UOF
标文通

中文办公软件文档格式规范

（1.0、1.1版）

使用指南

李　宁　吴新松　等著

12

湖南师范大学出版社

UOF
标文通

中文办公软件文档格式规范

（1.0、1.1版）

使用指南

李　宁　吴新松　等著

湖南师范大学出版社

图书在版编目（CIP）数据

中文办公软件文档格式规范（1.0、1.1 版）使用指南 / 李宁，吴新松等著. —长沙：湖南师范大学出版社，2010.7
ISBN 978-7-5648-0259-2
Ⅰ. ①中… Ⅱ. ①李…②吴… Ⅲ. ①汉字信息处理系统—文件—规范—指南 Ⅳ. ①TP391. 12-65

中国版本图书馆 CIP 数据核字（2010）第 131969 号

中文办公软件文档格式规范（1.0、1.1 版）使用指南

李 宁 吴新松 等 著

◇责任编辑：何海龙 李 妮
◇责任校对：蒋旭东 胡晓军
◇出版发行：湖南师范大学出版社
地址/长沙市岳麓山 邮编/410081
电话/0731. 88853867 88872751 传真/0731. 88872636
网址/http：//press. hunnu. edu. cn
◇经销：湖南省新华书店
◇印刷：湖南航天长宇印务有限公司

◇开本：889×1194 1/16
◇印张：23
◇字数：690 千字
◇版次：2010 年 8 月第 1 版 2010 年 8 月第 1 次印刷
◇书号：ISBN 978-7-5648-0259-2
◇定价：70. 00 元

前　言

随着全球信息化进程的不断深入发展，信息资源已成为与材料和能源同等重要的战略资源，在经济社会资源结构中具有不可替代的地位。中文办公软件是人们日常办公不可缺少的基础软件，大量办公软件生成的电子文档信息已成为信息化社会中一类重要的信息资源。中文办公软件在我国拥有着广阔的市场。

然而，长久以来办公软件产品使用了大量不公开的格式，导致不同编辑软件产生的文档难以交换，文档过度依赖软件产品，带来大量兼容性和安全性问题。为实现办公文档的交换和互操作，我国于 2007 年颁布了推荐性国家标准 GB/T 20916—2007《中文办公软件文档格式规范》（简称“标文通”）。该标准定义了文字处理文档、电子表格和演示文档三种主要文档格式的描述体系。

标文通以标准体例叙述的内容可能难以为一般读者所理解。为配合标准的宣贯实施，推广标文通的应用，我们编写了本使用指南。本指南得到了核心电子器件、高端通用芯片及基础软件产品科技重大专项“办公软件文档格式标准研制与测试”(2008ZX01044-001)、北京市教育委员会科技发展重点项目暨北京市自然科学基金“智能文档关键技术研究”（KZ200810772017）的基金支持。

本指南以最常用的中文办公软件功能需求为出发点，对每个功能点如何采用标文通描述进行了详细说明。本指南还配有大量的例子，深入浅出地介绍了标文通的使用方法。

本指南第一部分主要为文字处理文档格式相关内容，第二部分为电子表格文档格式相关内容，第三部分为演示文稿文档格式相关内容。

指南的第一部分主要包括：

- 关于标文通格式：介绍了标文通的制订背景，从一个最简单的例子认识标文通文字处理文档。
- 标文通的公共内容：涉及元数据、式样集、书签集、链接集、对象集、用户数据集、扩展区、数字签名，以及关于度量单位、锚点表示、线型表示的约定。
- 文字处理：从文档设置、章节设置、段落设置、字体设置、文字表格、式样、域、脚注/尾注、书签、批注、目录索引、项目符号和编号、修订、超链接、文本框等方面介绍了如何用标文通表示文字处理相关的信息。这部分内容有很多在电子表格处理和演示文稿处理中也会用到。
- 图形表示：包括图形的组合、填充、箭头表示、图形控制、路径表示等内容。
- 多语种置标支持方案：介绍了如何在标文通中支持置标的中外文转换。

