

●臧国全

数字迁移风险管理*

摘 要 数字迁移是数字资源长久保存的策略之一。为尽可能减少迁移过程和结果所带来的风险,应构建风险管理框架,对数字迁移风险进行评估和测度。图1。参考文献6。

关键词 数字资源 数字保存 数字迁移 风险管理

分类号 G250.76

ABSTRACT Digital migration is a strategy of the long-term preservation of digital resources. To minimize the risks caused by digital migration, we should construct a risk management framework and make evaluation and measurement for the risks of digital migration. 1 fig. 6 refs.

KEY WORDS Digital resource. Digital preservation. Digital migration. Risk management.

CLASS NUMBER G250.76

1 数字迁移缘起

数字资源的长久保存是图书馆信息资源数字化项目的重要目标之一。然而,保存数字资源的文件格式不断更新,使得采用当前浏览软件难以甚至根本无法阅读和访问以前格式保存的数字资源。为保证用户采用当前浏览软件进行有效访问,可以采用数字仿真法(Emulation)和数字迁移法(Migration)。

数字仿真法也可称环境改变法,并不改变数字资源本身,其目的是构建一个采用新的软硬件技术访问原格式数字资源的环境。数字仿真法的关键是提供采用原来浏览软件来访问原始格式数字资源的一个工具,这个工具称为仿真工具。仿真工具在新的环境中工作,使原来的浏览器和原来的工作平台适应新的环境。它的最大优点在于保持数字资源的外观,最大缺点是开发和维护这种仿真工具非常复杂,也很困难,成本太高。在实际应用中,要维护多个仿真工具,但即使这样也很难确保这些工具对未来计算机平台的有效性。

数字迁移法也可称格式转换法,着眼于数字资源本身,保证软件与硬件的发展不影响存取。这就要通过改变或更新数字资源的格式,使它适应于新的软件和硬件环境。数字迁移法的实施是在数字资源的文件格式出现过时危机时进行的。例如,如果技术发展表明PDF1.1版本将很快不能访问,所有用该格式保存的数字资源就要被转化为新的版本格式(如PDF1.4版本格式)。这样,数字资源就进入了下一个可被访问的生命周期,直到PDF1.4版本格式出现不可存取

危机,这时就要进行下一个数字迁移过程。数字迁移法的主要优点是:数字资源总是以能够被普遍接受的格式被人们使用,现行的硬件与软件能够毫无障碍地访问数字资源,且可以被复制。数字迁移法的主要缺点:其一是在格式转化过程中,数字资源的一些样式(layout)甚至一些数据将会丢失。如果数字资源的原始样式需要重点保存,迁移法或许不是最好的选择。其二是它的操作对象一般是数字资源集合中的单件资源,如果数字资源集合庞大,转化过程需很长,工作量很大。其三是如果数字资源的文件格式已经作废或所用的转换工具不再有效,那么采用这种方法进行格式转换就不可行^[1]。采用数字迁移策略,必须不断开发并运行转换程序,以免数字资源的格式过时。

2 数字迁移原因与类型

数字迁移的原因有很多,大部分与文件格式有关。非结构化的或没有格式的文件是一个简单的二进制数据流,软件开发商对数据文件进行结构化,从而允许他们的软件对数据文件进行有效地读写数据操作。软件应用越来越复杂,文件格式的种类也快速增长,新格式或格式的新版本不断地取代已有的格式或版本,使得以过时格式存储的数字资源必须进行不断地迁移,确保用户对其存取。

数字迁移主要有4种类型。

(1)数字文件的日常刷新。若干年前,数字资源从一种存储介质转存到另外一种存储介质是迁移的主要形式。随着存储介质稳定性和可靠性的提高,这

* 本文为国家社科基金项目“图书馆信息资源数字化建设模式研究”(编号05BTQ007)研究成果之一。

种形式的数字迁移越来越少。

(2)当数字文件从一种应用转到另一种应用时,需要改变数字文档的格式。这种迁移形式的一个例子是数字文档从苹果机操作系统的应用转到 Windows 操作系统的应用。

(3)数字资源格式本身的完全改变。比如,将 Word 格式文件转换为 ASCII 文件。

(4)从数字主文档格式制作副本。一些情况下,数字化主文档不适合公共存取,这时就要生成适合存取格式的文件副本。比如,存储的是直接扫描格式 TIFF,为了发布和公共存取需要,可能要转换为 PDF 格式的文件副本。

