

● 宛 玲

电子文档长期存取的可跨媒体开放文件格式*

摘 要 PDF/A 和 ODF 是采用与平台和设备无关的开放文件格式标准,有利于低成本高质量进行电子文档的长期存取。作为新出台的国际标准,它们在电子文档长期存取方面具有重要作用,可以实现异构系统的互操作,降低维护长期存取活动的成本,消除对特定软件厂商的过度依赖并促进办公软件的发展,在后续的发展上具有很好的透明性和可靠性。但二者在应用中仍然存在较大的差异,开放性也有一定的局限。参考文献 12。

关键词 PDF/A ODF 文件格式 国际标准 电子文档 文档存取
分类号 G253

ABSTRACT PDF/A and ODF are open file formats independent from platforms and equipments, and can be used for long-term preservation of electronic files with low costs and high quality. As new international standards, they will play important roles in long-term preservation of electronic files. They can realize interoperability of heterogeneous systems, lower costs for long-term preservation, eliminate dependency on specific software vendors, promote the development of OA software, and have transparency and reliability in future development. However, they still have some differences in applications, and have some limitation in openness. 12 refs.

KEY WORDS PDF/A. ODF. File format. International standard. Electronic file. File access.

CLASS NUMBER G253

电子文档长期存取的基本要求是:无论在什么时候,人们都可以方便、可靠、经济地利用当时的软硬件读取早期创建的电子文档,甚至在特殊条件下可以对其进行编辑处理。众所周知,数字环境下存在着连续不断的“过时”现象,包括软件过时、硬件过时和媒介过时等。目前已经出现多种应对的技术方法,如迁移、仿真、风干(Dessication)、UAF(底层抽象格式)^[1]、更新等,但每种方法都存在一定缺陷,且成本高,技术和管理复杂。为了以方便、经济的方式解决问题,解决文件格式问题以延长电子文档存取寿命不失为一项重要措施。

1 电子文档长期存取数字格式的问题及 PDF/A 和 ODF 的产生

目前解决文件格式问题的方法主要有文件格式转换,辅助工具为数据格式登记系统。如当原有文件格式存在过时危机时,可以通过数据格式登记系统的自动提醒系统或查询系统,获知应当转换的格式,并可利用其提供的转换工具和服务进行直接转换。数字格式转换属于迁移技术,从技术角度看,优点主要在于无论软硬件如何变化,总是可以通过迁移技术保证数字对象的可再利用,缺点是无法保证原有文件

的特征或数据全部迁移,甚至有可能破坏或改变数字资源的显示、结构和含义;从管理角度看,文件格式转换是针对个体文件进行的,当针对大量文件时,转换就成为密集型劳动,成本则水涨船高。上述问题提出了文件格式转换的风险管理要求,即对最佳转换时机的选择,而这种选择本身就是一个难题,另外转换工具也存在失效的危险,因此减少数字格式转换频率是重要的选择。

减少文件格式转换频率和成本的方法之一就是采用共用文件格式,或以共用文件格式转换中介。在市场经济环境下,封闭的私有格式很难成为基本的共用文件格式,应当采用开放的文件格式。目前这类格式已有 RTF、HTML、XML、TEXT、PDF 以及最近刚刚兴起的 PDF/A 和 ODF。RTF 文件体积庞大,易丢文件的样式和版面,丢失资料内容和逻辑,无法恢复原样;TEXT 格式、HTML 格式、XML 格式具有 RTF 格式的后三种缺点,同时 HTML、XML 文件冗长^[2],没有被业界列入长期存取用格式,但基于 XML 格式的平台无关性以及元语言性质,发展基于 XML 的开放文件格式成为趋势。PDF 是由美国 ADOBE 公司开发的一种通用的支持多种压缩方式的文件格式,具有原状保持性、兼容与开放性、以及支持比较、鉴别和多

* 本文为河北大学人文社会科学研究项目(编号 0509017)成果之一。

级别安全控制等特点^[3]。然而从电子文档长期存取角度看,PDF 格式仍然存在着相当大的问题,如 PDF 格式是由 ADOBE 一家公司所有和控制,属于私有格式,这就意味着 ADOBE 公司可以在任何需要的时候改变格式,引入新的功能或限制;另外,PDF 允许加密和批注,允许嵌入声音和动画,而这些对文档的归档处理可能带有不利因素^[4]。

