●刘家真 徐 曼

数字化过程中的管理决策与技术建议*

摘 要 在文献数字化的输入阶段,模数转换方式、扫描模式和分辨率的选择,是决定数字化源图像质量的关键,而源图像质量将直接影响创建于屏幕显示、打印及其派生的图像质量等。由此,提出扫描对象与最小分辨率的建议、图片数字化的格式化建议、文献数字化模式和技术指标。表3。参考文献8。

关键词 文献数字化 模数转换 扫描模式 分辨率 格式
分类号 G250.76

ABSTRACT The authors think that during digitization processes, the methods of D/A conversion, the patterns of scanning and the selection of resolutions are key factors for digitized source images, which will also affect the quality of screen display, printing and derived images. Therefore, the authors propose some suggestions for minimum resolutions, digitization formats, digitization patterns and technical indicators. 3 tabs. 8 refs.

KEY WORDS Document digitization. D/A conversion. Scan patterns. Resolution. Format.
CLASS NUMBER G250.76

数字化是将传统文献转换为数字形式以供存储、检索与传输的过程。随着网络传输的发展以及人们对数字信息需求的提高,数字化项目已越来越具有广泛性与多样性。人们期待数字化的技术规范与指南早日诞生,以便通过简单的结论来解决复杂的现实问题。

然而,要寻求适合各类文献数字化的统一技术指标是不现实的。但数字化项目的实践经验,却是人们在探索中前进的最佳指南。作者在研读文献的基础上,将数字化项目中成功的经验推荐给大家,以减少失误。

概括起来,文献数字化的基本要素主要有输入、处理、存储、传输与输出等。在输入阶段,模数转换方式、扫描模式与分辨率的选择,是决定数字化源图像质量的关键,而源图像质量将直接影响创建于屏幕显示、打印及由其派生的图像质量,这些指标的确定也关系到数字化工程的效用和费用等。

1 模数转换方式

在当前技术条件下,直接使用扫描仪扫描原件形成的数字图像是较为理想的,但由于文件大小或其他原因,并非所有文献都能直接通过扫描数字化。有时必须使用传统摄影技术制作原件的胶片拷贝作为数字化中间体,或使用数码相机进行拍摄实现数字化转换。

1.1 对原件直接扫描

数字化要获得良好的效果,首先必须选择扫描仪。

最常见的平台扫描仪对扫描的原件没有特殊要求,可允许扫描各种形式的原稿,包括书籍、三维物体等。平台扫描仪有接触式与非接触式(又称高架平台扫描仪)。对于稀有的、易脆的文献宜选用非接触式扫描仪。

如国家图书馆专门从德国购进一台非接触式的扫描仪,它的稿台通过液压控制,书脊可以自动调节,根据书的重量,机器自动调节稿台上下的空间,不会因为压力而造成纸张损坏。国家图书馆将它用于特别珍贵的宋元善本图书的数字化加工[1]。

滚筒式扫描仪使用光电倍增管作为感光元器件,能产生比平台扫描仪更佳的数字图像。所以多用滚筒扫描仪制作高质量的图像,"特别是滚筒扫描仪在复制高光、暗调细节很多的图像,及需放大至300%以上的原稿上有着得天独厚的优势。"[2]因而,滚筒扫描仪一般应用于专业的大幅面文献的扫描

* 本文系教育部人文社会科学基金资助项目(01JA870010)的研究成果。

上。滚筒扫描仪的不足是仅可扫描柔软的原件。

1.2　以文献的胶片拷贝作为扫描中间体

在以下几种情况下,可以考虑以胶片影像作为中间体制作原件的数字图像[3]。

(1)已经具有高质量胶片拷贝的文献,应使用胶片拷贝进行扫描。因为这将比直接扫描原件更便宜、更容易,不仅节省了成本与时间,还避免了对原件的损坏。特别是具有缩微拷贝片的原件,更应考虑使用胶片影像作为文献数字化的中间体。

(2)过于脆弱的原件可以考虑首先用普通相机拍摄原件,制成原件的底片拷贝,再用底片扫描仪将原件底片拷贝转换为数字图像。

(3)如果扫描对象的大小超过扫描仪可容纳的尺寸或拟以二维显示的三维对象,可考虑制作该原件的胶片拷贝,通过对胶片影像的扫描完成数字化。

(4)缩微技术比数字技术稳定,同时缩微胶片具有保存信息的稳定性,因而缩微胶片仍然被认为是保存纸质文献的首选介质。对于需要持久保存的文献应首先制成缩微胶片,再通过扫描缩微胶片制作其数字图像文件。这比既对文献缩微又对文献扫描要经济得多,同时有利于文献的保护。对于需要彩色或灰度信息的文献,可对这些页面单独扫描。

