●刘家真

馆藏文献数字化的原则与方法(下)*

摘 要 文献数字化对象的选择,主要应考虑文献价值、用户需求、文献形体状况、技术可行性及版权等因素。文献数字化转换技术的选择原则主要有:利于保护原件原则,满足用户需求原则,不同目的采用不同技术要求的原则,高仿真度原则,实行标准化和采用最优化方法原则,利于归档和长期保存原则,等等。对转格式、方式,以及采用光盘还是网络版,都要作出合理考虑。全信息采集是一种优秀的数字信息采集策略。数字化信息资源应予良好的维护。表 1。参考文献 11。

关键词 馆藏文献数字化 选择原则 转换技术 全信息采集
分类号 G250.76

ABSTRACT For the selection of the objects of document digitization, the main things we should consider include document values, user needs, physical attributes, technological feasibility and copyright. For the selection of conversion technologies, we have several principles. Besides, we should also consider formats and media. The author thinks that full information acquisition is an excellent strategy for digital information acquisition. 1 tab. 11 refs.

KEY WORDS Digitization of library collections. Choice principles. Conversion technology. Information acquisitions.
CLASS NUMBER G250.76

1 数字化对象的选择

不同类型的数字化工程,拟数字化的对象有所不同。在选择文献数字化对象时,主要应考虑文献价值、用户需求、文献形体状况、技术的可行性与版权等因素。

1.1 文献价值

只有具有持久价值的文献,转换后才有可能为用户长期使用,才有可能使其投资与其效益相符。文献是否具有持久价值是依人的主观判断的,主观判断又是可以随人与时间而改变的,因而文献价值的判断就十分令人棘手。

文献当前利用率的高低可否作为判断文献持久价值的客观因素呢?这个问题极为复杂。使用频繁的、当前利用率极高的文献,有的并不一定具有持久价值。例如,某些在当时、当地极令人感兴趣的新闻,一时存取率很高,但其使用价值可能是有限的。而利用率当前较低,也不一定意谓着该文献价值就低或今后仍然利用率低。有时很有价值的资料并非令许多人都感兴趣,例如,四书五经。又如,有价值的文献所处的位置太远、交通不方便也会造成利用率低下,或因其检索工具差也会造成文献不为人知而利用率低下。当这类文献数字化后以超越时空方式传送,形成新的用户群,则利用率就可能不再低了。因而,传统环境中文献利用率的大小只能作为判断文献价值的参考因素。为克服文献价值判断的主观性、局限性(仅由馆员决定),有必要在确定数字化对象前进行广泛的用户调研。

1.2 用户需求

使用频率高、用户需求大的文献应优先考虑数字化。这类文献的数字化,不仅有利于对原件的维护,也为用户的存取提供了便利,是受用户欢迎的。

1.3 文献形体状况

形体损坏或字迹模糊不清的馆藏应优先考虑数字化,以抢救文献的信息内容、避免信息在不稳定媒体上进一步地丢失,也有利于用户存取这类文献。

1.4 技术的可行性

不同物理形式或不同特征的原件,例如装订资料、散页资料、照片、底片、彩色或黑白文献等,模数转换技术对其信息转移的支持程度不同,有些暂无法

*本文为武汉大学重大科研项目研究论文。

达到满意的转换效果,也不宜数字化。数字化硬件、软件发展很快,价格也在不断调整。有时推迟对这类文献的数字化项目,以便等待技术发展到可以完好地解决这些问题时再去转换它更明智。例如,大辐面的地图数字化,若用相当高的解像力进行扫描,形成的数据如果超越了图书馆的计算机与网络的容量,则传送的效果必然不理想。等待宽带的发展与更大功率计算机出现,再去进行这类文献的转换,可能更恰当、更现实。

1.5　版权考虑

版权是制约文献选择的另一个问题。如果拟数字化对象受版权保护,要得到版权所有者允许后方可以数字化;如果受版权保护的文献未得到版权所有者许可,即使满足拟数字化对象的其他所有要求,也不可以数字化。数字化工程负责人必须对版权有充分的了解。IFLA 网站的"版权与知识产权"专栏上,有大量关于版权的讨论,包括有关论文、报告、白皮书、某些国家版权机构的有关信息以及有关知识产权的讨论等,可供我们参考。

