

电子文件管理与长期保存

学 院：信息资源管理学院
专 业：2013 级档案专业
日 期：2016 年 6 月 24 日
作 者：肖文、段言、李晨阳

2013级档案专业档案信息化课程

1.0 电子文件定义

电子文件是指在数字设备及环境中生成，以数码形式存储于磁带、磁盘、光盘等载体，依赖计算机等数字设备阅读、处理，并可在通信网络上传送的文件。它主要包括电子文书、电子信件、电子报表、电子图纸等等。

2.0 电子文件长期保存格式介绍

2.1 电子文件种类

文本文件、数据文件、图形文件、图像文件、视频文件、声音文件、命令文件、多媒体文件

2.2 电子文件格式

电子文件格式指将信息以某种格式进行编码，而且只有依靠特定组合的软硬件方能对文件实施读写操作。按电子文件的格式划分，有数字文本文件、数字影像文件、数字图形文件、数字语音文件、数字视频文件。文本文件常见格式有 XML、TXT、UC、RT、WPS、PDF/A、CEB、SEP；图像文件常见格式有 TIFF、JPEG2000、JPEG、GIF、PNG、Djvu；音频文件常见格式有 WAV、MP3、WMA、OGG Vorbis、FLAC；视频文件常见格式有 AVI、MOV、MPEG 等。

3.0 电子文件长期保存的标准保障

3. 解决电子文件长期保存格式问题的管理措施

3.1 限定接收进馆的电子文件格式类型

3.1.1 电子文件长期保存格式的选择标准

1、按格式种类分析

要作为文本型电子文件的归档格式，马春茂认为应该符合下列条件：（1）应该是公开的格式，不依赖特定的系统和软件存在，能够在可以预见的时间内被正常读取；（2）应该保持文件生成时的排版格式，即使经过办公软件的不断升级，也能输出同纸质文件一样的格式；（3）应该便于数据交换、便于查找和检索；（4）应能保证电子文件在处理过程中不被篡改和窃取。

为了对数字资源实施长期保存，对保存图像格式的选择尤其重要。叶新民认为应满足格式使用的标准化和广泛性；格式应是公开的，而非私有的，而且应公开已出版的技术资料；格式应支持元数据；格式应对存储介质无依赖性要求。

对于数码录音的电子文件长期保存格式，张文浩认为应支持真实性验证技术、数码音质无损、文件格式开放透明、不绑定软硬件、格式自描述、格式自包含、易于存储、格式紧凑、信息安全、代价最小等要求。

数字摄像(DV)、数字电视(DTV)和网络视频系统的广泛应用以及对现有模拟视频的数字化，视频电子文件的数量和类型不断丰富。项文新等提出从保证视频文件的原真性和档案保管利用的便捷性出发选择合适的视频电子文件归档格式，认为视频电子文件应尽可能保证视频文件的原真性。

2、综合性分析

对于长期保存人类文化遗产的档案馆、图书馆，最关心的是数字信息的内容不得丢失。刘家真针对电子文件保存格式的选择问题，认为该格式能在不同的环境下使用；应是通用的、非专用的格式；支持数据从专用格式的环境中移出；最好是标准格式；被业界或用户广泛支持；具有可扩展性；可真实完整地被用户读出并理解，无论何时何地，用户所查到的该文件信息都与该文件最初情况完全一致；具有自身的可证明性；具有可评价性。

档案是历史的真实记录，具有凭证价值与情报价值。李泽锋认为档案数字化过程中选用电子文件格式应考虑：保持档案原貌，利于网络平台下的传输交流，具有一定的标准性，具有较好的通用性和可移植性，选用电子格式的支持软件应支持数据从专用格式中方便地移出。

我国档案行业标准《板式电子文件长期保存格式需求》(DA/T 47-2009)明确规定了档案系统对电子文件长期保存格式的需求：支持真实性、格式透明、不绑定软硬件、格式自包含、格式自描述、固定显示、持续可解释、持续可用、可转换、易存储。

3.12 几种适于长期保存的电子文件格式

1、PDF/A 格式

PDF 格式规范已经公开，并可免费获取，具有广泛的支持性。该格式具有与平台、字体无关的特点，解除了电子文件与操作系统的依赖关系，用户可以通过它的通用浏览器 Adobe Reader 在 Windows、UNIX、Mac 等任一系统平台上原版原式地读取 PDF 文档，不受操作系统、网络环境、应用程序的版本、字体的限制，从而从电子文件自身的角度保证了电子文件的可读性。PDF 格式所具有的高度兼容性使用户可以轻松地将任何文档原版原式地转换成 PDF 文档，进而固化信息，使之不易更改，保持了原档信息内容和结构的完整、真实，为不同软件生成的电子文件提供了一个可完整迁移其信息内容的统一文件格式。PDF/A-1a 实现了对文档逻辑结构的保存和以自然阅读顺序保存文本。PDF/A-1b 可实现对文本和附加内容的正确显示。在各国政府机构包括美国、英国、德国、新加坡、印度、澳大利亚等的电子政务领域中，PDF 格式也被广泛应用。综合来说，PDF/A 格式的特点是支持数字签名、格式开放、不绑定软硬件、格式自包含、格式自描述、固定显示、不包含加密、可向其他文本格式转换等。