指南的第二部分主要包括：

- 页面设置：包括工作表属性中与页面相关的内容，以及可以在页面中出现的页眉页脚、图表，以及工作表的高层属性。这一节的目的是通过读者习惯的页的概念来了解电子表格，使之容易接受。
- 工作表设置：介绍了电子表格的主体内容，包括工作表的属性、内容、筛选设置、分页控制，并按照表、列、行、单元格的概念逐级介绍各部分内容。另外，对于单元格中经常使

用的公式与批注也进行了讲解。

- 单元格式样：这部分介绍了与格式相关的数字格式约定、边框和背景的描述方法。
- 公用处理规则：涉及 Schema 中相应部分的常用设置、处理规则或集中放置的内容，包括度量单位、精确度、日期系统、对迭代计算的规定、关于数据有效性检验的规定、条件格式、区域公式以及单元格引用等等。
- 其他内容：包括命名表达式和超链接的描述以及其他对象引用的方法。

指南的第三部分主要包括：

- 全局设置：涉及 Schema“公共处理规则”中相应部分的常用设置、处理规则或集中放置的内容，包括度量单位、页面设置、页面版式、文本式样、最后视图、显示比例、放映设置、页眉页脚等。
- 母版：介绍了母版的种类和描述方法。
- 幻灯片：介绍了幻灯片描述方法，重点包括：背景设置、以框对象形式出现在幻灯片上的内容、动画和幻灯片切换。
- 动画：介绍了演示文稿常用动画效果的描述方法，重点为时序和特效。
- 使用对象：介绍了幻灯片中使用域、超链接和多媒体对象的方法。

指南的附录中包括：域代码定义、数字格式码描述、数字格式枚举、图案纹理填充枚举、自选图形种类枚举、数字格式码描述、公式和函数等内容，便于与正文内容相互参照。

本指南所述内容对应《中文办公软件文档格式规范》Schema 的版本为 1.1 版本。本指南将根据 Schema 后续版本的修订情况适时推出后续版本。

本指南的预期读者为标文通的开发人员和有办公软件使用经验的用户。读者需事先对办公软件和文档编辑排版有基本的了解。

本指南的读者可以从地址 http://www.uof.org.cn/ 找到与本指南相关的 Schema 定义、支持软件、勘误表和其他相关资源。

参加本指南编写出版工作的有：李宁、吴新松、方春燕、梁琦、董慧、于进福、张云锋、文佳斐、滑淑然、吴倩、黎美秀、夏艳霞、樊凯、罗文甜、李东明、宋昊苏、陈金红、杨映玉等。吴志刚、赵菁华、王长胜、黄芳、成修治、李建萍、唐毕洪等曾提供大量帮助，在此表示衷心感谢。

中文办公软件基础标准工作组为本指南的编写做了大量的前期工作，特别得到了中国电子技术标准化研究所、中国科学院软件研究所、上海中标软件有限公司、无锡永中科技有限公司、北京红旗中文贰仟软件技术有限公司、金山软件股份公司、北京九州汇宝软件有限公司等单位的大力支持和协助。本指南中部分样例采用以上公司的产品生成，一并在此鸣谢。

由于编写时间仓促，并限于作者的水平，本指南必定会存在一些不足之处，欢迎读者给予批评指正。作者联系方式：ningli@public2.bta.net.cn。

2010 年 2 月 1 日

目　录

第二部分：电子表格文档格式

第三部分：演示文稿文档格式

附录及参考文献

第一部分：文字处理文档格式

第一部分　文字发展之历史

第一章　关于标文通格式

1.1　中文办公软件文档格式规范的制订背景

办公软件是一类特殊的软件，它的一个重要特点是面向日常办公应用，例如，文字处理、电子表格处理和演示文稿处理等，近年也深入到电子政务和电子商务等各个领域。由于应用面广，必然要求一个办公软件产生的文档能够被其他办公软件或应用软件理解和使用，文档在不同办公软件之间可以无障碍地交换。

