ICS 01.140.20

A 14

备案号

C A D A L 项 目 标 准

CADAL 20903—2012

数字图书馆安全标准规范集
第三部分：数字资源长期保存规范

Digital Library Security Standards

Part 3: Digital Resources Long-term Preservation Specification

第一稿

2012-05-01

请将您知道的与本标准相关的专利连同支持性文件一并发给 CADAL 项目管理中心

2012－05－08 发布　　2012－05－09 实施

CADAL 项目管理中心　　发 布

目　次

前　言

本标准包括以下四个方面的内容：

——第1部分：数字对象存储安全规范

——第2部分：　访问控制规范

——第3部分：　数字资源长期保存规范

——第4部分：　安全传输标准

本部分是《数字图书馆安全标准规范集》的第三部分。

本部分的制定依据了《GB/T 1.1-2009标准化工作导则第1部分：标准的结构和编写》的要求。

本部分是由大学数字图书馆国际合作计划（CADAL）项目管理中心提出并归口

本部分起草单位：数字图书馆教育部工程研究中心

本部分起草人：张鹏、张寅

引　言

数字资源长期保存是指为保证数字资源能够被长期存储和访问到而进行的管理活动。

数字图书馆数字资源的长期保存应该具有相应的安全规范。

本规范是在 CADAL 数字图书馆对资源长期保存的实践基础上编制的。

本标准清晰说明了符合 CADAL 项目要求的数字资源长期保存的方式。

数字图书馆安全标准规范集

第3部分：数字资源长期保存规范

1 范围

本标准规定了数字资源长期保存规范。本标准适用于CADAL项目中对数字资源的长期保存管理。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

《ISO 32000》

《GB/T 23286.1-2009》

3 术语定义

3.1

开放电子书 Open eBook **缩写：**OEB

用于格式化和包装电子书的一种行业标准。OEB基于XML，定义了电子出版物的文本如何被标记，以及一本ebook的各部分(封面、目录、正文、说明、索引等等)应如何包裹在一起。

3.2

电子发行本 Electronic Publication **缩写：**Epub

一种电子图书标准，由国际数位出版论坛（IDPF）提出；其中包括3种文件格式标准，文件的附档名为.epub，这个格式已取代了早期的的Open eBook开放电子书标准。Epub格式的内部结构域OEB基本一致，可以方便的相互转换。

3.3

便携文件格式 Portable Document Format **缩写：**PDF

一种电子文件格式，这种文件格式与操作系统平台无关，主要利于打印和跨平台使用。

3.4

都柏林核心元数据 Dublin Core **缩写：**DC

数字图书馆中使用的一组简单的包括 15 个“核心元素”的元数据元素集合，主要用于描述数字对象、馆藏管理和元数据交换。

3.5

标签图像文件格式 Tagged Image File Format **缩写：**TIFF

一种主要用来存储包括照片和艺术图在内的图像的文件格式。它最初由 Aldus公司与微软公司一起为PostScript打印开发。

TIFF文件格式适用于在应用程序之间和计算机平台之间的交换文件，它的出现使得图像数据交换变得简单。

3.6 DjVu 图像文件

主要用于存储扫描的文档。这种格式的特色包括图像分层、渐进载入、算术编码、对二进制图像进行有损压缩，从而以较小的空间，存放高质量的可读图像。渐进载入使得DjVu适合于应用于因特网。DjVu对于大部分的扫描文档，表现都优于PDF，故被作为PDF的替代品来进行推广。这种格式已经在文件共享网络中，被广泛使用于分发数学书籍。

4 数字资源长期保存流程规范

计算机技术的发展日新月异，当前流行的标准、格式可能很快就会升级、废弃、改动。此外，软件技术的发展速度也使得不可能总是使用当前的软件来访问数字资源。因此我们需要对数字资源做好长期保存工作。

首先，对扫描文件需要进行保存。将纸质文本扫描后以图像格式存储，存储格式要具有良好的结构，能够尽可能多的与各种通用图像格式相互转换，并保持高度的清晰度。因此，本标准建议使用tiff图像文件格式。

其次，对于数字资源的组织结构需要保存。扫描后得到的图像格式仅能反应某一页的内容，却不能反映出整本书的结构和组织方式。因此，也需要对整本书的组织结构进行长期保存，以保证多年后仍然能解析组织结构并完整的展示整本书。标准推荐使用OEB结构来保存数字资源。

最后，访问、解析数字资源需要使用额外的软件或者系统，没有它们，即使存有数字资源本身，也无法有效利用数字资源。因此，需要长期保存相应的软件、硬件系统，包括硬件结构（或对应虚拟机）、操作系统、各类软件、数据库、文件系统。

所有应长期保存的资源，都应该单独存放于外部存储设备，且任意时刻都有2份以上的副本，并定期检查内容的完整性，防止存储设备损坏、老化造成的数据丢失。同时，也可以考虑使用稳定可靠的云存储服务辅助存储。

5　OEB 和 EPUB 规范

表1　OEB 格式和 EPUB 格式的 CADAL 图书规范

01010243/ meta/ a.opf Catalog.xml dc.xml ptiff/ 00000001.djvu 00000002.djvu 00000003.djvu	01010243/ mimetype META-INF/ container.xml signatures.xml OEBPS/ content.opf toc.ncx style.css ptiff/ 00000001.xhtml 00000002.xhtml .....