2、扫描性电子文件格式

(1) **TIFF 格式**。TIFF(标记图像文件格式)是由 Aldus 和 Microsoft 公司为桌面出版系统研制开发的一种灵活的位图图像格式，用于应用程序之间和计算机平台之间交换文件。实际上已被所有绘画、图像编辑和页面排版应用程序所支持，而且几乎所有桌面扫描仪都可以生成 TIFF 图像。TIFF 支持多种编码方法，其中包括 RGB 无压缩、RLE 压缩、LZW 压缩、ZIP 压缩、CCITT 压缩等，6.0 版本又增加了 JPEG 压缩选项。TIFF 格式由于存储图像质

量高非常有利于原稿的复制。另外，所有绘画、图像编辑和页面排版应用程序以及大多数扫描仪对 TIFF 格式都提供良好的支持，这使得 TIFF 格式成为数字图像处理的最好选择。

(2) JPEG2000 格式。JPEG2000 于 2001 年正式成为国际标准。JPEG2000 完全采用小波变换代替余弦变换，与 JPEG 相比，具备了更高的压缩比以及更多新功能。在文件大小相同的情况下，JPEG2000 压缩的图像比 JPEG 质量更高，其压缩率比 JPEG 高 30% 左右，精度损失更小。

JPEG2000 的特点是：(1) 同时支持有损和无损压缩，而 JPEG 只能支持有损压缩。(2) 能实现渐进传输，先传输图像的轮廓，然后逐步传输数据，不断提高图像质量，让图像由朦胧到清晰显示。(3) 支持对图像中细节较多或较为感兴趣的区域进行高精度的无损编码、很强的容错性、支持水印等。(4) JPEG2000 的许多优点和新的功能都是建立在复杂的计算与较大的缓存基础上，因此速度与 JPEG 相比要慢许多。

3、音频电子文件

FLAC 格式。FLAC 属于无损音频压缩文件格式(文件名为.flac)。FLAC 格式文件较小，便于播放。OGG 计划支持 FLAC，也可将 FLAC 音频数据封装在 OGG 传输层中，生成 OGG FLAC 格式文件(文件名为.ogg)。OGG FLAC 格式文件较大(比 OGG Vorbis 格式文件大得多)，便于传输、编辑、播放。该格式的特点：支持真实性验证技术、数码音质无损、文件格式开放透明、不绑定软硬件、文件格式自描述、格式自包、易于存储、信息安全。

4、多媒体电子文件

MPEG 格式。MPEG 即 Moving Picture Expert Group(运动图像专家组)，用该编码器编码的影音文件扩展名可以是 MPG、MPE、MP4、M1V、M2V 等，其中 M1V 和 M2V 都表示该影音文件中不包含音频部分，只有视频部分。这种格式也流行了很长时间。未压缩的 AVI 格式文件经 MPEG-1、MPEG-2 和 MPEG-4 等编码压缩后，图像容量大幅度缩小，画质则降低不大，而且图像质量视编码参数而定。该格式具有压缩算法公开、不绑定软硬件、易于转换等特点。

3.2 建立电子文件格式登记系统

电子文件格式登记系统是指能对数字文件格式的属性进行描述、识别、记录和保存，支持信息系统通过公开发布和规范的机制去发现所需要的文件格式信息，进而识别、转换、展示和保存相应文件的系统。电子文件格式登记系统保存了各种常见的文件格式以及创建、运行和迁移这些文件格式的软件的相关技术信息。通俗地讲，电子文件格式登记系统就是给文件格式“上户口”。

(一) 电子文件格式长期保存的必要性

1、文件格式生命周期缩短

虽然各国档案馆在接收电子文件时都严格规定了电子文件接收格式，但由于文件格式具有生命周期，信息技术发展速度越快，文件格式生命周期就越短。当前新技术的飞速发展使得一些文件格式在较短的时间内就遭到废弃或被取代，文件格式的生命周期越来越短。所以，即使是电子文件长期保存格式，我们也无法确保几十年或几百年后它是否还适合长期保存。当这些格式被废弃或被取代后，以这种格式保存的无数的电子文件就会面临无法读取的危险。

