第 30 卷第 4 期 地 质 与 资 源 Vol. 30 No. 4
2021 年 8 月 GEOLOGY AND RESOURCES Aug. 2021
文章编号:1671-1947(2021)04-0512-05 中图分类号:P628 开放科学标志码(OSID):
DOI: 10.13686/j.cnki.dzyzy.2021.04.015 文献标志码:A

地质资料数字化工作方案研究

温雪茹[1,2],郭斯嘉[1,2],刘 冰[1,2]
1. 中国地质科学院 水文地质环境地质研究所,河北 石家庄 050061;
2. 自然资源部 地下水科学与工程重点实验室,河北 石家庄 050061

摘 要:对地质资料数字化的整体流程和技术要求进行了全面梳理,便于读者更加精准、科学地组织工作. 对数据化趋势和数据化方法的简要分析更加明确了地质资料管理的努力方向. 在信息化的大背景下, 数字化进而数据化是地质资料管理的必然趋势,能帮助我们更加智能、高效地把地质研究成果提供给社会服务.
关键词:地质资料;档案;数字化;数据化;信息化

STUDY ON DIGITIZATION WORK PLAN OF GEOLOGICAL DATA

WEN Xue-ru[1,2], GUO Si-jia[1,2], LIU Bing[1,2]
1. *Institute of Hydrogeology and Environmental Geology, Chinese Academy of Geological Sciences, Shijiazhuang* 050061, *China*;
2. *Key Laboratory of Groundwater Science and Engineering, Ministry of Natural Resources, Shijiazhuang* 050061, *China*

Abstract: The whole process and technical requirements of geological data digitization are sorted out for more accurate and scientific organization. The brief analysis of datamation trend and method makes the direction of geological data management more clear. Under the background of informatization, digitization and then datamation are inevitable trends of geological data management, which is helpful for the geological research results to serve the society more intelligently and efficiently.
Key words: geological data; archive; digitization; datamation; informatization

0 引言

目前, 数字化进而数据化在档案事业中发挥着重要的作用, 已成为档案工作发展的必然趋势. 如何将纸质档案转变为符合国家标准的电子格式, 是档案数字化的首要问题.

2006 年出版的 GB/T 20530—2006《文献档案资料数字化工作导则》[1],代表了规范档案数字化过程的国家标准的完善, 规定了文献档案资料数字化过程中涉及的标准与一般管理、数字化对象的确定原则、数字化工作的一般过程、数字化过程中适用技术的选择、数字

收稿日期:2020-10-10;**修回日期**:2020-12-18. **编辑**:李兰英.
基金项目:中国地质调查局地质调查专项"水文地质及环境地质调查数据集成与应用"(编号 DD20190433);中国工程院项目"地质专业知识服务系统"(编号 CKCEST-2020-1-16).
作者简介: 温雪茹 (1978—), 女, 硕士, 工程师, 主要从事地质科技档案信息化工作, 通信地址 河北省石家庄市新华区中华北大街 268 号,E-mail//124091339@qq.com
通信作者:郭斯嘉(1989—),女,工程师,主要从事地质图书资料管理工作,通信地址 河北省石家庄市新华区中华北大街 268 号,E-mail//825912122@qq.com

化成果的管理等. 国家档案局 2005 年出版了《纸质档案数字化技术规范》(DA/T31—2005),2017 年根据信息技术发展的新要求重新修订形成标准(DA/T31—2017),对纸质档案数字化过程和数字化成果管理进行了规定[2].

地质资料是地质工作中取得的对地质现象的认知、描述、总结及实物等信息,是实际工作中得到的第一手资料,具有档案和资料双重属性,是地质工作者认识地球所取得的重要知识性财富,具有重复利用、不断开发、长期提供服务的功能,可为经济社会发展、工程建设、地质找矿提供基础信息,价值巨大[3]. 地质档案的形成过程耗资无数,构成内容复杂. 各种地质勘探方法会形成大量复杂、综合的档案资料,包括野外记录、采样记录、物化探野外观测记录,地质、水文地质钻孔原始编录及综合成果,槽、井、洞及坑探原始编录图件,测量原始记录、成果及图纸,长期观测资料,岩矿鉴定成果、化验数据,各类设计及报告的透明图、薄膜图、底版图,设计及报告的审批意见,综合研究的材料以及专题研究论文、报告等[4].