从 1979 年出现第一个通用字处理软件和电子表格软件，及 1987 年第一个整体发售的办公套件至今已二十余年。在这期间使用的大部分办公软件文档格式是以二进制形式存储的、办公软件厂商所特有的格式，一般不对外公开。这样就使得不同的办公软件之间难以交换文档，例如 Microsoft Office 自 97 版至 2003 版未公布其格式细节。随着办公软件复杂度的提高和其他人为因素，其他办公软件对封闭文档格式的完全兼容变得越来越困难。

常见的主要办公文档格式及相关办公软件见表 1-1。

表 1-1 国际和国内部分不同办公软件的文件格式

公司/社区	办公软件	文字处理	电子表格处理	演示文档处理
Microsoft	Microsoft Office	Word (.doc)	Excel (.xls)	PowerPoint (.ppt)
OpenOffice.org 红旗中文 2000 中标普华	OpenOffice.org Red Office 中标普华 Office	Writer (.sxw/.odf)	Calc (.sxc/.odf)	Impress (.sxi/.odf)
Corel	WordPerfect Office	WordPerfect(.wpd)	Quattro Pro(.qpw)	Presentations(.shw)
IBM	Lotus SmartSuite	Word Pro(.lwp)	Lotus 1-2-3(.123)	Freelance (.prz)
金山	WPS Office	金山文字(.wps)	金山表格(.et)	金山演示(.dps)
永中	永中 Office	文字处理(.eio)	电子表格(.eio)	简报制作(.eio)

办公软件采用私有格式带来的问题是：

- 格式信息不完整。封闭的办公软件文档格式使得无法通过合法途径获得文档格式的完整、准确的描述信息。
- 文档依赖软件厂商的产品。由于文档格式封闭，只有办公软件的开发商具有访问文档结构数据的完全能力，并享有相关的软件著作权甚至专利权。即使某些办公软件文档格式是公开的，但是相关的许可限制可能会导致该文档格式实际上无法广泛使用。
- 扩展性差。由于大多数封闭的办公软件文档格式都是采用二进制方式存储数据，所以在遇到软件升级时通常都不得不重新设计文件格式，以至于一个软件的不同版本间所支持的缺省文件格式相差很大。如此一来，不仅不同办公软件间的兼容和互操作更加困难，甚至同一办公软件中反向兼容早期版本的格式都很难实现。
- 互操作性差。封闭的办公软件文档格式会导致不同办公软件间难以互操作，进而阻碍不同软件间的良性竞争和技术创新，使办公软件厂商固守现有的用户和市场，失去改进产品的动力。
- 安全性差。封闭的文档格式的一个巨大隐患是文档无法稳妥地长期保存。因为封闭的文档格式只有用特定的专有软件才能完全读取其内部信息，在长期保存的过程中无法确定未来仍然可以获得该特定的专有软件来访问这些文档。使用封闭的办公软件文档还可能存在严重的安全性和隐私权问题。例如，办公软件记录的某些信息可能会隐藏在文档内部，如在特殊环境下这些隐藏信息被发掘出来，会导致严重的后果。
- 庞大的文件尺寸。封闭的二进制办公软件文档格式通常尺寸庞大，这是因为其内部保存了诸多软件相关的冗余信息和隐藏信息。

为了改变办公软件和文档格式的现状，2003 年在国家“863”项目“中文 Linux 和办公软件相关标准与规范”的支持下，成立了中文办公软件基础标准工作组。中文办公软件基础标准项目以中文

办公软件普遍需求为出发点，在追踪、分析和消化国际办公软件相关标准、规范的基础上，借鉴各种先进技术，结合我国国情，从实际应用出发，经过七年多的努力，建立起一个自主可控的中文办公软件标准体系，包括以下部分：