2、特殊文件格式受专利保护

在一些专门领域，会形成很多特殊的、专门的格式，一些特定的电子文件只能通过这种格式才能被解码并显示、读取，而这种格式又只能通过特殊的软硬件设备才能有效地运行。但由于这些格式为某单位私有或受专利的保护，因此未向社会公开。当这些电子文件向档案馆移交时，由于档案馆并没有支持此格式运行的软硬件设备，所以以这种特殊的专门格式保存的电子文件在档案馆中就成为“死档”。另外，当这种格式由于更新换代遭到废弃时，因为没有关于此格式的相关信息，所以当用户需要利用这份文件时却不能读取它，使其失去利用价值。

3、历史文件格式未及时转换

传统的档案鉴定原则没有涉及电子文件格式的鉴定，因此早期的一些电子文件在向档案馆移交时并没有进行格式的鉴定和转换，从而使得档案馆中保存了大量的以不利于长期保存的格式存在的电子文件。由于文件格式的更新换代和硬件等设备的不断升级，以及对传统电子文件格式的忽略，一些特定格式的电子文件就同样面临成为“死档”的危险。

因此，针对这些问题，最有效的办法就是建立一个电子文件格式登记系统，把文件格式及与其相关的软件、技术等信息一并登记，以便电子文件格式及电子文件的长期保存和查询。

（二）电子文件格式登记系统的功能

1、防止格式过时带来的信息损失

电子文件格式登记系统能跟踪文件长期保存由于格式过时可能造成的潜在信息损失，当文件有可能出现过时的情况时，系统的查询和自动提醒系统能够帮助我们得到应当转换的格式，然后利用系统提供的转换工具和服务直接转换。

2、持久维护文件格式的描述信息

电子文件格式登记系统支持对文件格式的公共描述以及这些描述信息的持久维护，从而在被保存对象、保存技术与保存环境的不断变化中能识别、解析、迁移、转换和保存各种格式的数字资源。

（三）电子文件格式登记系统的案例

1、PRONOM。PRONOM 是提供关于电子文件长期保存所必需的关于文件格式、软件产品和其他技术要素等方面的公正、权威信息的在线技术登记系统。它最初是英国国家档案馆为

支持其自身的电子文件接收与长期保存而开发的,现在已经发展成为通过网络向所有人开放的资源库,它保存了软件产品信息与产品可读写的文件格式信息,包含格式的标式信息、环境信息、保存规划信息与其他技术参数,而且个人还可以通过在线方式提交新的文件格式到 PRONOM。

PRONOM 网站与英国国家档案馆网站的整体风格一致,简洁美观,而且使用非常方便简单,无需专门培训;PRONOM 运用权限控制,采用用户名和密码限制特定人员进入管理系统,安全可靠;PRONOM 的主系统和公共系统能够支持超过 100 名用户同时利用系统,性能强大;在检索方面,PRONOM 支持七种检索方式:简单检索、格式名称/扩展名检索、PUID(PRONOM 永久唯一标识符)检索、软件名称检索、组织名称(格式开发组织)检索、生命周期检索(即在某一个日期被支持的文件格式、软件名称)、迁移路径检索等。

2、MIME MTR。MIME MTR(多用途网际邮件扩展协议媒体类型注册-Multipurpose Internet Mail Extensions Media Types Registry)是目前最多被使用的格式登记系统,由 IANA(互联网数字分配机构 Internet Assigned Numbers Authority)管理维护,提供关于 Internet 上传文件的类型和格式的权威、标准的定义信息、编码信息、参数信息等。

MIME MTR 主要登记关于 MIME 文件的属性信息,目前共有 9 种媒体类型,分别是应用文件、音频、示例、图像、信息、模型、多媒体类型文件、文本、视频。每个媒体类型还有子媒体类型。登记的格式属性包括 MIME 类型名称、MIME 子类型名称、编码要求、必备参数、可选参数、发布说明、安全要求、联系人、互操作要求、所支持的应用等。

MIME 允许在线登记,用户可以通过浏览的方式来查看文件类型和类型定义。如果用户要了解相应的格式,链接的格式信息可以被用来定义文本,获得查验和解读格式技术参数。

3、GDFR。GDFR(全球数字格式登记系统,Global Digital Format Registry)在美国哈佛大学主持下建立,是为了提供持久的存储、发现、利用文件格式表征信息的分布式服务。它将文件格式解释成一种稳固的以位序列编码的信息模型,本质上是一种可交换的正式的知识表达,具体的内容流与属性可以通过一定的表示信息来说明。不管哪一种文件格式,都可通过某个格式分类体系注册于某个格式登记系统,用唯一的标识符来标记这三种文件格式。