全国地质资料馆历时十几年,累计投资上亿元实现了全部馆藏纸质资料的扫描数字化工作,并已启动全文数据库建设,数字地质资料馆建设基本完成[5]. 为确保图文地质资料扫描数字化的质量,使其生产过程规范化,全国地质资料馆与国土资源部信息中心在参考国家相关规范的基础上,于 2000 年制定了《图文地质资料扫描数字化规范(试行)》[6],规定了扫描数字化过程中的文件组织、数据制作、目录编制及数据保存的一般方法和质量要求. 全国 20 多家省级地质资料馆参照全国地质资料馆的工作模式已经完成了全部馆藏资料的数字化. 中国地质调查局局属单位的资料馆也陆续开展了数字化工作,一部分单位完成了全部馆藏资料的数字化,一部分对馆藏重要地质资料进行了数字化,也有一部分因为资金或重视程度的问题尚未开展.

1　纸质档案与数字化档案的优缺点对比

纸质档案:优点是历史的真实记录,具有凭证价值[7],阅读起来更加直观、舒适;缺点是易破损,且一旦丢失无法弥补. 数字化档案:优点是能够实现多地、永久存储,便于广泛共享,不易丢失,并且便于深入挖掘和集成利用;缺点是易篡改、易泄密. 在实际应用中,一般是两者结合,优势互补,更大的发挥档案作用.

2　数字化与数据化的阶段分析

地质资料管理有 3 个阶段,第一阶段是纸质资料,第二阶段数字化资料,第三阶段数据化资料. 数字化阶段的特点是建设目录数据库,对纸质地质资料扫描存储,两者对接形成网络化的图像浏览. 数据化阶段的特点是全文数据库、图像矢量化、数据整合和大数据平台. 数据化是将图像文件通过数据化技术,转化为计算机可以对其内容进行识读、分析和挖掘的文本信息文件[8-9]. 目前,全国地质资料馆和部分省级地质资料馆已完成数字化阶段,正在进行数据化建设. 中国地质调查局局属单位资料馆基本处于数字化或纸质资料管理阶段.

3　数字化工作方案设计

3.1　人员安排

人员是一切工作的基础保障,目前数字化工作有 4 种人员配备模式:外包、单位临时设置专职、内外组合、专设机构[10]. 外包式,是地质资料管理机构与专业数字化公司签订合作协议,公司承包数字化前处理、数字化扫描、数字化后整理等全套业务. 单位临时专职式,是抽调技术人员组成临时工作组,开展数字化工作. 内外组合式,是以单位专职人员为主,同时从社会上聘用临时工作人员,共同组成数字化工作小组,开展工作. 专设机构式,是在单位中设立一个专职部门,负责开展数字化处理工作. 涉密资料应避免由社会聘用人员处理,应由单位专职人员负责,或与有保密资质的专业公司签订保密合作协议后由其处理.

3.2　确定扫描范围

一般馆藏地质资料数量都很大,在人力或财力有限的情况下,可遵循一定的原则布置数字化工作顺序[11],如利用频率高、使用或收藏价值大、具有典型意义、形成时间早、保管期限长、破损程度大的资料首先数字化,但某个分类里面的资料在数字化时应系统完整,避免后期数字化时在核对上耗费太多时间. 当然也应当考虑到存在这种情况,即不同层面的人员对资料价值的认识不同,有时差异极大,并且随着时间变化,原先认为不重要的资料可能后来却变得很重要.

