 5, and is 7.2% in all the keywords, being the most frequently retrieved keywords by users.

It is worth to note that Hebei, which is not a geological major or giant province at all, has ranked the 3rd in being retrieved. According to analysis, we consider that this is critically related with the decision of the Central Committee of Communist Party of China and the State Council of China to establish Xiongan New Area, Hebei Province. After establishment of this new area, ministries of the State Council began to prepare New Area Planning, China Geological Survey carried out field geological survey in Xiongan New Area. All these activities increased users’ concern about Hebei Province and keyword search regarding Hebei Province. This shows that users’ retrieving behavior and need are closely related with government policies and current hot affairs. In our future work, it is advisable to provide high quality service products to meet the specific needs of users.

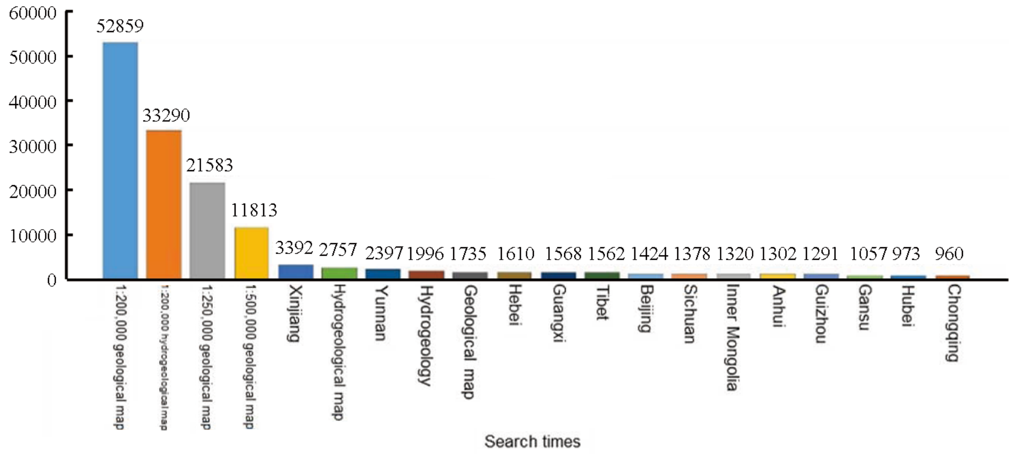

Fig. 4 User's retrieving keywords statistics

3 Data sample description

The Dataset of User's Accessing Log to DGL of NGAC is of Access format database, including two Access format database files, (1) Digital Geological Library's website accessing IP address registration database, and (2) Digital Geological Library's website keywords searching record database. Among them, the first one records the Digital Geological Library's visitors IP address, accessing data content, whether or not the real name registered users, user's accessing geographical location and accessing time during the years 2014—2017; while the second one records the Digital Geological Library's visitors IP address, searched keywords, search time and other data items during the years 2014—2017.

Table 3 Digital Geological Library's website accessing IP address registration database

No.	Field Name	Format	Example
1	Only string	Short text	2293B1F0579C7019E05341015A0A617B
2	IP address	Short text	14.215.222.217
3	Access path	Long text	/Data/FileList.aspx?MetaId=E928A0F55D2F7A73E0430100007F3D67&type=zw
4	User account	Long text	631795983@qq.com
5	User location	Long text	China, Guangdong, Foshan
6	Access time	Data/Time	2015−10−21 9:21

Table 4 Digital Geological Library's website keywords searching record database

No.	Field Name	Format	Example
1	IP address	Short text	59.71.224.2
2	Keywords	Long text	Zhejiang Huzhou
3	Access time	Data/Time	2015−10−17 19:16
4	Only string	Short text	224C06EB72A00920E05341015A0A506A
5	User account	Short text	Anonymous User
6	User location	Long text	China, Hubei, China University of Geosciences (Wuhan)

4 Data Quality Control and Evaluation

This user's accessing log dataset of master station of Digital Geological Library of NGAC contains all the data since the digital data logger started to record the data, with two data tables in the dataset recording a total of 603 million pieces (6, 034, 025) of data. Due to the twice power-supply upgrading maintenance, the master station's accessing log records are not continuous, respectively, from 18:40 on March 17, 2017 to 20:17 on March 18, 2017, and from 14:37 on October 5 to 12:06 on October 7. However, the incomplete records only take a very small percentage of the total records, and do not affect the completeness, reliability, applicability and accuracy of the entire dataset.