3 数字迁移风险分析

作为一种数字资源保存策略,数字迁移的过程和结果都具有不确定性。为尽可能减少这种不确定性所带来的风险,方法之一是构建一个风险管理框架。该框架将整个迁移过程分解为若干个可描述的且可计量的步骤,风险评估就是将风险分析过程进行结构化的一种方法。如果这种方法描述和使用得当,不同人对数字文件的相同信息进行评估,所得到的风险值应该基本相同。

数字迁移风险主要有 3 个方面。其一是与数字资源管理相关的风险。这类风险包括缺乏机构支持、资金、系统软硬件和数字文档管理人员等。这些都是数字资源管理的重要组成部分,数字资源乃至其拥有者都在不同程度上受数字迁移影响,与数字资源有关的法律政策也会增加迁移的风险值。其二是与数据文件格式相关的风险,包括在迁移过程中容易被修改的文件内部结构要素。其三是与格式转换过程相关的风险。格式转换软件的转换结果可能理想,也可能不理想,转换错误或有或无,或大或小^[2]。

如果将上述 3 个方面进行细化,数字迁移风险包括 8 项具体内容。

(1)内容稳定性风险(指二进制数据流的结构)。包括:软件虫、对存储介质误操作和设备机械故障等因素导致二进制数据流的破坏;新的压缩方式所产生的文件格式对二进制数据流配置的改变;文件头信息没有迁移,或迁移不完整,或迁移错误;数字图像的质量(分辨率、颜色等)受到二进制数据流配置改变的影响;新的文件格式改变了字节顺序等。

(2)安全性风险。格式迁移影响了采用水印和数字戳等加密技术对原始数字资源进行加密的信息。

(3)内容连贯性和完整性风险(与其他相关文件

的关系,或与数字环境中诸如软硬件依赖性之间的关系)。包括:由于对不同软硬件的依赖性,浏览和访问新的文件格式需要新的配置;与其他文件的链接(如元数据文件、脚本等)在迁移过程中被改变;由于文件格式组织或新的压缩方法导致新格式的文件体积减小,使存储更密集,并产生潜在的文件目录结构问题;由于新存储介质或操作系统的文件组织协议导致文件存储更密集,影响标签和文件结构等。

(4)参考性风险(在数字图像集合中检索特定数字图像的能力)。包括:由于文件格式的更新和受 URL 的影响,文件扩展名被改变;迁移活动没有被完整记录,导致来源信息不完整或不精确,这对未来的迁移活动带来潜在问题。

(5)成本风险。包括:数字迁移性质的差异可能导致每个迁移周期涉及不同步骤,因此,长期保存的数字迁移成本具有不可预测性;由于信息资源的价值难以精确衡量,因此迁移的优先顺序无法设置;迁移成本有时难以精确计算等。

(6)人力资源风险。包括:在整个数字迁移决策过程中,员工的轮换和缺乏连续性,导致长期计划难以实施,尤其是在元数据抓取不充分以及缺乏完整迁移过程记录情况下;员工缺乏足够的技术技能;迁移周期的难以预测性,使得人力资源需求规划(如技能、时间、资金等)更加困难等。

(7)功能风险。包括:新文件格式对原文件格式数字资源的一些功能产生影响;数字迁移增加或减少原始文件的性能以及需要对检索界面进行修改,如新格式要求对 Web 界面的支持;新文件格式不支持的原文件格式中一些功能在迁移过程中丢失;原文件格式的一些信息增值在迁移过程中丢失等。

(8)知识产权风险。产权管理限制新格式产生的一些数字资源的应用,比如一些数字资源的产权许可仅仅限定在一定分辨率范围内的图像发布,而新格式生成的图像超出了这个分辨率范围^[3]。

4 数字迁移风险评估与测度

数字迁移过程是原文件通过转换软件转换成目标文件。

转换软件有 3 种类型:其一是为一个特定数字迁移项目由技术人员编写的程序;其二是为一个特定目的而编写的商业软件,如抽取不同格式原文件中的一些数据生成一个新格式数据文件;其三是大众化商业转换程序,如在苹果机和 PC 机文件格式之间相互转

换的软件。它们各有优缺点。第一种类型针对性强,但开发周期长且费用高,商业性软件虽然成本低廉,但针对性一般来说较差。

转换软件的主要功能在于实现数字资源的格式转换,在实际应用中风险值最高,对这种类型风险进行评估包括4个方面。

(1)同构数字资源转换过程风险评估。其方法是对比数字文件转换前后的格式属性,关键在于需要知道原始文件格式的所有特征,然后在转换后的新格式文件中寻找这些特征的存在情况。比较过程由人工进行,虽然很费时,但最准确。