对于存在不同版本的资料,如正本、定稿、草稿等,根据利用目的的不同,有的单位选择扫描全部版本,展示资料的变化形成过程;有的单位考虑数据资源性、权威性,选择只扫描正本,无正本的扫描定稿,无正本无定稿的扫描草稿.

扫描范围应遵守国家相关保密管理规定,绝密文件一般不建目录数据库、不扫描;机密级和秘密级文件在数字化、存储和使用过程务必严格执行保密规定.

3.3 核对目录数据库

扫描数字化前一般先形成地质资料目录数据库,并确保目录与纸质资料吻合.传统档案强调文件内容间的有机联系[12],关系非常紧密的文件常常整合、装订在一起,作为一件来保存和利用,数字化时建议拆件后以纸质档案的“自然件”为单位进行编目和扫描,根据规范要求补充档号,在保证新生出的每条目录都有编号的基础上,保持文件之间的有机联系,以及目录与纸质档案的良好对应关系.著录时注意标明密级等必要信息,充分利用好计算机不怕细的优点.著录的同时做好摸底调查和登记、统计工作,例如对破损、霉变、虫蛀情况以及严重程度做出登记,并统计数量和修复的工作量.

如果前期已经形成了目录数据库,数字化前需要与纸质资料核对.目录信息不完整、漏缺的予以补充,不准确的重新著录[13].

3.4 修复处理

纸质地质资料存在的折皱、卷曲、污损、字迹不清楚等问题,需要采取合理的措施进行修复,并且要求工作人员有一定的地学专业知识,必要的情况需请教相关专家鉴定.对于价值比较高的地质图件,需要做到修旧如旧,保持原貌.对于霉变较严重的老旧地质图件,在进行扫描作业前,作业员需要根据图件资料具体霉变情况,选用水、有机溶剂、氧化剂对脏污进行清理.对于折皱、卷曲,我们采用压平以及用金属电熨斗从资料背面轻轻熨烫的方式尽可能减少折痕(要保证不伤害资料)[14].破损资料修复时,根据传统的修裱经验,黏合剂一般选择除去面筋的小麦淀粉制作成的浆糊,这种浆糊性柔黏性适度,用以修裱可以取得柔软、平整的效果,此黏合剂不破坏纸张酸碱度并具有可逆性[15].

拆件后需恢复的资料,必要时可做好记录,以便有序地恢复成原貌.

3.5 资料分类

在扫描之前,需要对地质资料原件进行整理分类,为正式开展扫描做最后的准备.一方面按照涉密情况分类,分为涉密资料和公开资料,另一方面根据扫描技术要求不同分类,分为文本和图件.也可再进一步细分,如手写稿和打印稿,胶片和纸质,1960 年前和 1960 年后等,便于针对不同的类别做不同的扫描技术要求.

3.6 确定扫描工作技术标准

3.6.1 扫描色彩模式的选择

扫描色彩模式分为彩色、灰度、黑白 3 种.灰度和黑白模式是早期大多采用的,主要是因为当时存储设备、电脑系统运行速度、网络带宽等条件有限.彩色模式存储容量大无法大规模实现,而灰度模式和黑白模式存储容量小便于实现.目前随着设备、网络条件的提升,主要采用彩色模式,对于字迹清晰、不带插图的黑白文本可用黑白模式扫描.

3.6.2 扫描分辨率的选择

分辨率的选择以栅格文件的清晰度为准,应最大限度地接近扫描原件,具有较好的还原度.凡原件中可识别的内容(除污迹外),在栅格文件的打印结果和屏幕显示结果中应可识别、无歧义[15].根据《纸质档案数字化技术规范》[15]要求,图纸分辨率一般不小于 300 dpi;文字、照片一般不小于 200 dpi,如某些文字偏小、密集、清晰度较差时则不能低于 300 dpi;重要照片建议用 500 dpi 以上扫描,以确保照片的层次性和色彩的丰富性;需要进行高精度仿真复制的档案,建议不小于 600 dpi,具体应根据实际情况调整分辨率及其它相关参数.对于年代久远的或质量不佳的老资料最好用 600 dpi,一步到位,以后尽量使用电子版或打印的纸质版,减少原件的翻阅.