GDFR 建立了特有的格式对象数据模型,包括四类属性:特征属性、一般描述属性、操作属性、管理属性。每个格式的基本属性包括:标识符、别名、分类、关系、创建者、所有者、维护者、工具、状态、定义说明、签名、来源和注释等。

GDFR 主要提供四类服务:管理服务,发现新格式和弃用格式,评审、发布新格式,修改、删除废弃格式;检索服务,依照检索请求提供特定格式的描述信息,向特定的存储库推送登记的格式信息;表示服务,确定并验证特定电子文件的格式;中介服务,转换格式,提取特定格式的电子文件中的元数据。

4、UDFR。由于 GDFR 和 PRONOM 的格式登记不是由同一地区主导,具有一些局限性:PRONOM 是英国国家档案馆的一个部门来承担的,所以协作相对容易,但是它的实施范围不够广;而 GDFR 是由哈佛大学领导,多个机构共同参加的项目,它的沟通不够充分,技术性也没有 PRONOM 好。这种情况下,国家 IT 和档案领域相关专家认识到建立一种全球统一的文件格式登记系统是有必要的。

UDFR(统一数字格式登记系统, Unified Digital Format Registry)由美国加州大学内容管理中心带头研发,加州数字图书馆提供资金支持。它融合了 GDFR 和 PRONOM 的优点,数据来源于 MIME 和 PRONOM。2009 年 4 月正式成立,2012 年 7 月投入生产使用。UDFR 是一个开源的、基于语义可用平台的系统,该系统是为了收集、长期管理和传播与人们利益相关的具有重要意义的文件格式属性。UDFR 登记的格式属性包括文件格式、字符编码、压缩算法、标识符、外部签名、内部签名、MIME 类型、代理商、软件包、软件过程、知识产权报表、关系等。

4.0 电子文件长期保存的技术策略

4.1 电子文件长期保存的层次

4.1.1 比特流保存

存即保存纯粹的数据,只存储原始数字对象,维护原始比特流的完整性和长期可存取性,不考虑数字对象的外观、感知和相关功能。一般是利用复制技术,采用多重设备,异地备份方式,提供一种简单的数据备份存储能力。通常不能直接为用户提供访问服务,在需要将备份数据上载到一定系统上提供服务,这种保存有时也被称为隐晦存储。

4.1.2 内容保存

内容保存是指保存内容数据的同时,还保存与内容数据存储和使用相关的元数据。由于数字资源的利用依赖于对其格式和处理方式的理解,因此仅保留纯粹的内容数据不能保证数字对象在未来的可用性和可靠性,为避免“死数据”的出现,除了对内容数据进行备份存储外,还必须保存相关的元数据,并对内容数据进行定期上载、确认、使用检验,以保证数据及其相关功能的完整性和可使用性,实现对数字对象外观、感觉、语义等方面的完整保存。

4.1.3 服务保存

服务保存指除了保存内容数据及其相关的元数据之外,还保存数据的支撑运行环境,以确保能重现原有的信息服务,实现对数字资源的再利用。数字资源与相应的服务系统已经很难区分开来,知识组织、链接机制、交互处理机制等信息服务功能能够使数字资源得到实质性利用,只有“复原”原始系统提供的信息服务功能才能真正实现数字资源可持续使用。

与前两种数字保存层次相比,保存“支撑运行环境”更为彻底,难度也最大。数字保存的上述 3 各层次由保存对象的类型及技术特性、保存目标等因素决定。实现不同层次的数字保存技术主要采用不同的技术策略。如采用复制或者更新技术策略可实现对数字对象的比特流保存,而采用仿真、迁移策略一般能实现对数字对象的内容保存。

4.2 电子文件长期保存技术策略及其选择

4.2.1 技术的改进

4.2.1.1 改善载体状况

由于电子文件是承载于磁盘和光盘上的。因此还在数字世界初期人们将电子文件保存的关

键归于载体的保存上。于是人们为电子文件保存创造了适宜的环境诸如环境温度应控制在 14℃—24℃之间相对湿度控制在 45—60 之间存放载体的柜、架及库房应达到有关标准的要求载体应直立排放并满足避光、防尘、防变形的保存条件远离强磁场和有害气体等等。另一方面由于载体的腐坏与脆化会带来信息的丢失即使保存条件再好载体的使用寿命也是有限的。于是人们呼吁厂商提高产品(载体)质量增强载体的耐久性、贮存能力及使用寿命以便极大地减少电子文件信息的丢失及降低贮存费用。有研究表明磁光载体的使用期限为 10—30 年。磁性载体磁粉不稳定易变形而感光载体又容易受温度、湿度激烈变动影响以及尘埃颗粒污染的危害。因而人们十分看重一种新型的载体——HD—ROM 光盘这种高密度的只读光盘是 CD—ROM 光盘存贮能力的 180 倍而价格只是它的一半。它采用离子束将电子文件信息写在不锈钢、铱等载体上。HD—ROM 不易受材料变质的影响并能以人们可读形式描述解释数据。这种方法展现了在新存贮技术下保存载体的发展前景。不过增强载体的耐久性、贮存能力及使用寿命这只能作为一定时期内的保存措施而非永久的解决方案。