2　转换技术的选择原则

2.1　利于保护原件的原则

目前,数字文献是不能被看成可以永久保存信息的。为防止数据丢失后文献的再次转换,必须保护好原件。在数字化过程中,应仔细地、慎重地操作原件,以免损伤原件。对于形体状况较好,无损伤的原件,在数字化过程中应尽量减少直接操作,以免损伤。例如扫描仪与数字相机都具有模数转换能力,但扫描仪对纸张具有直接压力,对易脆的纸质文献极不安全,使用数码相机更可取。这就是转换技术选择中的利于保护原件的原则。

2.2　满足用户需求原则

使用什么硬件来转换传统文献,还必须遵守满足需求原则。例如,对于文献价值而言,如果彩色信息极为重要时,就不宜使用黑白扫描仪处理,以保证满足用户需求。值得注意的是,使用什么转换设备除应考虑用户需要外,还必须考虑经费问题,因为数字化设备是昂贵的。

2.3　不同目的采用不同技术要求的原则

模数转换可以达到两种不同效果,一是转换成的数字拷贝将作为原件;二是数字化拷贝仅仅是作为原件的代用品,提供用户利用。对于前者,如脆化书刊数字化的目的就是使数字拷贝代替原件使用,这时的数字拷贝实际就成为了该文献的源文献。这类文献的数字化要求极高,应将原件中全部关键信息转换至满意的效果,以满足研究、法律与财政的所有要求。如果转换只需达到作为原件代用品的效果,则只需有效地采集原件信息内容,使数字拷贝比原件能更及时、方便地提供存取即可。

2.4　高仿真度原则

转换技术应使生成的拷贝对原件有较高的保真度。这需要在转换中仔细地关注文献的重要细节,以免丢失。为了使数字拷贝尽量地再现原件,除非弥补设备的缺陷,一般不宜对图像进行处理,图像增强不应过分,而应寻求使数字化后的文献达到准确的复制效果即可,不能造成原始信息的丢失。

2.5　实行标准化和最优方法原则

数字化信息在进行技术平台转换中,标准格式很少出问题;标准能推进与简化部门间的合作。但标准也存在一定问题,例如技术发展太快,使得花费多年认可的标准,不可避免地落后于技术发展。此外,商业化竞争促进了专利工具的开发,使标准难以实现。在当今这样一个高度分布与可变化的环境中,十分完整地实现标准化是理想化的。为了适应以上情况,人们越来越寻求建立最优方法解决这些问题。为了保证数字化后的主文件将来可以广泛使用,为了便于数字拷贝的管理与利用,像缩微胶片的制作一样,模数转换必须符合标准。在不存在标准的地方,如元数据保存、持久稳定标识符号领域,应采用最优方法。

2.6　利于数字文献的归档与长期保存原则

数字技术变化太快,硬件、软件很快就淘汰,只有在数字化处理过程中充分地注意这一问题,才可能有利于数字信息的长期保存与存取。数字文献归档在图书馆是一个较新的领域,但极为重要,特别是学术研究图书馆。许多图书馆在数字化处理过程中,不注重数字归档的技术处理,而依赖计算服务中心,后患无穷。当前,英国许多机构已经在研究这一问题,如大学研究图书馆联盟(CURL)承担了数字文献归档的研究任务,对 CD-ROM、网址、动态 email 列表或其他数字资源的归档进行了研究,确定了各种资源在数字图书馆中的归档模式。

与数字化文献长期保存及归档有关的另一个问题是,文献数字化处理过程中应慎用压缩技术。文件压缩可以充分地减少文件大小,便于文件的传送、存取及拷贝,但压缩程度越大,数据丢失风险越高。

因此,对于数字文献的归档存储不提倡压缩。除非对于非常大的文件,如数字视频,方可考虑使用压缩格式,以利于传送或存贮。

3 转换格式与方式的选择

3.1 格式要求

馆藏文献的数字化处理中,有多种可供考虑的格式,如 word、TIFF、JIPEG、GIF、PDG 等,这些格式各有其特点。文件格式发展很快,但淘汰也很快。回溯文献的数字化处理的费用是相当高的,我们期望它们能有长期效用,因而在格式的选择上首先考虑的问题是这种格式是否有利于数字文献的长期存取。标准化格式能使数字文献在不同技术平台上迁移的风险减至最小,并可缩小数字文献保存的风险与费用。此外,使用被广泛采纳的格式,也可以最大限度地减少文献长期存取的风险,因为业界会为该格式提供迁移路径或反向兼容,以有利于技术平台的转换。但应当指出的是,反向兼容只能维持上一代或两代的版本兼容,对更早的版本是无兼容能力的。因而,馆藏文献数字化格式的确定,最好采用标准格式或采用广泛使用的格式。