1.1 电子文档存档国际标准 PDF/A 的产生

为了充分利用 PDF 的优点以及人们对 PDF 的熟悉,同时解决 PDF 存在的上述问题,2002 年美国印刷出版及纸品加工设备器材供应商协会(NPES)和美国国际图像信息管理协会(AIIM)共同发起,由 ADOBE 的代表和来自其他团体尤其是美国政府和法律界的代表组成了团队,联合研制一个新的经过瘦身的 PDF 版本,即 PDF/A。该版本于 2005 年 5 月被 ISO 批准为正式的电子文档存档标准格式(ISO19005-1),并于当年 10 月正式出版。

PDF/A 实际上是 PDF1.4 版规范的一个子集,其目的是定义一个文档格式以保存电子文档静态外貌,提供一个记录电子文档的元数据框架,以及提供定义电子文档逻辑结构和语义属性的框架^[5]。它规定所有文档的字体必须包含在一个 PDF/A 文件中,不可利用其他商业性字体,没有加密和批注功能,不可嵌入多媒体等,这些规定保证了 PDF/A 文件对系统软件和应用软件的独立性。

PDF/A 包含有知识产权,ADOBE 公司通过 PDF reference 和 XMP specification(元数据规范)声明了这些知识产权,并许可这些产权可被人们永久、免费和非排他性地利用,从技术和权利上实现了任何人都可以利用 PDF reference 和 XMP specification,对 PDF/A 文件进行阅读、呈现、书写或其他处理。PDF/A 中包含基于 XML 和 RDF 规范的元数据平台 XMP,提供描述、起源、保存和管理的功能。XMP 元数据作为纯文本被嵌入到每个文件中,以减少内容丢失的可能性,并且可以通过元数据进行简单检索。

目前 PDF/A 提供了两个层次保存规范:PDF/A-1a 符合 ISO19005-1 的全部要求,包含了结构和语义标签;PDF/A-1b 符合 ISO19005-1 最基本要求,保存 PDF/A 文件的表现形式。现在公布的标准是 PDF/A 的第一版本,PDF/A 团队已经开始了以 PDF1.6 为基础的 PDF/A 版本 2 的开发工作,希望在新版本中实现更多的功能,比如交互式表格、某种安全功能等^[6]。

1.2 办公应用电子文档格式国际标准 ODF

ODF(Open Document Format)是基于 XML 的电子文档格式,涵盖了文本、电子表格、图表和绘图文档所需要的特征。2003 年,OASIS 联合澳大利亚国家档案馆等十几家办公应用机构组织了一个技术委员会,研发基于 XML 的办公文件格式标准,其工作基础是 OpenOffice.org XML 文件格式。经过多次改版,2005 年 5 月 ODF 正式成为 OASIS 的标准,并被提交到 ISO,2006 年 5 月被 ISO 批准作为国际标准,标准号为:ISO/IEC 26300。

OASIS 是一个非营利联盟,长期以来致力于公共 XML 标准的开放发展。这次它发展 ODF,目的就是为了创建一个开放的、基于 XML 的文件格式规范,改变当前办公软件相互封闭、文档格式互不兼容的不良状况。ODF 格式可以让人们在不同程序、平台之间自由地交换文件,而无须理会文件是由何种应用程序产生,保证了基于 ODF 格式的文档在未来仍然可以被任意一款支持 ODF 格式的最新办公软件打开使用。目前主要的支持厂商有 Corel、IMB、Opera、甲骨文、红帽、国内软件厂商中文 2000 公司等,SUN 开发了免费下载办公软件 OpenOffice.org,其缺省文档格式即为 ODF。

2 PDF/A 和 ODF 的应用价值

PDF/A 和 ODF 文件格式具有多种功能,在技术方面对长期存取电子文档有很大的优势。本文主要是从二者的开放性、标准性和跨媒体等方面分析它们的应用价值和存在的问题。

2.1 实现异构系统的互操作和降低维护长期存取活动的成本

IT 行业充满诱惑,吸引着无数有才智和理想的人们为之奋斗,人们研制出了大量且不同的文件格式以及相应的软件工具,实现了表现和处理各种内容数据和利用的需求。然而,事物的发展总是有着两重性,文件格式的泛滥以及它与特定软件的捆绑,严重地影响了人们对不同软件系统下产生的数据的共享,影响了人们对特定格式内容数据的长期使用。PDF/A 和 ODF 都是独立于系统软件和应用软件的数字格式,它们的出现正好迎合了人们共享数据的需求^[7],满足了人们在组织内部以及不同组织之间交换信息的需求。