利用胶片扫描仪直接扫描胶片影像就可以达到原件数字化的目的。依扫描对象的不同,胶片扫描仪可分为底片扫描仪、缩微片扫描仪和幻灯片扫描仪等。底片扫描仪专门用来扫描定影后的底片,扫描的图像十分清晰,使用成本很低,可以有效减少误差,提高精确度,节省时间,还能够通过曝光补偿、色彩修正、艺术效果处理等手段获得高质量数字图像。缩微胶片扫描设备可以对缩微平片、卷片、开窗卡、封套片进行扫描。

利用胶片影像作为原件数字化的中间体,其数字图像的质量取决于胶片影像的质量与扫描的精度。影像质量好的胶片是完全可以替代原件完成扫描任务的,但胶片影像质量差、成像不清晰或聚焦不实,或有擦伤、退色等问题,扫描成像的质量就会差。因而,在采用胶片影像作为原件数字化的中间体时,必须注意:

(1)制作高质量的胶片拷贝。拍摄与冲洗质量的控制是获得高质量胶片影像的关键,应极为小心谨慎地进行,以保证准确、完整地再现原始对象。

(2)胶片影像质量不高或胶片损坏严重的不宜作为文献数字化的中间体。我国早期的一些缩微胶片,限于当时的技术条件,有些影像质量并不理想,如有资料介绍敦煌文献的某些缩微胶卷就发现有影像不清的问题,像这类缩微胶片是不宜用作扫描对象的。此外,已经发红、发黄、退色或生彩斑或有霉斑、水迹、乳剂起皱、药膜脱落等问题的胶片拷贝也不可作为原件替代品。

(3)尽可能使用底片扫描。直接使用原件的底片扫描可以获取较高质量的数字图像。这是由于,底片扫描仪扫描底片是目前扫描精度最高的扫描方式,特别是当底片质量好又采用高分辨率的胶片扫描仪扫描,可以获得最佳的扫描图像质量。将负片复制成照片后,图像的细节与分辨率多少都会有所损失,这必然会影响到扫描成像的质量。例如,正常情况下用中档135镜头拍摄的底片,它的分辨率相当于3000dpi左右,冲扩后的成品却只有200dpi左右(专业冲扩店可达600dpi以上)。面对低分辨率的原件,扫描仪自身分辨率再高也将无济于事。可见扫描底片会比扫描照片可能获得更好的数字图像质量。使用负片扫描的另一个优点是,负片提供在动态范围内修复曲线的工具,使得阴暗区域与明亮区域可以处理得更好[4]。

1.3　数码相机直接拍摄

从理论上讲,直接拍摄原件所获得的图像质量应是最好的,但高质量的民用数码相机难得。有人对数码相机直接获得的数字图像与对原件底片拷贝扫描获得的数字图像进行过质量比较,发现在输出分辨率相同的情况下,数码相机生成的图像质量不如底片扫描仪扫描底片,而且目前主流数码相机的输出分辨率仍然低于底片扫描仪的水平。因而,建议在民用数码相机质量尚不理想的情况下,谨慎使用数码相机捕获需长期保存的数字图像。这方面的教训是有的,如台湾故宫博物院在1996年就提出将清代档案用数码相机摄制成光盘保存,并用数码相机建置了30万件图像文件,后发现数码相机摄制的影像文件不够清晰且较模糊。在建立影像数据库时,不得不将这类不够理想的影像文件重新扫描再建影像文件。

有些部门为了解决超大尺寸文献的数字化问题,将原件挂在墙上,用数码相机一块块拍完后进行拼接。这种方法其实是不妥的。也许采用传统相机拍摄原件图像,再以原件底片作为扫描中间体,得到的数字图像效果更好。

2 扫描模式的选择

扫描模式与扫描分辨率直接关系到捕获原件信息的多少,关系到数字图像的主文件是否可以精确地再现原始信息。扫描模式决定了从原件中捕获到的颜色信息的数量,也直接关系着形成的数字图像文件的大小。要达到较好的扫描品质,前提是选择正确的扫描模式。

一般扫描仪都可提供三种扫描模式,即黑白扫描、灰度扫描与彩色扫描等。文献以哪种模式扫描为好,以下两点是需要考虑的:拟扫描对象的类型;扫描结果的用途,是彩色显示、黑白显示还是准备使用OCR处理等。