4.2.1.2 电子文件的更新

日新月异的数字技术使电子文件的技术环境及存贮结构更加复杂。新设备、新的处理方法与软件以 2—5 年为一个周期更新换代因而使得记录、存贮与检索电子文件的手段与产品迅速发生变更。又由于技术更新太快。大量非标准的信息剧增使得电子文件离开了特殊环境就失去了可读性。在这种背景下再增加电子文件载体的寿命已显得不太实际了。事实上技术淘汰要比电子文件载体的腐坏对电子文件长期保存具有更大的威胁。针对这一点人们便采用“更新”技术即通过复制将电子文件信息从一个载体拷到另一个载体上。更新技术改变了电子文件因载体问题无法读出的状况从而使电子文件能够延续地保存下去。更新技术涉及到了许多问题如法律、经济、文化、技术、社会道德等。因此要使更新能够顺利地进行还有赖于多方面的合作和支持。

4.2.1.3 电子文件的转换

如今人们生活在一中纸张、缩微胶卷、录像制品、磁盘和光盘等多种媒体共存的环境里这样就必然要面对格式转换的问题即将电子文件信息从各种各样的格式上转换到更易管理的标准格式中。因此人们需要迅速快捷、价格低廉、质量更好的转换技术。目前可采用以下三种方法来进行转换。一是在电子文件归档时把电子文件转换成各种平台上都能适应的文本文件格式以消除由于技术演变所带来的影响二是目前许多数据库管理系统的制造商都在将自己的产品向通用型、开放型发展这些产品可以在多种不同类型的硬件及操作系统平台上得到支持因此可以将电子文件建立在这种通用的数据库平台上三是采用或开发对应的转换软件。每当所应用的技术与所保存的电子文件在技术上的一致性发生变化时就对所保存的电子文件进行相应的转换使其始终保持同新技术的兼容性从而保证电子文件逻辑上的一贯准确。在强调转换技术的同时人们还建议使用一些标准格式采用标准格式有利于电子文件格式的转换有利于电子文件的保存和利用。

4.2.1.4 利用再生性技术

这里所述的再生性保护是指将技术过时的电子文件适时地转移到缩微品或纸上。转移到缩微品上可能更好一些因为它存贮量大便于保管。不再使用计算机软硬件技术读取。这种方法为长期保存电子文件信息提供方便并避免计算机软硬件技术过时所带来的任何麻烦。缩微胶片的寿命理论值 500 年国际上已有 100 多年的实践经验国内也有几十年的实践经验。这对我们来讲可能是最希望采取的方法。但这种方法也存在一些现实问题一是有些信息无法转移到缩微胶片或纸上如声音信息、超文本信息、多媒体信息等二是信息一旦转移到缩微胶片或纸上也就失去数字记录的风格和魅力如失去了数字化的传递与使用的灵活性。但在缩微摄影技术和设备大发展的今天缩微品记录的模拟量信息能很快地进入计算机网络系统又恢复它原有的风格和魅力。

4.2.1.5 电子文件的迁移

在今天硬件、软件高度变化的环境中对电子文件长期保存多数情况应采取“迁移”的方法。迁移能保留电子文件信息的完整性保留客户对电子文件信息的检索能力、显示能力及其它利用能力。当然把电子文件信息迁移到新技术上的费用是非常昂贵的但对于利用率高、具有长期保存价值并符合长期标准的电子文件应当有理由实施数据迁移技术。

4.2.2 管理的改善

4.2.2.1 建立计算机技术档案馆

建立一个过时技术的计算机档案馆收集技术过时的计算机软硬件提供利用读取那些现在已经过时的技术在当时记录在存贮载体的信息例如收集过时的磁带驱动器和过时的操作系统。这种方法要求广泛收集、保管过时技术的全套软硬件还必须保护操作这些过时技术的某些技能。由于设备的老化原生产厂家不再生产这些设备的备件存贮载体在不断损坏等原因使得代价十分昂贵不是文件保管部门自己力所能及的可以采取的措施。而且即使采取这种措施也很难永远维持下去。