数字格式转换,或多或少都会产生某些方面的失真,这种失真随着转换次数的增加而增大。同时,在实践领域,数字格式转换的成本与文件量密切相关。一个开放标准不会轻易过时,因而 PDF/A 和 ODF 格式比私有格式更可靠和稳定,人们可以开放和利用各种系统软件长期存取 PDF/A 文件和 ODF 文件,减少格式转换频次,实现迁移成本下降和数据及特性损失的减少。对最终用户来讲,可以方便地读取甚至加工以往各时期的文档数据。

长期存取活动需要尽可能避免原有电子文档服务能力的降低,一个有效的措施就是保存数据的支撑运行环境,包括其知识组织机制、链接机制、交互处理机制等提供的丰富服务功能^[8],但这种措施存在技术难度大、经济成本高以及知识产权复杂等问题。采用开放文件格式标准,在保存可完整描述支撑运行环境的元数据条件下,未来系统开发者利用开放文件格式规范和必要的元数据,可以构建同样甚至更先进的支撑运行环境。另外,为了保证在未来可以定制读取软件,保证长期可靠地读取购买的数据库内容,要求数据库商提供数据格式参数是图书馆的必须选择。目前出现了开放文件格式标准,图书馆可以仅要求数据库商提供符合该标准

的数据,在未来,由于对标准的采用,市场上会存在多种支持平台,图书馆可从中选择而不必自行研制。从经济学角度看,这种现象加强了社会分工,降低了社会整体生产成本,提高了数据处理能力,保证了软件产品的质量。

2.2 消除对特定软件厂商的过度依赖并促进办公软件的发展

开放性的一个突出特点就是消除人们因对特定软件产品和厂商的依赖而产生的恐惧感。当依赖一种私有的文件格式时,支持它的软件产品和厂商的兴衰直接影响它的使用寿命。而开放文件格式标准,不仅能够而且也使更多的厂商放心地开发满足不同应用需求的产品,用户也无需担心万一开发这些语言或格式的厂商出现问题下一步如何应付^[9]。由于可以有許多厂商自由开发支持软件,人们可以选择不同的软件产品,不用担心特定支持软件的过时和变化。同时由于它的开放性,人们特别是政府机构不必担忧文档中存在什么隐藏信息,安全性得到提高。

虽说市场上分割不均,但开放文件格式有助于打破办公软件系统市场的垄断局面,避免文件格式受一两家厂商所左右,给大量从事办公软件开发的人们提供了更大的发展空间和竞争机会,研发出更为优秀的产品。对我国来讲,也有利于发展自己的软件行业,提高我国信息化程度。

微软的 OFFICE 已成为世界著名办公软件系统,使用广泛,功能强大。我们可以相信,作为世界顶级 IT 企业,在近期甚至在当代人有生之年,微软都不会放弃对 OFFICE 文档的支持。然而,微软 OFFICE 文档格式毕竟是私有格式,我们无法预料在遥远的未来,人们依然可以方便地使用微软 OFFICE 文档。因此作为国际标准的开放文件格式,应为未来提供强有力的保险。

2.3 后续发展具有很好的透明性和可靠性

PDF/A 和 ODF 都是由一个联盟性质的团体进行维护和发展。PDF/A 是由包括了软件专家、法律代表、政府人员等组成的团队进行研制,并不完全属于 ADOBE 公司。ODF 是由 OASIS 技术委员会和各类办公产品卖家参与发展,目前 ODF 开放文档技术委员会由来自不同国家的 17 个基金委员会组成,该团队特别注意收集各方面的反馈信息,并且欢迎人们与技术委员会合作共同发展 ODF。这种不是由一两家主导的研制形式,更能反映市场和社会的需求。更为可靠的是,由于二者都是国际标准,因此后来的版本无论增加和修改什么功能和条件,都需要经过 ISO 的批准,保证与早先版本的兼容。

技术是不断发展的,我们不能说 PDF/A 和 ODF 就是电子文档长期存取的唯一选择,不能说未来不会有更适合的文件格式,但由于目前二者已经被认定为国际标准,且具有了一定的市场。因此,即使未来出现新的更完善的格式,也容易在市场和社會的需求下,相互兼容。如目前微软正是在这种压力下,同意让 OFFICE 系统支持 ODF。

2.4 PDF/A 和 ODF 之间的差异

PDF/A 与 ODF 同是刚批准的国际标准,同样具有维护电子文档长期存取的功效,但它们在应用中仍然有着较大的差异。

正如 PDF/A 标准名称所言,PDF/A 格式是电子文档存档格式标准,即其主要功效是保证所保存内容在时间轴上的一致性。PDF/A 自带字体和呈现内容所需的组件,在浏览器上不可编辑等功能,很好地实现了这种一致性。因此 PDF/A 主要应用于拒绝被修改和编辑的档案管理领域和文件公开发布领域,如图书馆、档案馆等机构对电子文献的保存,团体机构或个人对文件、报告和数据进行发布或传递。

