ICS 01.140.20
A 14

C A D A L 项 目 标 准

CADAL 20601—2012

数字资源发现规范

Digital Resources Discovery Standards

第一稿

2012-05-08 发布　　2012-05-09 实施

CADAL 项目管理中心　　发　布

目　　次

前　言

《数字资源发布标准规范集》包括以下 3 个方面的内容：

——第 1 部分：数字资源发现规范；

——第 2 部分：数字资源访问规范；

——第 3 部分：数字资源传输规范。

本标准是其中的第 1 部分。

本部分的制定依据了 GB/T 1.1—2009《中华人民共和国国家标准》的要求，同时参照了国际数字出版论坛（International Digital Publishing Forum，IDPF）2007 年正式发布的 EPUB 标准。

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 CADAL 项目门户组。

本部分起草人：尹彦飞、张寅、边科。

引　　言

本标准充分分析了 CADAL 项目中数字资源发布的流程和规范，全面比较了现有的数字资源发布过程，规范了数字资源发现的概念与实现方式。

数字资源的发现是在数字资源平台上添加新的数字资源的过程，主要关注对数字资源进行元数据提取的过程，至于数字资源的访问与传输，在其他规范中都做了相应规定。

数字资源发现规范

1 范围

本标准确立了 CADAL 项目数字资源发布规范。
本标准规范的数字资源对象为图书数字资源。
本标准特种资源指印刷型文献的数字化衍生物。
本标准不适用不符合规范的数字资源的发布。
该规范适用于 DjVu、pdf 和 tiff 格式的图书数字资源。

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—1：2008 *Document management—Portable document format*
ISO 15836：2009 *Information and documentation—The Dublin Core metadata element set*
The EPUB 3.0 specification

3 术语和定义

3.1 数字资源发现 Digital Resource Discovery 缩写：DRD

用户从原始的数字资源中获取到数据的元数据，并且将这些元数据加入系统中对外提供服务的过程。

3.2 开放封装格式 Open Packaging Format 缩写：OPF

将已经编码压缩好的视频轨和音频轨按照一定的格式放到一个文件中，也就是说仅仅是一个外壳，或者把它当成一个放视频轨和音频轨的文件夹。

3.3 元数据 Metadata 缩写：MD

用于描述要素、数据集或数据集系列的内容、覆盖范围、质量、管理方式、数据的所有者、数据的提供方式等有关的信息。

3.4 文件格式 File Type 缩写：FT

文件格式是指电脑为了存储信息而使用的对信息的特殊编码方式，是用于识别内部储存的资料。比如有的储存图片，有的储存程序，有的储存文字信息。每一类信息，都可以以一种或多种文件格式保存在电脑存储中。每一种文件格式通常会有一种或多种扩展名可以用来识别，但也可能没有扩展名。扩展名可以帮助应用程序识别的文件格式。

3.5 CSV 文件 Comma Separated Values 缩写：CSV

CSV 文件，即字符分隔值文件。CSV 是分隔的数据格式，有字段/列分隔的逗号字符和记录/行分隔换行符。字段包含特殊字符(逗号、换行符或双引号)，必须用双引号括住。行内包含一个项目是空字符串，可以用双引号括住。字段的值包含双引号时，要双写这个双引号(就像把一个双引号当作转义符一样)。CSV 文件格式并不需要特定的字符编码、字节顺序，或行终止格式。

4 数字资源发现规范

4.1 图书数字资源

4.1.1 图书发现规范

图书发现是指从磁盘等存储介质上存储的原始图书数字资源中，获取到图书相关元数据信息的过程。在这个过程中，需要扫描与分析磁盘等存储介质，确定各个图书所在的存储路径，并且获取到其中的图书元数据，用于图书数字资源的发布。

图书发现有以下两种情况：

第一种情况是图书数字资源的管理人员明确新的图书数字资源的存在，并对它们的存储信息有充分的了解。在这种情况下。用户图书发现的工具将会根据管理人员所告知的存储路径对指定路径的图书资源进行扫描。

第二种情况是图书数字资源管理人员并不知道有新图书数字资源的添加，也不知道具体的存储目录的情况。在这种情况下，需要通过对全部的图书数字资源目录进行轮询的方式对所有的图书数字资源进行扫描，并对其中新添加的数字资源进行分析，以获取其中的元数据信息。

4.1.2 图书发现的方式

图书发现根据以上两种情况可分为按需准确发现和定时全局发现两种类型。两种类型的图书数字资源的发现方式均为需要提供的发现方式，具体如下所示。

4.1.2.1 按需准确发现方式

按需准确发现方式对应于第一种情况，此时图书数字资源的管理人员对新添加的图书数字资源有充分的了解，能够制定资源的路径，这种方式效率较高、发现的速度快，但是需要数字资源的管理者参与。