The specific working process of data collection is as follows: starting from the very beginning of the user's accessing to the website, the data logger begins to record the user's IP address, accessing location, accessing content and other information, and meanwhile preprocess the recorded data according to the instruction to check whether there is abnormal recording or non-expected performs. According to the time sequence, accessing log data is stored to ensure the accuracy and effectiveness of warehousing. In order to ensure the safe and stable operation of the database, a regular database updating and maintenance strategy is formulated and strictly implemented, and the contents of the database are periodically evaluated. In the process of data extraction and analysis, the data items are unified based on format to ensure the overall consistency of the data content. In order to further improve the data quality, the project team will make more efforts in improving the mechanisms of dataset updating and maintenance, increasing the volumes of new information, and enhancing the versioning of accessing log data.

5 Analysis on Visitors

Statistical analysis on domestic and foreign users' search keywords has revealed that the top 5 search keywords are '1:200, 000 geological maps' (52, 859 times), '1:200, 000 hydrogeological maps' (33, 290 times), '1:250, 000 geological maps' (21, 583 times), '1:500, 000 geology map' (11, 813 times) and 'Xinjiang' (3, 392 times). 'Geological map' has become the main content searched by users. The search ratio of 'geological map' reaches 97.2%, being a dominant keyword in the top 5. The keyword '1:200, 000 geological maps' accounts for 43% search times of the top 5 keywords and 7.2% search times of all keywords, being the most frequently searched keyword.

In terms of geological visitors, the first tier cities of China and developed cities in Yangtze River Delta region and Peal River Delta region have a higher visitor volume, while western regions of China have a lower visit volume. In general, geological visitors of China are normal in access behavior, while many foreign visitors have abnormal access behavior. Analysis on search keywords, registration information, search language and search frequency has revealed that these search behaviors are mostly considered as web redirection to foreign server to hide real IP address by domestic agents. Therefore, among foreign visitors accessing the website, suspects of malicious request, theft and capture of information are not excluded.

6 Conclusions

(1) The Digital Geological Library has used the information system to comprehensively transform and upgrade the traditional geological archives. The paper-back predominated

processing pattern has been updated to the digitalization predominated pattern, the transformation of which has realized the technical innovations in China's geological data management and service. Consequently, the operation mode and organization style of the national and provincial geological data archives have both experienced vital changes.

(2) The user's accessing log dataset of master station of Digital Geological Library of NGAC is of a database that has digitally recorded the user's access to the archived digital resources. The relevant records can be used to track the whole process of user's access to the geological data network. Through statistical and correlation analysis, the multiple internal relations among such many factors as accessing users, accessing time and accessing content can be detected; the user's needs and searching habits can hence been better digitally supported.

(3) According to the analysis of the dataset content and the user's interested fields, the thematic data service is offered, and meanwhile the website home pages is newly arranged; all these efforts have further upgraded the service level of the Digital Geological Library of NGAC, reaching the purposes of precise, efficient and quick services.

(4) The content of the dataset, the operation patterns, and data analysis samples can provide references for the socialized service of the other digital archive agencies; it is especially valuable in promoting the qualified services of China's geological data.

Acknowledgments: The accomplishments of this dataset is very much obliged to the joint efforts of all the project team members. In the process of the design and construction of the Digital Geological Library management system, we have also been firmly supported by Sinosoft Co., Ltd., Turui (Beijing) Information Technology Co., Ltd. and Esri China Information Technology Co., Ltd.; we are here very glad to express our whole-hearted thankfulness to them.

Notes:

❶ Development Research Center, China Geological Survey. 2013. NGAC Core digital system R & D results report [R].

References

Gao Xuezheng, Kong Zhaoyu, Qi Fanyu, et al. 2016. Research into the development of geological data collection and service in the National Geological Archives of China [J]. China Mining Magazine, 25(S2): 73−76 (in Chinese with English abstract).

Kong Zhaoyu, Shang Yuntao, Gao Xuezheng, et al. 2016. Research on construction of national geological data center [J]. China Mining Magazine. 25(S2): 92−96 (in Chinese with English abstract).

Wang Xinchun, Qi Fanyu, Li Xiaolei, et al. 2016. Research on the geological data integration and service: A case study of geological work in the equipped exploration area [J]. Geology in China, 43(2): 691−697 (in Chinese with English abstract).

Xie Xiaoping. 2016. Analysis of the Application of Web Visiting Statistics in the Archival User Research [J]. Lantai World, 18:31−33 (in Chinese with English abstract).

Yu Shiyang, Wang Jiandong. 2013. Government website analysis enters big data era [J]. E−Government, 8:79−85(in Chinese).

Zhao Guohong. 2010. Analysis of traffic impact factors of government portal websites: based on the comparison between Chinese and Japanese government websites [J]. E−Government, 5:62−68 (in Chinese).