ODF 是办公文档开放格式。在办公环境中,人们很有可能需要对很久以前的文件进行内容更新,对其中的数据或格式进行调整,或者利用其中的数据进行排序、绘图、索引和分析等,这些都是 PDF/A 所不能和不允许的。而 ODF 作为一种办公文档格式,支持办公软件实现这些要求。另外,ODF 是系列格式,包括文本、报表、表格、图像、数学公式、幻灯片等格式,实现办公中的不同需求。很显然,ODF 更多是从办公应用角度设计的,适用于办公环境。

3 应用前景及应注意的问题

3.1 应用前景

PDF/A 和 ODF 都具有广泛的应用空间。PDF/A 是 PDF 的一个子集,基于 PDF 的广泛应用,人们对 PDF/A 格式不会有陌生感觉。哈佛大学数字图书馆经理 Stephen Abrams 参与了 PDF/A 的研制,他认为 PDF/A 的发布对电子文件保存产生重大影响,新的标准将鼓励图书馆、文献馆和其他数字信息使用者,有效地运用标准格式保存珍贵文献和传递重要文件^[10]。

据 ODF 联盟网站(<http://www.odfalliance.org/>)提供的信息,为了推广 ODF 在全球范围的应用,IBM、Sun 等公司联合创建了 ODF 联盟,刚开始仅有 36 个成员,到 2006 年 7 月份 Google 加入后已有 241 个会员。ODF 联盟的迅速壮大说明了 ODF 为社会接受的程度。另外,自 2006 年 5 月 ODF 成为国际标准后,短短 2 个多月,已有不少政府机构提出将 ODF 作为政府办公用标准格式,而且这个数量在不断增加,如比利时政府和马来西亚政府都提出要将 ODF 作为政府内部交换文件的标准格式,西班牙政府提出近 10 年里要将电子文档都转换为 ODF 和 PDF 格式,美国马萨诸塞州政府要求政府机构最迟在 2007 年 1 月 1 日前皆需安装支援 ODF 格式的应用支持软件,并淘汰其他产品。欧盟信息通讯管理委员会(EU/TAC)建议欧盟各国采用以 XML 为基础的公开文件格式标准,而 ODF 符合了这一要求。在我国,ODF 开放文件格式已被一些有远见的厂商使用,比如中文 2000 公司一直致力于中文版 OpenOffice.org 的开发,其主打产品 RedOffice 3.0 办公套件就使用 ODF 格式。据新闻报导,目前 Sun 和其他 5 家公司开发了提供 Office 文件格式与 ODF 之间兼容性的插件,微软则表示它将免费向开发这类插件的第三方

开发商提供技术文档和知识产权^[11]。

由于技术、市场、面市时间等方面的原因,现在的软件对 PDF/A 和 ODF 还没有足够的支持。目前 ADOBE 的浏览器可以浏览 PDF/A 格式文件,但生成 PDF/A 格式文件的软件还不太成熟。2006 年 7 月 13 日 PDF Tools AG 宣布已经开发了支持创造和管理 PDF/A 文件的工具 3-Heights™ PDF Producer,不过该工具对 PDF/A 的支持还仅限于 PDF/A-1b。美国 Apago 宣布开发了可以将 TIFF 和 PDF 文件转换为 PDF/A 的工具 PDF Appraiser,利用该工具还可以鉴别一个 PDF 文件是否符合 PDF/A 的要求。PDF 首创者 ADOBE 开发的 Acrobat 7.0.7 可以支持 PDF/A。不过除了 ADOBE 的浏览器,其他的都不是免费工具。对于 ODF,目前支持软件包括 OpenOffice.org 2.0, StarOffice 8 Office Suite, IBM Workplace, KOffice, TextMaker, AbiWord, WordPerfect,但要人们去熟悉使用这些软件还需一个过程。针对 ODF 开发的免费产品 OpenOffice.org 2.0 具有高度可用性、开放性、跨平台等特点,可以生成和读取包括 ODF 在内的多种格式文档。不过目前它还不完善,如笔者利用它将 DOC 文档转换为 ODF 文档时,就出现了有的文档样式出现偏差,无法利用金山词霸对 ODF 文档进行屏幕取词等问题。