4.2.2.2 对电子文件进行选择 and 鉴定

目前各种类型的电子文件在层出不穷地产生 ■ 我们要对浩如烟海的电子文件进行保存是不可能也没有必要的。因为在众多的电子文件中有不少属于“垃圾文件”、“伪文件”、“有害文件”等等档案部门对此根本就无收藏的必要。而对于有价值的电子文件归档保存则是档案部门的历史责任。因此对电子文件档案保存的第一步就是要对电子文件进行选择以决定这些电子文件是否有必要进行保存。初步选择的电子文件还必须请有关专家进行鉴定经过鉴定有长期保存价值的电子文件可转化为电子档案进行保存对于保存价值尚不明确的则不需要进行全面的保存。电子文件处于频繁的变化之中电子文件选择的难度首先在于怎样寻找适合于馆藏原则的电子文件。然后才是对符合用户需要的潜在资源进行选择和鉴定。这样对电子文件应找出哪些重要的电子文件正在产生这些电子文件有无保存价值及由谁去保存。总之对电子文件进行选择和鉴定可避免电子文件长期保存中的盲目性和随意性从而保证信息资源的合理开发与利用。

4.2.2.3 确保电子文件的真实性、完整性和安全性

由于电子文件随时可以变换载体或被不目痕迹地更改所以为保证电子文件的真实性必采取专门措施。如有的办公软件用电子印章或电子签名的方法对电子文件进行签发处理。还有的办公自动化网络中可自动对每份电子文件的修改、存取、传输等任何操作都进行实时登记以密码形式存储并有防删改措施。为了使电子文件保存完整性 ■ 应当从电子文件形成起就进行妥善保存定期把符合归档条件的电子文件信息联机传递在计算机网络上集中并按档案管理标准格式存储到可长期保存的脱机载体上作为档案长期保存的电子文件应与相同内容的纸质文件或硬拷贝一同归档。具有保密要求的电子文件上同时必须符合国家或部门的规定 ■ 要有必要的安全保密措施以确保电子文件的安全。

4.2.2.4 电子文件的利用

电子文件长期保存的目的最终是为了利用为了用户对电子文件的目前尤其是长期的检索需要。与传统文件不同的是能长期保存电子文件并不等于能保证被保存的电子文件具有可存取性和可读性。被保存完好的电子文件若在当前设备中无法读出实际上与该文件被损毁无异。所以要时时通过电子文件的利用来检验该文件长期保存是否有效。电子文件的利用除了把它拷贝后供用户阅读外更重要的则是通过网络提供给用户存取。通过电子文件的利用可以使保存单位从保存费用和保存价值的角度出发而把那些失去保存意义的电子文件淘汰把有保存价值的电子文件继续保存下去。同时通过利用还可以及时发现保存工作和利用系统中存在的问题与不足从而促使保存单位不断地调整和改进工作方法与技术措施。这样保存与利用的相互促进既能使电子文件最大限度地利用又能使电子文件信息资源可靠、安全、有效地长期保

存下去。

4.2.2.5 构建电子文件长期保存标准规范体系

标准规范体系的构建是推动电子文件长期保存理论与实践 发展的重要动力之一。从长远考虑电子文件的长期保存和有效利用必须考虑标准化和 规范化问题。我国电子文件长期保存的标准制定工作相对滞后 ■ 虽然 早在十多年前就先后发布了《CAD 电子文件光盘存储、归档与档案管 理要求》(GBT17678121999)与《电子文件归档与管理规范》(GB / T188942002)两个国家标准但后续研究一直不成体系也没有形成广泛接受和遵循的国家或行业标准规范。因此现阶段应根据我国电子文件管理工作实际借鉴国外电子文件长期保存先进经验和技 术抓紧制定我国电子文件长期保存标准构建标准规范体系。

4.3 电子文件长期保存的元数据策略

电子文件形成、捕获、登记、分类、存储和保管、利用、跟踪、处置、传输、归档移交及长期保存等过程中都需记录在元数据中,并应保持连续、一致,以确保电子文件的真实性、完整性与有效性。为实现电子文件全过程管理与监控,电子文件及其元数据应结构稳定、定义良好的元数据标准相一致,特制定本标准。本标准的制定遵循国际文件管理的系列标准规范:按 ISO15489《信息与文件—文件管理》规定的文件管理过程要 求定义电子文件管理元数据,采用 ISO/TS 23081《信息与文献—文件管理流程—文件元数据》定义的电子文件管理元数据概念模型、实体及关系模型;采用 ISO14721《空间数据和传输系统—开放档案信息系统—参考模型》定义的信息包(SIP、AIP 及 DIP)数据模型,实现电子文件内容、背景、结构信息的封装,以确保在分散与集中管理过程中电子文件的真实性、完整性、有效性,确保电子文件在不同应用系统间的安全传输;采用 ISO14721 定义的数字信息长期保存模型,实现电子文件的长期保存与利用。本标准支持元数据值的再利用、元数据元素的互操作及可扩展性和模块化。元数据来源于机构文件处理方案、主题词表或关键词表、编码方案、电子政务系统、办公自动化系统、电子文件管理系统、档案管理系统、E-mail 系统、网站。