作为这种评估方法的一个变通,选择具有代表性的测试文件进行格式转换,然后人工计算这些文件转换前后格式属性的吻合情况,并以此为参照点评估整个数字资源集合的转换风险。这种评估方法适合于原始文件的格式是一种类型的数字资源集合。

(2)异构数字资源转换过程风险评估。可以将异构数字资源集合分解为若干个同构数字资源子集合,采用上述方法对每个子集合进行风险评估,再汇总为整个集合的风险值。如果这个数字资源集合包括的文件种类和数量庞大,这种方式费用昂贵又缺乏效率。

另一种方法是采用一个文件阅读器。首先界定每种类型格式文件迁移风险的格式特征,训练文件阅读器完全准确地识别这些特征,然后用该阅读器遍历检查每个文件迁移前后的特征值。检查过程中,如果发现目标文件与原文件格式特征值不一致,就记录下来。记录内容包括文件的名称、文件的位置、与文件相关的风险特征的类型和数量等。检查一批文件后,将这些记录汇总成报告输出。一个好的文件阅读器应该具有足够的灵活性,能够阅读几乎所有格式的文件,能够识别定义的所有风险特征。

这两种方法的出发点都是要完全准确地获取文件的格式特征信息,然而大多数文件格式都具专有性,致使获取这类信息非常困难,而这又正是数字迁移风险评估的关键。所以,应该提倡文件格式的开放化和标准化。

(3)转换软件风险评估。如果存在多个转换软件供用户选择,且每个软件都能提供主要的乃至全部的核心性能,同时又提供一些可选性能,这时就需要设计一些指标对这些软件进行评估,以便选择出适合于特定数字迁移项目的转换软件。

一个理想的数字迁移软件至少应该具备下述性能:能够分析原文件格式与目标文件格式的差别,能

够识别和报告迁移风险的等级,能够精确地将原格式文件迁移到目标格式文件,能够进行单个文件迁移也能够进行成批文件的迁移,能够提供迁移记录等。

(4)元数据风险评估。在对数字资源主文档迁移时,必须考虑相应的元数据迁移。离开元数据,主文档可能就没有任何信息意义。元数据的文件格式可能与主文档的格式完全不同,比如,主文档格式是TIFF,而描述该主文档的元数据文件格式是HTML,如果主文档的名字或地址在迁移过程中发生变化,那么元数据中的相应内容就必须改变,否则,主文档就无法被存取^[4]。

数字迁移风险的测度有两个指标:风险概率和风险影响。Gregory W. Lawrence等人将风险概率分为5个等级^[5]:非常高(5级,风险概率在26%~99%之间),高(4级,风险概率在11%~25%之间),中(3级,风险概率在6%~10%之间),低(2级,风险概率在1%~5%之间)和非常低(1级,风险概率低于1%)。风险影响也分为5个级别:灾难性(E级,信息完全不可逆转丢失,无法从其他资源中再生成),非常严重(D级,信息部分不可逆转丢失,并且不可能从其他资源中再生成),严重(C级,信息完全丢失,可从其他形式的资源中完全再生成),重大(B级,信息部分丢失,可从其他形式资源中完全再生成),较小(A级,信息部分或完全丢失,但都可从其他数字文件中复制)。

这样,针对某件特定数字资源进行数字迁移的风险主要有5个等级:5E(风险概率非常高,且将导致灾难性影响),3D(风险概率中等,且导致非常严重的影响),2C(风险概率较低,且导致严重影响),1B(风险概率非常低,且导致重大影响),1A(风险概率非常低,且导致较小影响)。图1是一个详细的数字迁移风险测度二维图^[6]。

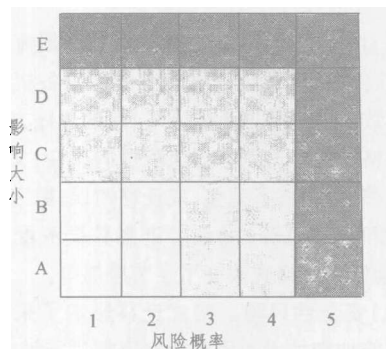


图1 数字迁移风险决策 (下转第86页)

据。有的采用 CNMARC, 有的用 MARC21, 还有的采用 DC 元数据。这一切不利于资源共享, 不利于加强 OPAC 的查找、识别、选择以及获取功能, 更不利于全面地揭示作品、内容表达、载体表现和单件之间的层次关系, 方便用户查找特定的信息资源。