为了兼顾利用与保存,欧洲国家一些图书馆、档案馆针对数字格式是否有利于长期保存,将格式分成三个层次:可以接受的格式、最优格式与不可接受格式。鉴于保存与利用的不同要求,将存档格式与提供给用户使用的格式分开,同一文献用不同格式提供利用与保存。如 TIFF 主文件可用于存储高解像力的数字图像,可作为归档存储格式;而小容量的文件可用 JPEC 格式(存取格式)用于网络传输的分发。然而 JPGE 图像由于涉及到难以避免的数据损失,不可用于存储。转换格式的选择可参考以下选用原则:(1)无论是存储格式还是存取格式,均应使用非专利化格式;(2)有选择的使用完善的、已广泛使用的、市场上的"事实标准"格式;(3)尽可能将归档存储格式、对用户传递格式区分开来并区别使用;(4)尽量减小使用格式的种类,以简化管理过程与降低管理费用;(5)对归档存储的文件,尽量不要加密或压缩。

3.2 转换方式考虑

数字化最简单的途径是使用扫描或数码相机形成原始文献的数字图像文件,其中用扫描仪是最经济的。

另一种转换方式是将文献转换成文本,以文本方式存贮文献内容。这种转换方式有直接键入法与格式转换法。直接键入方式可以形成 ASC Ⅱ数字化文本文件,ASC Ⅱ文本很适于使用关键字或短语检索。对某些数字化工程而言,回溯文献数字化的目的就是为了便于检索。当数字文件可以用字、词进行检索时,目录、字典与索引就会特别便于使用,数字文本很易于其他用户共享、共操作。但用键入方法形成的数字文本难以复制原始文献的结构与版式,如果没有专门代码,用户难以直接查到某章某节的内容。

也可以将扫描形成的图像文件,经 OCR 软件处理形成文本。经 OCR 软件处理形成的文本,可产生便于检索的索引,实现全文自动检索。此外,还可用 HTML 语言将 OCR 形成的文本编辑上网,进行全文检索。但其缺点是 OCR 只能较好地识别印刷体,并不能识别所有其他字体,对于文字极为复杂的中文,识别力有限,会造成部分不识别字体的数据丢失,只适用于某些字体的文献的数字化工程。

应当说明的是,将回溯文献转换成文本格式要比转换成图像文件所占内存少,但费用高得多。此外,无论以哪种方式形成馆藏文献数字化的文本形式,均无法再现原文献的版式。为了弥补这一缺陷,有人将文献正文以 ASC Ⅱ格式或置标文件形成,然后链接到原文献的数字图像文件上,让读者按需存取或利用这种方式去由读者判断转换后的文本是否可靠。

3.3 光盘版还是网络版

在考虑转换方式的同时,还应考虑数字化后信息的传递方式。是以 CD-ROM 还是以网络传递。CD-ROM 与因特网上可存取信息的存储与分发方式是不同的,其区别主要体现在硬件、软件要求与使用的难易上。有时,CD-ROM 与检索、分析软件捆绑在一起,这点比因特网上的文献易于使用。但 CD-ROM 存取仅限于专用工作站或小网络,而网上文献可以适于广大用户。此外,与 CD-ROM 不同的是,不需要用户要求,因特网就可以更新而取代淘汰款式。因特网的主要问题是,没有好的检索工具,查找十分困难。究竟采用什么方式,应考虑到文献数字化的预期目的、文件大小、人员与设备等。

4 全信息采集

全信息采集(fall Information capture)是国外推荐的一种信息采集策略,它是可以确保数字图像的高保真、良好的功用性,以及实施付费合理的数字信息