3.6.3 图像存储格式的选择

《纸质档案数字化技术规范》[15]要求:纸质档案数字图像长期保存格式为 TIFF、JPEG 或 JPEG2000 等通用格式,图像压缩率的选择可根据实际应用的需求而定.JPEG 是如今最常用的图片格式之一,其优点为兼容性高、容量小、传输速度快.JPEG 是一种有损压缩格式,允许用不同的压缩比对文件压缩,方便在图像质量和文件大小之间找到平衡点.在实现相同清晰度的情况下,JPEG 的容量相对较小.JPEG2000 既支持有损压缩也支持无损压缩,在获得相同图像质量的情况下

可以比 JPEG 的压缩比更大，而且能够实现图像的渐进传输[16]. 虽然 JPEG2000 在技术上有一定的优势,但是目前还不是很普及，在档案部门使用得很少. TIFF 的优点是可以实现对图像的无压缩存储或无损压缩存储,图像失真度极小,可以保存分层和透明信息,并且可以将多个图像合并成为多页的一个文件，其缺点是占用存储空间很大. 因此目前档案行业最常用的是 JPEG 图像存储格式,一般使用 75%压缩比.

图像一般保存两类格式：存储格式和利用格式. 存储格式是上面提到的 JPEG、TIFF 等精度高、用于长期存储的格式. 利用格式是 PDF 等便于网络传输浏览、容量较小的格式. PDF 格式的优点是网络传输快、适合网络浏览，可以把多个图像整合成一个 PDF 文件,适合文本和图集的制作,缺点是图像文件经过了再压缩处理,图像质量远远达不到印刷要求.

3.6.4　扫描设备的选择

扫描仪一般有平板扫描仪、快速扫描仪、高拍仪、手持扫描仪、大幅面扫描仪等. 平板扫描仪的优点是清晰度高，扫描质量好，缺点是需要手动一页一页进行,效率相对较低;快速扫描仪优点是自动进纸,速度快,扫描质量高,缺点是适用范围小,只能针对纸张质量好、能够拆成单页的文本；高拍仪的优点是操作方便,相比平板扫描仪速度快,缺点是对于文本中特别细的线、特别浅的颜色、比较小的字符捕捉不清晰或有缺失;手持扫描仪的优点是便于携带,缺点是文本夹缝处容易漏扫；大幅面扫描仪是针对大幅面图纸扫描的最有利工具. 在实际工作中，文本的扫描采用平板扫描仪和高拍仪结合使用的方式，适用于高拍仪的使用高拍仪，不适用于高拍仪的使用平板扫描仪，小于等于 A3 幅面的图件使用平板扫描仪,大于 A3 幅面的图件使用大幅面扫描仪. 在档案扫描中，很少使用数码相机,主要是图像容易变形、质量差,而且容量大.

3.7　图像处理

图像处理主要包括对扫描文件进行旋转、纠斜、调整页边距、去污、修补、拼接、勘误工作. 整饰过的字符、颜色等属性应尽可能与原件保持一致. 整饰后应保持图文地质资料的原意，不能违背. 少部分原件本身就不清楚且无法考证的信息，原则上不做整饰. 分类、拼接、加页、重编页码等工作应形成工作记录,便于后续查证.