4.3.1 元数据概念模型

文件管理业务系统的各个流程 需要通过元数据实现对文件或档案的管理。元数据最主要的用途之一是用来描述业务系统中实体。关键实体如下:文件实体:文件本身,不管是单份文件还是文件集合体;责任者实体:业务环境中的人或组织结构;业务实体:业务办理。根据元数据类型,可将元数据分为下列几类:关于文件自身的元数据、关于责任者的元数据、关于业务工作或过程的元数据、关于业务规章制度、政策以及法规的元数据、关于文件管理过程的元数据。

子文件元数据模型的建立是以文件连续体理论为基础的。文件实体、责任者实体和业务活动实体,是用元数据帮助确保文件真实性、完整性和有效性,实施档案化管理必须充分关注的要素。文件作为交流、传递、贮存、利用信息的工具,其生成、处理、运转必然与文件责任者处理某项事务相关。对该事务的办理,形成文件的业务活动,构成了文件的来源...。这种业务活动构成了文件的背景。 责任者作为文件的形成和处理者,为业务活动或事务处理活动的实施者。作为业务活动的主体,责任者与业务活动的推进息息相关,人们的业务活动自身也是划分为一定层次并且与相关文件群体(文件集合)的层次相对应。从微观到宏观区分为以下层次:

- 1) 事务处理:即处理每一件具体事务的业务活动,在这种业务活动中形成的文件一般情况下可以构成一个文件组合(案卷)。
- 2) 业务活动:即处理某种类型事务而进行的职能活动,有可能是由某个机构的一个职能部门进行;也可能该职能活动属于某个相对独立的业务活动项目,是其中的一个子项目。在这种职能活动中形成的文件,往往构成全宗内的一个类别(系列)。
- 3) 业务职能:即具有相对独立性的业务活动过程。

其中以主体为核心的,是由一个立档单位进行的全部业务活动,在该项业务活动中形成的全部文件往往构成一个主体全宗;以客体为核心的相对独立而又具有相当规模的业务活动,则是围绕着该客体而跨立档单位进行的,所形成的文件一般情况下可以构成一个客体全宗。人们进行的业务活动,同时也决定了文件所拥有的和将要拥有的各种有机联系。除了在处理同一项业务活动过程中形成的所有文件相互之间的联系以外,还有:业务活动与文件的联系;业务活动承担者(即文件形成者和处理者)与文件的联系。它们是与文件的形成产生过程同步共生的。因此,(由业务活动承担者进行的)“业务活动”所诱发形成的“与文件有关的各种有机联系”是文件来源信息的重要组成部分,因而也是文件背景信息的重要组成部分。

4.4 电子文件长期保存的格式问题及其对策研究

电子文件的长期保存与利用需要法律支撑,没有法律支持的电子文件就没有保存意义。我国档案行政主管部门正会同政府信息产业部门,联合制定电子文件管理办法,确认电子文件的法律地位,规定用统一的软硬件设施来管理电子文件;同时,加强归档管理,增设相应的机构,行使监督、检查、指导的职权。

4.4.1 使电子文件归档时改变结构,脱离原来的软、硬件环境

针对电子文件对设备依赖的特性,一些国家采取相应措施,使电子文件脱离原来的软、硬件环境,达到长久保存的要求。如挪威承担了北欧国家共同的“电子文件管理工具(Team)”项目的开发。该工具软件包将 ORACLE 数据库文件以展开方式存储,去掉了元数据库复杂的索引结构,使电子文件的保管不再受元数据库的条件限制。在日后需要使用这些电子文件时,可通过 SQL 结构化查询语言检索。虽然效率较低,但通过 SQL 仍可重建原始数据库的逻辑结构。美国国家档案与文件署与国家科学基金会合作开发了 ISO 标准,并已在 2007 年试行,整个系统预计到 2011 年 4 月正式完成。该系统从统一性和综合性层面支持电子文件管理,而忽略对实现这个系统的信息技术构架的定位。它的功能(服务)在于:无论身处何地,公众都能通过因特网利用政府电子文件,并能从国家档案馆、总统图书馆和联邦文件中心发送、检索电子文件。该系统在支持档案馆的电子文件管理过程中,支持所有电子文件原创者和档案馆保存电子文件的集纳方法,且支持范围延伸到所有电子文件管理的基本过程。与系统支持电子文件管理的方法和如何定位其他类型的电子文件不同,这个系统还支持对利用电子文件所产生的信息、信息记录者以及形成电子文件的全过程管理。总部设在加拿大的大不列颠哥伦比亚大学的多学科合作项目——“永久保护真实电子文件国际研究”(INTERPARES),其重点研究方向就是保存真实电子文件的必要条件。我国目前还未发现采用上述方法的尝试,原因是我国信息化建设起步较晚,信息技术相对落后,加之幅员辽阔,情况复杂,用一个软件包对多种数据库的电子文件进行展开处理非常困难。另外,被展开存储的电子文件,能否恢复其原有的模式尚存在争议。