2.1 黑白扫描模式

黑白扫描模式可以捕获到没有丝毫色调浓淡变化的纯黑与纯白双色图像,是三种扫描模式中产生文件最小的。黑白双色图像可以高效地被压缩,并可产生多种便于激光打印或喷墨打印的输出效果。但黑白扫描模式常常无法复制出能充分显示多样反射比值的页面,如原件的亮度、黑度与色彩是无法通过黑白扫描反映出来的。因而,该模式不适于需要保留色彩信息的文献的扫描,也不适于扫描手稿或年代久远的印刷文献。年代久远的文献,纸张发黄,颜色深浅或状况各不相同,有时在同一页面上也会出现这类差异。如果选用黑白扫描,捕获的图像可能效果极差,如图像对比度非常低,不利于识别等。至于手稿,在同一页纸上,字迹或笔画的宽度、密度也不相同。若采用黑白扫描模式,这类文件的复制件就有损失原始信息的风险,或从外观、视觉感受上蒙受极大损失。一般说来,黑白扫描模式主要适于下列文献选用:不带图表、插图的黑白印刷文本文献;几乎没有或完全没有色调浓淡变化的黑白线条图。

2.2 灰度模式

该模式能够提供从纯黑到纯白间256级灰度,可以记录和显示原件更多层次的明暗色调。扫描出来的图像有较高的清晰度,不仅有黑白两色,还可完整保留原件的灰度层次。以下类型的文献可以考虑选用此模式:黑白负片印在相纸上的图片;拟用黑白打印的彩色图片;一般的工程设计图、建筑图纸等;字迹有浓淡变化的手稿,如铅笔或钢笔字迹、绘画临摹等;年代较早的印刷文献,特别是档案文献。

从1994年开始,美国的一些图书馆就开始研究年代久远的印刷品和手稿的数字化。对于一些珍贵的资料,图书馆一般是采用无压缩的灰度图像来保存。

2.3 彩色模式

彩色扫描可以从拟扫描的对象中捕获到最多的色彩信息,最多可达1670万种。故彩色扫描形成的图像文件最大。彩色文件的数据量是同等条件下灰度文件数据量的3倍,是黑白文件数据量的24倍。可用黑白扫描的选用彩色模式,不仅文件的数据量会增大、识别时间加长,而且扫描时间也要长得多。但如果必须采用彩色扫描的,选用了其他模式,就有可能会损失原件的许多重要信息,使数字图像无法真实地再现原始信息。因而,根据原件类型来选择扫描模式就很重要了。一般说来,适于彩色扫描的对象是那些必须保留色彩的文献,如:(1)如果文献的色彩所代表的信息具有史料价值,则必须选用彩色扫描模式。如手稿,珍贵、古老的文献等。(2)文献色彩所代表的信息具有重要价值,如文献上必须保留的有色标记(有色重点区、有色边框或划道等)。(3)文献色彩所代表的信息具有美学价值的。(4)原件是彩色的,且希望以彩色来显示、打印或编辑的,如彩色负片印在相纸上的照片。

3 分辨率

分辨率是辨别精细空间内细节的能力,是决定数字图像质量的最重要因素之一。扫描分辨率表示扫描仪在既定文档中捕获像素的模式与数量,它决定了从原件中所采集信息的精细程度,扫描分辨率越高,所获得的图像越精细。扫描阶段,应尽可能多地捕获原始图像信息,以便在后面的转换处理、打印输出过程中即使丢失部分信息,仍然可以保持一定的图像信息总量,保证数字图像的相应品质。

扫描分辨率的设置是极为重要的。分辨率设置过低,捕获的信息量过小,最终输出的图像质量低劣;分辨率设置过高,扫描图像的信息量很大,占用的硬盘空间增大。一味追求扫描分辨率是没有实际意义的,这是因为:数字图像最终输出的质量要受计算机处理能力与输出打印机分辨率限制,并非扫描分辨率越高,输出的图像质量越高;用于扫描的原件分辨率如果很低,使用再高的扫描分辨率也是枉然的。此外,高分辨率极大地增加了文件尺寸,这对于存储、传输都是无益的。在扫描分辨率的设置上,必须权衡图像质量、成本、必要性与可能性。

扫描分辨率的设置尚无统一的尺度,主要取决

于被扫描的原稿、扫描需求以及图像处理等。最佳扫描分辨率的选择,首先必须考虑输出的数字图像中,可否保证最小字符或最有意义的信息清晰可读。这两点在实际工作中常常也是最难确定的。对于印刷文本(如印刷书刊、文件),最小的字符常常是上标、脚注等,但手稿的最小字符就难以确定了。影响手稿字迹读出的因素太多,如墨色浓淡、字迹大小等等。最有意义的信息的确定取决于该数字图像的用途与主观因素。对于照片、图片与地图等文献,最有意义的信息就很难确定。如一幅图片上哪个信息更为重要,是随使用目的(欣赏、证据、资料等)及用户需求(普通用户、研究人员、鉴赏家等)而变化的。