数字资源的管理者使用图书目录的扫描工具并对所要扫描的目录做具体的规范，即可实现采用该种方式发现图书数字资源。

4.1.2.2 定时全局发现方式

当图书数字资源的管理人员对数字资源缺乏了解的时候，使用定时全局发现的方式更为便捷。该种方式使用定时的机制，通过查看所有的图书目录来判断哪些是新增加的图书，并对这些图书进行元数据的分析和获取。

4.1.3 元数据导入系统规范

通过图书数字资源的发现过程所得到的数字资源的元数据需要满足《资源检索协议规范集》中有关“元数据存储”的规范，具体的规范内容请参见对应的章节。

对符合元数据标准的元数据，能够保存为 CSV、XML 格式的中间文档，然后使用元数据存储仓库中的数据导入工具来导入这些中间文档。

4.1.4 支持的格式

4.1.4.1 PDF

便携式文件格式(portable document format，PDF)是由 Adobe Systems 在 1993 年由文件交换发展出的文件格式。

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。它的优点在于跨平台、能保留文件原有格式(layout)、开放标准，能自由授权(royalty-free)、自由开发 PDF 相容软件。

对普通读者而言，用 PDF 制作的电子书具有纸版书的质感和阅读效果，可以“逼真地”展现原书的原貌，而显示大小可任意调节，给读者提供了个性化的阅读方式。由于 PDF 文件可以不依赖操作系统的语言和字体及显示设备，阅读起来很方便。这些优点使读者能很快适应电子阅读与网上阅读，无疑有利于计算机与网络在日常生活中的普及。

PDF 是一个开放标准，2007 年 12 月成为 ISO 32000 国际标准。2009 年 9 月 1 日，作为电子文档长期保存格式的 PDF/Archive(PDF/A)经中国国家标准化管理委员会批准已成为正式的中国国家标准。

4.1.4.2 OEB 与 EPUB

OEB 是电子书阅读器、PDA 等载体上的文书格式，已经被新修订的 EPUB 格式取代。出于兼容考虑，CADAL 数字资源仍然支持 OEB 格式的发布。

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 档案内部使用了 XHTML 或 DTBook (一种由 DAISY Consortium 提出的 XML 标准)来展现文字，并以 zip 压缩格式来包裹档案内容。EPUB 格式中包含了数字版权管理(digital rights management，DRM)相关功能可供选用。

EPUB 于 2007 年 9 月成为国际数位出版论坛(International Digital Publishing Forum，IDPF)的正式标准，以取代旧的开放 Open eBook 电子书标准。

EPUB 解决了 PDF 和开发人员友好性有关的所有瑕疵。一个 EPUB 就是一个简单的 ZIP 格式文件(使用.epub 扩展名)，其中包括按照预先定义的方式排列的文件。除此以外，EPUB 非常简单：

(1) EPUB 中的所有内容基本上都是 XML。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。

(2) EPUB 内容(eBook 的具体内容)基本上都是 XHTML 1.1。

(3) 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。

最关键的点在于 EPUB 元数据是 XML，EPUB 内容是 XHTML。如果您的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。目前，以 Google、Apple 为代表，众多公司都以 EPUB 作为数字图书的格式。

4.2 音频数字资源

音频数字资源的发现方式与前述图书数字资源的方式相同。具体支持的音频格式如下：

4.2.1 MP3

动态影像专家压缩标准音频层面 3(Moving Picture Experts Group Audio Layer Ⅲ，MPEG-1 Audio Layer 3)，经常被称作 MP3，是当今较流行的一种数字音频编码和有损压缩格式，它被用来大幅度地降低音频数据量，而对于大多数用户的听觉感受来说，重放的音质与最初的不压缩音频相比没有明显的下降。它是在 1991 年，由位于德国埃尔朗根的研究组织 Fraunhofer-Gesellschaft 的一组工程师发明和标准化的。

MP3 的特点如下：

(1) MP3 是一个数据压缩格式。

(2) 它丢弃掉脉冲编码调制(pulse code modulation，PCM)音频数据中对人类听觉不重要的数据(类似于 JPEG 是一个有损图像压缩)，从而使得文件体积小得多。

(3) MP3 音频可以按照不同的位速进行压缩，由此提供了在数据大小和声音质量之间进行权衡的一个范围。MP3 格式使用了混合的转换机制将时域信号转换成频域信号。

(4) 32 波段多相积分滤波器(polyphase quadratus filter，PQF)。

(5) 36 或者 12 tap 改良离散余弦滤波器(modified discrete cosine transform，MDCT)；每个子波段大小可以在 0，…，1 和 2，…，31 之间独立选择。