3.2 应当注意的问题

虽然 PDF/A 和 ODF 都属于开放文件格式,但它们的开放有一定局限性。开源代码的开放,一个重要的特点就是其他人可以有条件地对源代码进行更改。但 PDF/A 和 ODF 的开放性,主要表现在格式规范是公开的,别人没有改版的权利,仅能表达自己的建议^[12]。

根据 ISO19005-1,将原文从其他格式转换为 PDF/A 格式时,并不能保证内容再现的精确性,可能会发生字体变换、图像有损压缩。这时就需要利用其他方法,制定审视表现原文精确性的政策并实施是非常重要的。另外,目前存在大量不同的文档格式,如何低成本高质量向标准的开放文件格式转换,还需要从技术和管理层面上进行研究。对图书馆而言,在采购数据库时应当主动积极要求数据库商提供符合这类开放文件格式标准的数据。

PDF/A 和 ODF 虽然已经被批准为国际标准,但如同其他国际标准一样,需要各国的认可,甚至要与各国制定的相关标准兼容才能得到广泛的推广。据计算机世界网(www.ccw.com.cn)2006 年 8 月 18 日的报道,目前我国中文办公软件基础标准工作组正在探讨新制订的国家标准《中文办公软件文档格式规范》与 ODF 互相兼容的途径。微软 OFFICE 得到普及,不仅仅与它的产品性能有关,各种传播方式,如学校教育、职称考试、业务培训等也起到重要作用。国家和社会也需要积极通过多种途径鼓励开放文件格式标准的应用。作为开放获取运动的成果,互联网上已出现多家开放信息资源网站,如中国科学技术信息研究所与国家科技图书文献中心联合建立的中国预印本服务系统、教育部科技发

展中心建立的中国科技论文在线、中国国家图书馆建立的图书馆学情报学开放文库等。如果这些开放网站能够采纳和支持 PDF/A、ODF 等各种开放文件格式标准,不仅可以促进开放文件格式标准的应用和发展,更有助于网站长期保存和开放所存储的信息资源。PDF/A 和 ODF 本身在技术上是否完善还需深入研究,但它们的开放性和标准化迎合了时代的需求,反映了社会对异构系统下共享信息、保存信息的需要。

参考文献

- 1 吴振新,张智雄,郭家义. 数字信息资源长期保存技术策略分析. 现代图书情报技术,2006(4)
- 2 洪宗胜. ISO ODF 标准应用展望. [2006-08-18]. <http://linux.tca.org.tw/seminar/20060627/03ISOODF%BC%D0%B7%C7%BBP%C0%B3%A5%CE%AE%B1%E6.pdf>
- 3 付斌. PDF 文档格式与电子文件管理. 北京档案,2003(3)
- 4 Marcus Uneson. Tomorrow's File Endings: On Archiving Principles and Archiving Formats. [2006-08-18]. http://www.sciecom.org/sciecominfo/artiklar/uneson_05_2.pdf
- 5 Stephen Abrams, Betsy Fanning, Diana Helander, Susan Sullivan. CRM. PDF/A The Development of a Digital Preservation Standard. August 14-21, 2005. [2006-08-18]. <http://www.aiim.org/documents/standards/PDF-A.ppt#285,26,Pdf/A>.
- 6 Adobe PDF/Archive 通过 ISO 批准成为国际标准. [2006-08-18]. <http://it.sohu.com/20050713/n240159609.shtml>.
- 7 Edward M. Corrado. The Importance of Open Access, Open Source, and Open Standards for Libraries. [2006-08-18]. <http://www.isrl.org/05-spring/article2.html>.
- 8 宛玲,张晓林. 数字资源长期保存过程中的知识产权问题分析. 中国图书馆学报,2005(3)
- 9 皮卓丁. 数字出版与跨媒体解决方案. [2006-08-18]. <http://www.myadobe.com.cn/newdoc/file.php?articleid=738>
- 10 奥多比公司. Adobe PDF/A 成为 ISO 国际标准的文献档案格式. [2006-08-18]. <http://www.adobe.com/tw/aboutadobe/pressroom/pr/oct2005/NR-051006.Pdf>
- 11 Sun,微软回应 ODF/Office 转换程序的要求. [2006-08-18]. http://www.ccw.com.cn/news2/software/htm2006/20060713_196750.htm.
- 12 Document management—Electronic document file format for long-term preservation—Part 1: Use of PDF 1.4 (PDF/A-1). [2006-08-18]. http://committees.standards.org.au/COMMITTEES/IT-021/N0001/ISO_19005-1-2005.pdf.

宛玲 河北大学管理学院副教授。通信地址:河北保定五四东路 180 号河北大学管理学院。邮编 071002。

(来稿时间:2006-08-22)