注：（左）OEB格式（右）EPUB格式

数字资源应遵循OEB开放图书标准。EPUB是一个自由的开放标准，属于一种可以“自动重新编排”的内容,也就是文字内容可以根据阅读设备的特性，以最适于阅读的方式显示。EPUB是OEB的后续标准，可以实现程序自动转换，并加入各种辅助信息，最后打包成ZIP压缩格式的单一文件。表1示例了OEB格式和EPUB格式的CADAL图书,CADAL图书主要由DjVu图像组成,在进行OCR识别和人工校对后,转变成xhtml格式的文本。

一个 EPUB 就是一个简单 ZIP 格式文件。EPUB 文件可使用标准 XML 工具创建，不需要任何专门或者私有的软件。大多数 EPUB XML 模式都来自现成的、可免费获得的、已发布的规范。EPUB 元数据是 XML，EPUB 内容是 XHTML。如果文档构建系统产生的结果用于 Web 和/或基于 XML，那么也可用于生成 EPUB。

6 发布图书格式—DjVu

数字资源使用DjVu作为主力图像发布文档格式，Djvu由AT&T的一个研究小组开发完成，它使得高质量的扫描图像可以轻易地在因特网上进行发布。DjVu格式背后的一项主要技术是将图像分为背景层

（Background Layer，包括纸的纹理和图片）和前景层（Foreground Layer，包括文本和线条），如图1所示，并且有针对性地使用JB2和IW44这两种压缩算法。

一般来说，要确保文字和线条的清晰度需要较高的分辨率（通常为300dpi），而反映连续色彩图像和纸张的背景机理则不需要那么高的分辨率（通常为100dpi）。因此，要提高清晰度，最好的方法就是将这些元素分为不同的层来进行处理。通过将文字和背景分离开来，DjVu可以用高分辨率来还原文字，使锐利边缘得以保留，并最大限度地提高可辨性，同时用较低的分辨率来压缩背景图片，从而使整个图像的质量得到了保证。使用DjVu格式，用户首先会很快得到页面的一个最初版本，这个版本主要是含有文字的前景层。随着后续信息的到达，图像质量不断提高。例如，一张普通杂志页面上的文字在56Kbps调制解调器的连接下只须3秒钟就可出现。在其后的1-2秒内，背景图片的初级版本也将出现。然后，再过几秒钟，最后的完整页面就可全部出现了。

DjVu作为一种新的彩色文件压缩技术，在纸质世界和比特世界之间搭起了一座桥梁。它使得高质量的扫描图像可以轻易地在因特网上进行发布。专门针对网络发行而设计的DjVu技术，以其友好的用户界面和网络功能博得了越来越多的商业和非商业用户的垂青。DjVu以其科学有效的压缩模式，使图片传播的硬件和带宽瓶颈得以突破。

7　长期保存图像格式—TIFF

数字资源使用TIFF作为长期保存的图像格式，标签图像文件格式（Tagged Image File Format，简写为TIFF）是一种主要用来存储包括照片和艺术图在内的图像的文件格式。它最初由Aldus公司与微软公司一起为PostScript打印开发。TIFF与JPEG和PNG一起成为流行的高位彩色图像格式。TIFF格式在业界得到了广泛的支持。

TIFF 是一个灵活适应性强的文件格式，通过在文件头中包含“标签”它能够在一个文件中处理多幅图像和数据。标签能够标明图像的如图像大小这样的基本几何尺寸或者定义图像数据是如何排列的并且是否使用了各种各样的图像压缩选项。例如，TIFF可以包含JPEG和行程长度编码压缩的图像。TIFF文件也可以包含基于矢量的裁剪区域（剪切或者构成主体图像的轮廓）。使用无损格式存储图像的能力使TIFF文件成为图像存档的有效方法。与JPEG不同，TIFF文件可以编辑然后重新存储而不会有压缩损失。其它的一些TIFF文件选项包括多层或者多页。