(6) MP3 不仅有广泛的用户端软件支持，而且有很多的硬件支持，比如便携式媒体播放器(指 MP3 播放器)、DVD 和 CD 播放器。

4.2.2 WMA

WMA(Windows Media Audio)是微软公司开发的一种数字音频压缩格式。一些使用 Windows Media Audio 编码格式编码其所有内容的纯音频 ASF 文件，也使用 WMA 作为扩展名。

WMA 格式最初为微软公司私有，但是随着苹果公司的 iTunes 对它的支持，这个格式正在成为 MP3 格式的竞争对手。它兼容 MP3 的 ID3 元数据标签，同时支持额外的标签。另外，一般情况下相同音质的 WMA 和 MP3 音频，前者文件体积较小。

WMA 可以用于多种格式的编码文件中。应用程序可以使用 Windows Media Format SDK 进行 WMA 格式的编码和解码。一些常见的支持 WMA 的应用程序包括 Windows Media Player、Windows Media Encoder、RealPlayer、Winamp 等。其他一些平台，例如 Linux 和移动设备中的软硬件，也支持此格式。

WMA 7 之后的 WMA 支持证书加密，未经许可(未获得许可证书)，即使是非法拷贝到本地，也是无法收听的。同时，微软公司起初宣称同文件比 MP3 体积小一倍而音质不变，这一承诺也得到了兑现。另外，微软公司在 WMA 9 大幅改进了其引擎，实际上几乎可以在同文件同音质的情况下比 MP3 体积小 1/3 左右，因此非常适用于网络串流媒体及行动装置。

4.3 视频数字资源

视频数据资源的发现方式与前述图书数字资源的方式相同，具体支持的视频格式如下：

4.3.1 FLV

Flash Video(FLV)，是一种流行的网络视频格式。随着视频网站的丰富，该格式已经非常普及。

FLV 流媒体格式是随着 Flash MX 的推出发展而来的视频格式。它形成的文件极小、加载速度极快，使得网络观看视频文件成为可能。它的出现有效地解决了视频文件导入 Flash 后使导出的 SWF 文件体积庞大，不能在网络上有效使用等缺点。

一般 FLV 文件被包在 SWF PLAYER 的壳里，并且 FLV 可以很好地保护原始地址，不容易被下载到，从而起到保护版权的作用。但还是有些视频格式转换软件可将 FLV 转成一般的视频格式，如中国大陆的软件格式工厂。

4.3.2 MP4

MPEG-4(MP4)是一套用于音频、视频信息的压缩编码标准，由国际标准化组织(International Standard Organization，ISO)和国际电工委员会(International Electrotechnical Commission，IEC)下属的“动态影像专家组(Moving Picture Experts Group，MPEG)”制定，第一版于 1998 年 10 月通过，第二版于 1999 年 12 月通过。MPEG-4 格式的主要用于网上串流、光碟、语音传送(视讯电话)，以及电视广播。

MPEG-4 包含了 MPEG-1 及 MPEG-2 的绝大部分功能及其他格式的长处，并加入扩充对虚拟现实模型语言(Virtual Reality Modeling Language，VRML)的支援，面向对象的合成档案(包括音效、视讯及 VRML 物件)，以及数字版权管理(digital rights management，DRM)及其他互动功能。而 MPEG-4 比 MPEG-2 更先进的一个特点，就是不再使用宏区块作影像分析，而是以影像上个体为变化记录，因此尽管影像变化速度很快、码率不足，也不会出现方块画面。

由于 MPEG-4 是一个公开的平台，各公司、机构均可以根据 MPEG-4 标准开发不同的制式，所以市场上出现了很多基于 MPEG-4 技术的视频格式，例如 WMV 9、Quick Time、DivX、Xvid 等。MPEG-4 大部分功能都留待开发者决定是否采用。这意味着整个格式的功能不一定被某个程式所完全涵括。因此，这个格式由所谓配置(profile)及级别(level)定义了 MPEG-4 应用于不同平台时的功能集合。

4.3.3 WMV

WMV(Windows Media Video)是微软推出的一种流媒体格式，它是由 ASF(Advanced Stream Format)格式升级延伸而来。在同等视频质量的情况下，WMV 格式的文件可以边下载边播放，因此很适合在网上播放和传输。

WMV 的主要优点在于：可扩充的媒体类型、本地或网络回放、可伸缩的媒体类型、流的优先级化、多语言支持和良好的扩展性。

参 考 文 献

[1] 中华人民共和国国家质量监督检验检疫总局,中国国家标准化管理委员会. GB/T 1.1—2009 中华人民共和国国家标准[S]. 北京:中国标准出版社,2009.

[2] EPUB electronic publication 数位版权管理[N/OL]. http://idpf.org/epub/30. 2007.

[3] 侯瑞芳. OEB(开放式电子图书)格式标准与安全性研究[J/OL]. http://www.cnki.com.cn/Article/CJFDTotal-JSTS200303005.htm.