采集策略。现简介如下,以便国内同行参考。

模数转换的费用是很不一致的,有关的存贮要求、处理要求都会影响到费用。全信息采集是处理模数字转换的,既考虑形成图像的高质量与功用性,又使费用最小的经济模式。只要设法使模数转换处理与原始内容的关键信息相匹配,既不过多也不太少地进行信息采集,就可以达到图像的高保真度。如,对于文献上的某个点,即使不断地提高解像力,可能对图像质量也无法产生可感知的效果,而费用无疑是提高了。要做到"恰当"采集文献信息,使形成的图像文件具有较高的保真度,关键在于怎样寻找与确定拟数字化对象的关键、重要信息,以及如何妥当地采集到这些关键、重点的信息。使它能准确、真实地再现原件。

为达到上述目的,在着手转换之前,应对拟数字化文献进行仔细分析,找出能充分显示文献信息内容的、关系到文献意义的关键特征。这项工作类似于原始文献的鉴别,对图书馆员与技术人员的鉴别力有较高要求。由他们所提供的关键特征,经采集后能使原件的主观特征与控制数字转换的客观规范(如解像力、位深度、图像增强与压缩等)相关联,使其能恰如其分地真实、完整地再现原件信息,使原始文献的重要信息全部采集下来。表 3 列出了对于印刷资料与照片来说具有重要意义的特征信息。

表 1　对存取具有重要意义的原始文献特征

	装订或未装订的印刷资料	照　片
关键特征	文献尺寸(宽 ×高,英寸)	版式(35mm,4 ×5 等)
	细节尺寸(最小值)	细部与边缘再现
	文本特征(亲笔文件、印刷品)	干扰
	插图(内容与处理)	动态范围
	色调(包括颜色)	色调再现
	动态范围,密度,对比度	颜色再现

5　对数字化后的信息资源导航与维护

印刷资料可凭目录的章、节、页码与索引为用户导航,数字化后这类导航工具不再方便了,因而应以易为用户存取与操作的方法进行整理。对已经被一页页扫描的文献,每页都必须单独标记、存贮。该文献的关键页码、例子、书名页与每章首页都应链接到电子导航工具上,以便定位。数字化工程的链接程度取决于它预计中的利用情况,馆藏文献数字化款目可与目录文件链接,也可与国家目录数据库或其他查找工具链接等。为使馆藏文献数字化收到较好的效益,图书馆应制作馆藏文献数字化索引。

馆藏文献数字化最低要求是原始文献的数字化拷贝应比原件更方便地为用户提供利用,而且应随时通过以下方式扩展它的使用:(1)为了能在新的传送背景下使用,应将其处理成新的或不同的数字版本;(2)将辅助检索工具与对"增殖"资料的解释增加到联网传送环境中;(3)随时编码与校对可检索文本;(4)编排与并入"增殖"的分析与评论;(5)添加增强级索引与导航;(6)添加与其他数字资源的链接与互操作;(7)只要可能,就用新的标准与数化方法对数字文献进行处理。

馆藏文献数字化具有可促进文献存取传递及达到保护文献的作用,受到广泛的关注。在我国,图书馆馆藏文献的数字化处于幼年时期,经验不多,为了使馆藏文献数字化真正达到预期目的,使投入与获得的价值一致,还必须很好地组织决策,不断地探索前进。

参考文献

1　邹家华. 加快推进国家信息化. 求是,1997(14)
2　http://www.thames.rlg.org/Preserv/matrix.heml
3　董焱,刘兹恒. 图书馆馆藏文献数字化. 图书情报工作,2000(7)
4　D. Hazen. Selecting research collections for digitization (Washington,1998), p. 18.
5　A Case for Fall Information Capture: http://www.dlib.org/cornell/10chapman.htmL
6　National library of Australia Digitization Policy 2000～2004. http://www.nla.gov.auy poliay/digitization.htuil
7　NDLP Project Planning Checklist: http://Lcwbz.loc.gov/ammem/prjplan.html
8　Guidence for Selecting materials for digitizstion. http://www.thames.rlq.orq/joint/cofpaper.html
9　A case for full Information Capture. http://www.dlib.orq/dlib/10chapman.html
10　University of California Selection Criteria for Digitization. http://www.library.ucsb/ucriaq.html
11　Preservation Digital Reformation Program. http://Lcweb.locqov/preserv/prd.presintro.heml

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

(来稿时间:2001-05-08)