因而,最佳扫描分辨率不仅是个技术指标,也凝聚着管理的决策。从技术的角度看,清晰易读取决于扫描分辨率与位元深度(或称位元浓度 bit depths)的结合。位元深度将简单问题复杂化,即在灰度扫描、彩色扫描中,使用比黑白扫描更低的分辨率,也可以获得相同程度的清晰度。因而,专家们认为对于某些文献使用彩色与灰度技术可能要比单靠增加扫描分辨率更能提高某些黑白图像的质量[5]。

表 1 是根据美国 Image Quality Working Group of ArchivesCom, joint Libraries/AcIS committee 共同提出的《数字成像工程的技术建议》中的数据而编制的,也许可供我们参考。

表 1　扫描对象与最小分辨率的建议

扫描模式	扫描对象		最小分辨率(dpi)	备注
黑白扫描	现代印刷文献(不带图表与插图)		600	可确保所有的细节被捕获
	黑白线条图	线条细密	600	
		线条粗宽	可低于 600	需测试
灰度扫描	手稿、打字稿、半色调文献及类似资料		300	若文献所有字符相当大,可考虑使用 300dpi 以下的分辨率,需测试。
	黑白照片		应以 300dpi 进行测试,观察是否够用。	带有重要细节的照片,需要更高的分辨率。
彩色扫描	地图、海报等印刷文献		200	需测试
	彩色照片		应以 300dpi 进行测试,观察是否够用。	有重要细节的,需更高分辨率。
	历史久远的手稿		600	可捕获到纸张纹理等细节

加利福尼亚数字图书馆在数字图像收藏标准中,对用胶片作为数字化中间体的扫描分辨率提出了以下建议:调整胶片拷贝清晰度,达到 600ppi(8 位灰色,24 位彩色)[6]。同时建议:"如果你想通过胶片作中间转换媒体,然后把胶片进行数字化,一定要考虑原件的尺寸。例如,从 10 英寸的原件上以 1200dpi 标准产生一个 4×5 英寸的底片,结果是 6000ppi 或 600dpi 的影像。要想在最长边为 12 英寸的原件上产生 6000 像素的影像,需要以 500ppi 标准进行数字化,而同时会丢失一些细节信息。"[7]

扫描分辨率的确定还得考虑形成的图像文件是否需要光学识别。分辨率设置不当,低版本的 OCR 可能根本无法识别文字材料。对于不同的扫描模式,OCR 的识别能力与要求也不同。尽管目前 OCR 软件一般都具有识别彩色稿件的功能,但从效果看,黑白扫描模式的识别率较高。对于大多数黑白扫描的文字材料,300dpi 是可以进行识别的最低值,若被扫描的字体太小,分辨率就需加大,特别小的字体加大到 500～600dpi 才可较好识别。反之,被扫描的字体大,分辨率可考虑缩小。对于灰度模式的扫描,OCR 对扫描分辨率的要求不宜低于 200dpi。

尽管今天的 OCR 不尽人意,但光学字符识别代表了扫描与图像处理同时进行的发展趋势。

4　存储与传输

不少的数字化项目的产物是用于单机浏览的,特别是一些享有版权的印刷文献的数字化版本(如某些大型工具书)的图像数据。但更多的数字化产物是提供在线利用的。文件传输的速度直接关系到远程访问的成败,因而图像数据的传输在数字化工

程中就备受重视。

压缩可以为文件减肥,但也或多或少会有损图像质量。尽管人们在图像质量与传输速度上做了不少尝试,但至今完美的方案尚未产生。大多数解决方案都提出了要求创建多样化的数字图像,以分开解决图像质量与传输速度的问题。

可利用档案图像文件来解决原件高质量图像的长期保存问题。档案图像可精确地再现原始信息,在某些情况下可替代原件。其主要作用有:提供用户拷贝,如通过档案图像可获取高质量的打印图像或印刷图像等;用于创建供编辑、浏览、传输、复制的派生文件或转换为其他格式文件;用于将来的再处理,如转换为供网络传输的格式或以便技术更新后的再迁移等。因而,档案图像文件应是原件经扫描后未经压缩的高质量源图像。