3.8　数据化加工

数据化分为文稿和图纸两种类型. 文稿类采用 OCR 技术对图像进行光学字符识别[17],形成可编辑的文本. 图纸类采用 MapGIS、ArcGIS、AutoCAD 等制图软件进行矢量化处理形成可编辑的文件. 目前全国地质资料馆和一些省级地质资料馆已经大规模开展数据化工作. OCR 识别能够解决档案全文检索的问题,图像分辨率在 200 dpi 以上便能够达到 OCR 识别要求,识别率比较高. 识别成全文文本后，可对重要信息进行人工核对,如标题、作者、段落名称、表格名称等,能够提高关键信息检索的准确性，同时具有全文的检索深度. 如果对全文进行人工核对,需要的成本太高,目前不太现实，性价比较低. 识别后的文本与图像结合制作成双层 PDF 文件,上层是原始图像,下层是文本,既可以 100%保留原始版面效果，又便于建立索引数据库，是目前常用的模式. 我们平时看到 PDF 文件是扫描版的,却可以复制其中的文字(偶尔会有错字)就是双层 PDF 文件的效果.

3.9　数字化存储

充分进行数据使用和存储的风险评估，建立长效保护机制[18-19]. 一般保存两套成果:优化后的图片和图像合成的 PDF 文件[20]. 存储介质方面,光盘最稳定,硬盘便于读取. 目前硬盘塔等存储设备可以自动进行多份存储,能够较好地保证数据安全,可以采用光盘与硬盘结合的方式进行备份,U 盘因为不稳定不建议作为存储方式. 档案级光盘较普通光盘的容量大,有 10 G、20 G、50 G 等规格，寿命可达到 20 a 以上. 在实际工作中,保存一式 3 套,硬盘塔 1 套,档案级光盘 2 套(采用只读模式，避免因能够修改造成数据权威性降低). 存储介质外部需要标注外标签,注明编号、形成时间、负责人等基本信息便于查找、追溯.

3.10　数据共享

纸质地质资料数字化、数据化后为网络共享提供了基础数据支撑，为地质资料提供社会服务创造了更加便利的条件. 地质云、全国地质资料馆网站等网络平台提供了部分公开地质资料(或资料公开部分)的在线阅览和下载服务,对于敏感或涉密的地质资料(或资料的非公开部分）基本都是通过部门内部的涉密电子阅览室，在物理隔离的小局域网环境下提供限制性的服务.

4 结语

通过梳理地质资料数字化的整体流程和技术要求,帮助地质资料管理者更加精准、科学地组织工作,对数据化趋势和操作方法的分析更加明确了地质资料管理的努力方向. 在信息化的大背景下,数字化进而数据化是地质资料管理的必然趋势,能够使我们更加智能、更加高效地把地质研究成果提供给社会服务.

参考文献(References):

[1]刘芳. 中外档案数字化政策比较与启示[J]. 浙江档案, 2010(10): 27-28.

Liu F. Comparison and enlightenment of Chinese and foreign archives digitization policies[J]. Zhejiang Archives, 2010(10): 27-28. (in Chinese)

[2]国家档案局. DA/T 31—2017 纸质档案数字化规范[S]. 2017.

National Archives Administration. DA/T 31—2017 Specification for digitization of paper-based records[S]. 2017. (in Chinese)

[3]颜世强, 连健, 丁克永, 等. 地质资料内涵与特征分析[J]. 中国矿业, 2013, 22(7): 45-48.

Yan S Q, Lian J, Ding K Y, et al. The connotation and features of geological data[J]. China Mining Magazine, 2013, 22(7): 45-48.

[4]余四星. 试论地质档案的特点及开发利用[J]. 档案学通讯, 1989(3): 46-48, 32.

Yu S X. On the characteristics and development of geological archives [J]. Archives Science Bulletin, 1989(3): 46-48, 32. (in Chinese)

[5]张立. 地质资料服务进入数字化时代[N]. 中国矿业报, 2016-09-28(002).

Zhang L. Geological data service enters digital era[N]. China Mining News, 2016-09-28(002). (in Chinese)

[6]王春宁, 单昌昊. 全国地质资料馆数字化工作回顾[J]. 中国档案, 2010(6): 53-55.

Wang C N, Shan C H. Review of digitization of national geological data center[J]. China Archives, 2010(6): 53-55. (in Chinese)

[7]尚武. 地质资源数字化信息建设的原则及若干问题的探讨[J]. 中国矿业, 2005, 14(7): 38-40.