4.4.2 对电子文件法律地位问题的解决。

档案的基本价值是凭证价值和参考价值,尤以凭证价值最为重要。虽然,目前人们对数字化档案具有参考价值这一点已无疑义,但是,由于一系列技术和立法问题还未解决,数字化档案目前还不具备凭证效力。在电子文件的法律效力问题还没有很明确的解决方法之前,面对大量产生的电子文件,不少国家已对此作出规定,即以严格的管理程序来保证电子文件的凭证作用。如加拿大等国家规定,符合国家规定的管理程序产生的电子文件具有凭证作用。不过在实际操作中,并不是所有的法庭都认可电子文件的凭证作用。目前我国对电子文件的凭证作用也没有很明确的规定,只是把这个问题包含在计算机信息系统的整体法律认定上。例如,具有法人资格部门的计算机系统产生的电子文件记录对该部门的活动具有凭证作用。很多国家当前也是采用了这种对策。看来,解决这个问题除了行政办法外,还要依靠更加严格的技术措施才行。

4.4.3 加强对电子文件管理知识的培训,提高档案工作人员的业务素质。

面对大量出现的电子文件，档案人员光懂纸质档案管理知识是远远不够的，必须学习电子文件管理知识，才能跟上时代步伐。可从三个方面进行培训：

- (1) 分期分批对档案人员进行电子文件管理知识的培训；
- (2) 采取“走出去请进来”的方法，组织档案人员学习外单位在电子文件管理方面取得的经验；
- (3) 召开现场会，及时总结经验和教训，拓宽业务见识，增强业务素质。

4.4.4 加快档案信息资源建设进程，提高电子文件利用的质量

现阶段，在做好电子文件归档管理工作的同时，要继续采用通用标准和先进技术，做好传统载体档案的数字化工作，整合安全完备的信息资源。针对传统用户与网络用户并存的现实，在构建档案信息资源体系时，找到实体资源建设和数字化资源建设之间的“虚”“实”平衡点，在基于用户需求的网络服务平台上，满足多样化用户的利用需求。此外，提高用户信息素质，也是档案信息资源服务部门的职责。在构建复合服务模式时，应培育和引导用户提高信息利用素养：(1) 培养信息意识。通过日常宣传和网站推广等措施对用户进行档案信息意识教育，提升全社会档案意识，增强用户利用档案信息资源的自觉性。(2) 培养信息获取能力。针对档案信息资源的特点和利用要求，提升用户利用档案信息的能力，让用户既可在传统利用服务模式下查阅所需信息，又能在网络环境下检索所需的资源。(3) 培养信息道德。通过对用户进行信息道德教育，提高用户信息道德水准，减少档案信息利用风险。

4.4.5 构建知识创新与服务的大平台——知识网络。

知识网络是伴随着互联网的迅速发展而创建的服务平台，它使整个互联网变成一台巨大的超级计算机，像电力网供电一样，能将相兼容的任何设备，在任何时间、任何地点插入网络，实现计算、存储、数据、信息、知识、专家等资源的全面共享。它的本质特征不是网络规模，而是协同工作、资源共享，消除资源孤岛。在知识网络中，灵活的、可扩展的信息服务体系结构，能够保证资源信息提供者和信息服务者的广泛分布性，避免由于单个信息服务实体的失败导致其他资源信息服务不能正常提供服务。知识网络中的所有元数据构成元数据目录，这种目录结构系统具有层次和分布式的特性，不但不会破坏现有系统的元数据描述方法，还能与它们很好地交互、融合。总之，电子文件的长期保管和开发利用，是档案界不可回避的使命，机遇和挑战并存，档案工作者应以积极的态度，努力提高自身素质，更新知识储备，完善电子文件长期保存和利用的管理方案与技术措施，以饱满的热情和充足的信心，迎接知识网络时代的到来！



中國人民大學
RENMIN UNIVERSITY OF CHINA

2013級檔案班檔案信息化課程