为了实现 FRBR 用户任务, 强化 OPAC 功能, 控制数据库质量, 避免现有的编目规则可能导致数据库建设的缺憾, 笔者建议: 第一, 摆脱中国传统分立目录的束缚, 将《中国文献编目规则》和《西文文献著录条例》合二为一, 统一中西文文献编目规则。第二, 加强联机数据的质量控制和规范文档建设工作, 为实现 FRBR 用户任务做好基础工作。第三, 以 FRBR 书目实体模式为目标, 开发基于 FRBR 模型、深度揭示信息资源的 OPAC。

只有这样才能进一步加强信息资源编目的标准化, 促进检索点的规范化, 加强数据库建设的整体化。最终达到实现 FRBR 理念, 完善 OPAC 功能, 共建共享信息资源的目的。

参考文献

- 1 富平. 继承与变革: 谈《中国文献编目规则》的修订. 国

家图书馆学刊, 2005(2)

- 2 中国图书馆学会《西文文献著录条例》修订组. 西文文献著录条例: 修订扩大版. 北京: 科学技术文献出版社, 2003
- 3 国家图书馆《中国文献编目条例》修订组. 中国文献编目规则: 第二版. 北京: 北京图书馆出版社, 2005
- 4 国家广播电视总局. 广播电视音像资料编目规范 第1部分 电视资料. 2005-03-31 http://www.abp.gov.cn/bi-aoshun/soft_show.asp?id=344 [2005-08-27 查询]
- 5 中国高等教育文献保障体系管理中心. 关于联合目录项目. <http://www.calis.edu.cn/calis/Ihtml/Ihtml.asp?fid=FA01&class=5> [2005-08-17 查询]
- 6 IFLA Study Group on the Functional Requirements for Bibliographic Records. Functional requirements for bibliographic records; final report. [S. l.] IFLA Section on Cataloguing, 1998. <http://www.ifla.org/VII/s/frbr/frbr.pdf> [2005-08-15 查询]
- 7 谢琴芳. FRBR 用户任务对 CALIS 联合目录编目实践的影响. 见: 中国图书馆学会. 以人为本 服务创新. 北京: 北京图书馆出版社, 2005

殷明莲 北京大学信息管理系教授. 通信地址: 北京. 邮编 100871. (来稿时间: 2005-10-20)

(上接第 56 页)

图 1 中, 如果测度值落在 1A~B 至 2A~B 之间, 数字迁移风险较小, 迁移可进行。如果测度值落在 1C~D、2C~D、3A~D、4A~D 之间, 数字迁移过程风险很高, 迁移应该推迟, 直到这些风险降低。如果测度值落在 1E、2E、3E、4E 和 5A~E 之间, 数字迁移应该禁止。

数字迁移的风险可以计量, 风险值的大小在不同的迁移条件下会有所不同, 有时差别可能会很大。产生风险的主要原因在于原文件和目标文件格式特征的差异。可以采用不同方法, 评估和计量格式特征迁移风险的等级。对大多数文件格式来说, 文件的基本结构具有共性, 所以对一种格式的特征分析可以用来理解其他格式。

相对于文件格式迁移风险存在较大区别来说, 组织机构、软硬件和元数据的迁移风险大同小异。数据迁移项目的各种要素风险在项目实施之前都必须评估和计量。

在数字迁移项目中, 最大的困难是对风险的性质和风险值的解释, 比如, 什么时候某种风险可以接受, 能够接受的风险值是多少等。风险评估工具不能代替经验, 更不能代替有效且正确的判断。风险评估工

具不是万能的, 即使最完善的工具也不能确保数据迁移万无一失。

参考文献

- 1 Caplan, Priscilla. Building a digital preservation archive: Tales from the front. The Journal of Information and Knowledge Management Systems, 2004(1)
- 2 Euhliir, Paul. Framework for the Preservation of and Public Access to USDA Digital Publications. <http://preserve.nal.usda.gov:8300/npp/frameprt.html> [2005-07-25 查询]
- 3,5 Lawrence, Gregory W. et al. Risk Management of Digital Information: A File Format Investigation. <http://www.clir.org/pubs/reports/pub93/pub93.pdf> [2005-07-29 查询]
- 4 Haynes, David, et al. Responsibility for Digital Archiving and Long Term Access to Digital Data. <http://www.ukoln.ac.uk/services/elih/papers/supporting/#blric>. [2005-08-10 查询]
- 6 McNamee, David. Assessing Risk Assessment. <http://www.mc2consulting.com/riskart2.htm>. [2005-08-10 查询]

臧国全 武汉大学博士后, 郑州大学信息管理系教授, 博士. 通信地址: 郑州市大学路 75 号郑州大学信息管理系. 邮编 450052. (来稿时间: 2005-09-23)