Shang W. The rules and related issues on the development of the digitalization geological data[J]. China Mining Magazine, 2005, 14(7): 38-40.

[8]杨来青. 再信息化: 档案馆发展战略的思考[J]. 浙江档案, 2019(9): 15-18.

Yang L Q. Reinventing informatization: a study of the development strategy of archives[J]. Zhejiang Archives, 2019(9): 15-18.

[9]李宝玲. 数字档案馆建设的机遇、挑战与思考[J]. 档案管理, 2020(2): 27-28.

Li B L. Opportunities and challenges of digital archives construction [J]. Archives Management, 2020(2): 27-28. (in Chinese)

[10]李红梅, 张栋. 纸质档案数字化前处理工作探析[J]. 档案学研究, 2015(4): 105-108.

Li H M, Zhang D. Analysis of pre-processing of paper archive digitization[J]. Archives Science Study, 2015(4): 105-108.

[11]刘兢兢. 档案数字化前处理工作的思考[J]. 档案与建设, 2012(6): 17-19.

Liu J J. Thinking about pro-processing work of archives digitization [J]. Archives & Construction, 2012(6): 17-19.

[12]韩李敏. 档案数字化攻略[J]. 浙江档案, 2019(1): 56-59.

Han L M. The strategy of archives digitization[J]. Zhejiang Archives, 2019(1): 56-59.

[13]徐叶黎, 钟秀梅. 基层档案室纸质档案数字化工作的实践探讨[J]. 山西档案, 2015(5): 67-68.

Xu Y L, Zhong X M. Digitalization of paper archives at the local level[J]. Shanxi Archives, 2015(5): 67-68.

[14]李敏, 纪惠英, 李秀荣. 原始地质资料数字化工作中的问题探讨[J]. 地质与资源, 2013, 22(5): 431-434.

Li M, Ji H Y, Li X R. Discussion on the digitalization of primary geological data[J]. Geology and Resources, 2013, 22(5): 431-434.

[15]于瑞洋, 贾国锋, 郭慧锦, 等. 破损纸质地质图件修复研究[J]. 中国矿业, 2019, 28(S2): 67-71.

Yu R Y, Jia G F, Guo H J, et al. Research on restoration of damaged paper geological maps[J]. China Mining Magazine, 2019, 28(S2): 67-71.

[16]傅荣校, 翁敏曦. 档案数字化扫描与存储格式比较研究[J]. 档案学通讯, 2007(2): 61-64.

Fu R X, Weng M X. Comparison of paper-based archives digital scaning and its storing formats[J]. Archives Science Bulletin, 2007(2): 61-64. (in Chinese)

[17]庞莉. 手稿与图纸档案数字化过程比较研究[J]. 档案与建设, 2018(1): 26-29, 51.

Pang L. A comparative study on the digitization of handwritten manuscripts and drawings archives [J]. Archives & Construction, 2018(1): 26-29, 51.

[18]孔昭煜, 郭磊, 李海龙, 等. 大数据背景下地质资料电子数据长期保存技术探究[J]. 中国矿业, 2019, 28(6): 69-72.

Kong Z Y, Guo L, Li H L, et al. Exploration on the long-term preservation technology of geological data under the background of big data[J]. China Mining Magazine, 2019, 28(6): 69-72.

[19]朱丽梅. 馆藏档案数字化风险应对研究[J]. 档案与建设, 2013(9): 18-21.

Zhu L M. Discussion on risk response of digitalization of the collecting archives[J]. Archives & Construction, 2013(9): 18-21.

[20]郭俊杰. 科研纸质档案数字化的规范化流程[J]. 矿业工程, 2018, 16(1): 61-62.

Guo J J. Standardization flowsheet of digitalization of scientific researches paper archives[J]. Mining Engineering, 2018, 16(1): 61-62.