供用户获取的文件不宜过大。除浏览不便,它所占的巨大空间会延缓打印速度,使大型图像难以打印。可见,压缩供用户存取的图像文件,无论是在线存取还是单机浏览都是必要的。为解决视觉效果与文件大小的矛盾,与书目一同显示的微图(预览图)可采用大压缩比的压缩技术,以便提高网上传输的速度。供用户正常存储与显示的浏览图(又称参考图),为了保证视觉效果和便于存取,也须些微压缩,使其图像质量略低于档案图像,高于预览图。

存储格式与压缩一样,也直接关系到原始信息留存的多少。人们在这方面也做过大量探索。美国国家数字图书馆、国会图书馆向文献管理部门提供了数字化技术措施的建议,表2是笔者依据相关的推荐意见编制而成的[8]。表3是ArchivesCom的图像质量组、joint Libraries/AcIS committee提出的格式选择与压缩建议。

表2 图片数字化的格式建议

图像类别		色调深度(bit/pixel)	格式	压缩	空间分辨率	实例大小
预览图		8	GIF	Native to GIF	约150×100~200×200	20kb图像
参考图	灰度参考图	8	TFIF(JPEG的交换格式)JPEG	灰度图以10:1压缩	中级:500×400~1000×700 高级:1000×700~4000×300	81kb图像
	彩色参考图	24		彩色图以20:1压缩		
档案图像	灰度图	8	TIFF	不压缩	中级:500×400~1200×1000 高级:3000×2000~5000×4000 (仅高级图像用于存档)	1.3mb图像
	彩色图	24				

表3 文献数字化模式与技术指标

媒体类型	转换方法	分辨率	档案格式	屏幕显示格式	打印显示格式
黑白印刷文字文献	平台扫描仪或数码相机	1bit,600dpi	TIFF W/CCITT Fax4压缩	GIF,4bit,120~200dpi	PDF,1bit,300或600dpi

续表

媒体类型	转换方法	分辨率	档案格式	屏幕显示格式	打印显示格式
图表、地图、手稿等	平台扫描仪或数码相机	8bit 灰度或 24bit 彩色,200～300dpi	TIFF	Multiple JPEG, 24bit, 512×768, 1025×1516, 2048×3072, 质量级 50	JPEG, 24bit, 2048×3072, 质量级 50～100
拟以二维显示的三维对象	数码相机	24bit 彩色, 200～300dpi	TIFF	同上	同上
35 毫米黑白或彩色幻灯片或负片	PhotoCD 或幻灯片扫描仪	24bit, 2048×3072	PhotoCD 或 TIFF	同上	同上
大幅照片、幻灯片、负片或彩色透明片	ProPhotoCD 或滚筒扫描仪	24bit, 4096×6144	PhotoCD 或 TIFF	Multiple JPEG, 24bit, 质量级 50	JPEG, 24bit, 4096×6144 质量级 50
黑白缩微胶卷	缩微胶卷扫描仪	1bit,600dpi	TIFF W/Fax4	GIF,4bit 120～200dpi	PDF,1bit 300 或 600dpi
		8bit,300dpi	TIFF	GIF8bit 120～200dpi	PDF,8bit 300 或 600dpi

以上的这些技术指标并非是恒定的,随着转换技术的提高或新的文件格式的发展,这些数据都可能再次发生变化。这是一个不断变革的领域,只有与时俱进地研究与总结,才可得到较为理想的数字化效果。

参考文献

1 李慧.扫描数字图书馆.中国计算机用户,2001-06-19
2 扫描仪与数码相机共建数码影像输入立体化网络. http://www.sina.com.cn 新浪科技 2003-03-26
3 Technical Recommendations for Digital Imaging Projects. Prepared by the Image Quality Working Group of ArchivesCom, a joint Libraries/AcIS committee. http://www.columbia.edu/acis/dl/imagespec.html
4 Ester, Michael. Digital Image Collections: Issues and Practice. Washington, D.C. Commission on Preservation and Access, 1996. To order a copy, http://www-cpa.stanford.edu/cpa/publist.html
5,8 Digitization: A Literature Review and Summary of Technical Processes, Applications and Issues. http://www.library.ualberta.ca/library_html/libraries/law/digital.html
6 California Digital Library Digital Image Format Standards. http://www.cdlib.org/news/pdf/CDLImageStd-2001.pdf
7 Technical Recommendations for Digital Imaging Projects, Prepared by the Image Quality Working Group of ArchivesCom, a joint Libraries/AcIS committee. http://www.columbia.edu/acis/dl/imagespec.html

刘家真　武汉大学信息管理学院教授,博士生导师。通讯地址:武汉。邮编 430072。

徐　曼　武汉大学信息管理学院硕士研究生。通讯地址同上。

(来稿时间:2004-03-02)