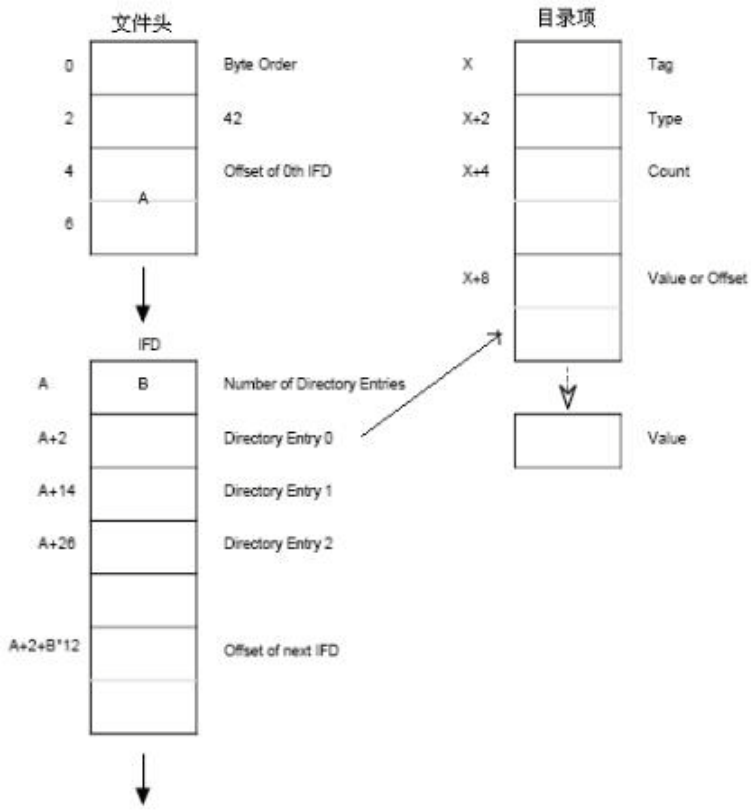

图1　TIFF 数据结构图

TIFF一般来说由四部分组成：文件头、文件目录、目录内容、图像数据。在每一个TIFF文件中第一个数据结构称为图像文件头或IFH，它是图像文件体系结构的最高层。这个结构在一个TIFF文件中是惟一的，有固定的位置。它位于文件的开始部分，包含了正确解释TIFF文件的其他部分所需的必要信息。IFD是TIFF文件中第二个数据结构，它的每个项指向一个目录内容。该结构前两个字节说明其项的总数，之后每个项都是12字节。第三个数据结构是目录项，它是一个名为TAG的用于区分一个或多个可变长度数据块的表，标记中包含了有关于图像的所有信息。这种方法允许数据字段定位在文件的任何地方，且可以是任意长度，因此文件格式十分灵活。最后一个数据结构存储图像数据，目录项中的定义了它的其实地址。

尽管现今它是一种被广泛接受的标准格式，当TIFF最初出现的时候，它的可扩展性带来了很多兼容问题。程序员可以随意定义新的标签和选项，但是并不是所有的实现程序都能支持这些所有这些创造出的标签。作为结果，它的一个最小特性集成为了“这个”TIFF，即使是在今天大量的TIFF文件和读取它们的代码都是基于简单的32位非压缩图像。

8 常用图书格式—PDF

PDF文件格式与操作系统平台无关，也就是说，PDF文件不管是在Windows，Unix还是在苹果公司的Mac OS操作系统中都是通用的。这一性能使它成为在Internet上进行电子文档发行和数字化信息传播的理想文档格式。PDF文件格式可以将文字、字型、格式、颜色及独立于设备和分辨率的图形图像等封装在一个文件中。该格式文件还可以包含超文本链接、声音和动态影像等电子信息，支持特长文件，集成度和安全可靠性都较高。

PDF文件使用了工业标准的压缩算法，通常比PostScript文件小，易于传输与储存。它还是页独立的，一个PDF文件包含一个或多个“页”，可以单独处理各页，特别适合多处理器系统的工作。此外，一个PDF文件还包含文件中所使用的PDF格式版本，以及文件中一些重要结构的定位信息。

图2 PDF 文件结构

PDF文件结构主要可以分为四个部分：文件头、文件体、交叉引用表、文件尾。文件头指明了该文件所遵从的 PDF 规范的版本号，它出现在 PDF 文件的第一行。文件体是PDF 文件的主要部分，由一系列对象组成。 交叉引用表是为了能对间接对象进行随机存取而设立的一个间接对象的地址索引表。文件尾声明了交叉引用表的地址，即指明了文件体的根对象，从而能够找到 PDF 文件中各个对象体的位置，达到随机访问。另外 还保存了 PDF 文件的加密等安全信息。

9 保存运行工具

除了保存数字资源本身外，对于查看、编辑、访问数字资源的软件工具也应该保存。

计算机技术的发展日新月异，当前流行的标准、格式可能很快就会升级、废弃、改动。此外，软件技术的发展速度也使得不可能总是使用当前的软件来访问数字资源。因此，需要对运行工具也进行保存。

需要保存的运行工具包括：

- 访问数字资源的软件或系统
- 存放数字资源的文件系统及数据库
- 解析各类文件格式的软件及格式互换软件
- 操作系统或对应虚拟机。

这些工具应该单独存放于独立可靠的外部存储设备，并提供各工具使用方式的文档。