- 《中文办公软件文档格式规范》。以中文办公软件需求为出发点，在分析、借鉴国际相关标准的基础上，结合我国国情，从实际应用出发，制定出的针对文字处理文档、电子表格和演示文档三种主要文档格式的描述体系。支持不同的办公软件之间文档的兼容和互换。
- 《中文办公软件用户界面标准》。针对各中文办公软件的用户界面特点，统一操作对象的标识、操作步骤的菜单提示信息，归纳中文办公软件最基本的交互方式和界面要素。为国内办公软件用户提供统一的用户界面和工作环境，降低学习成本、提高工作效率。
- 《中文办公软件应用编程接口规范》。规定与中文办公软件文档格式相适应的中文办公软件应用编程接口（API）。为中文办公软件的二次开发制定一个统一的开发接口，实现平台无关、语言无关、产品无关，利于产品的移植和软件重用，提高二次开发的效率，便于与其他应用系统集成。

《中文办公软件文档格式规范》也称为“统一办公文档格式”（中文简称“标文通”，英文为 Uniform Office Format，英文缩写 UOF），参见文献[1]。该标准共分 6 章，9 个附录，共近 554 页。主要内容分为 3 个部分，即文档格式规范说明、文档存储格式规范说明以及规范性附录。其内容结构如下：

- 文档格式规范说明。该部分为 W3C Schema 语言（参见文献[2]）定义的文档格式标准，并针对描述文件结构的元素和属性加以说明，定义符合标准的文档必须满足的体系结构。每个规范性描述文件构成独立的命名空间，作为管理和控制元素和属性集合的基本单元。
- 文档存储格式规范说明。该部分描述以压缩包为基础的存储格式，针对文档的存储处理和文档交换，描述了物理文档的基本结构。

《中文办公软件文档格式规范》的附录主要包括：

- XML Schema。这是采用 W3C Schema 定义的标文通结构。
- 用户 XML 数据的支持方法。用于支持用户定义的 XML 数据，使得标文通可以带有用户定义的逻辑内容，并与格式描述相关联。
- 多语种置标支持方案。用于支持标文通多语言置标版本的转换。标文通文档的主版本是中文，通过标识符映射机制，可以将其与其他语言置标的版本正确地交换内容，从而支持标准的国际化。
- 标文通支持功能扩展的策略。用于支持用户个性化扩展，以解决不同办公软件功能差异性的问题，满足不同用户的需要。
- 域代码定义。规定了各种功能域的代码形式。
- 数字格式的格式码描述。规定了各种数字格式的描述方式。
- 文字处理数字格式枚举说明。对文字处理中数字格式的枚举类型进行了描述。
- 图案（纹理）填充枚举。对图案（纹理）填充枚举类型进行说明。
- 自选图形种类枚举。对自选图形种类枚举进行说明。

标文通体现了如下的特点：

- 基于中文办公软件功能需求，充分反映中文办公软件的特点。
- 尽量采用正式的国家标准、国际标准或行业规范。
- 中文办公软件文档格式基于可扩展置标语言 XML。采用 W3C XML Schema 作为文档格式标准定义语言，尽量采用成熟的开放标准，例如，将 SVG 作为统一的图形描述规范，MathML 作为数学公式描述规范，等等。
- 采用中文作为标准定义的基础语言，支持多语言置标版本。
- 形成独立、完整、开放和可扩展的文档描述体系架构，方便用户扩展。
- 可以嵌入用户数据。本标准通过 UOF 元素与用户 XML 实例元素的对应，可以方便地从中文办

公软件文档中提取用户数据，或导入用户数据到中文办公软件文档之中。实现内容和格式的分离与融合，方便数据交换和应用集成；

- 文档标准体系架构支持模块的可重用性，减少文档描述的冗余，保证文档简洁易用。

需要指出的是，文档格式标准是随着信息化建设的需求而不断变化的。需求的改变、技术的进步，以及错误的更正均使标文通所规定的 Schema 可能会不定期更新。读者可以从地址 http://www.uof.org.cn/获得标文通 Schema 的最新版本。标文通 Schema 版本的更新将遵循下述约定：除非指定了新的命名空间，否则同一命名空间的新版本 Schema 将完全兼容旧版本的 Schema，这意味着用户所有旧版本标文通格式的文档无需修改均可通过新版本 Schema 的验证。

1.2 标文通文件的存储格式

标文通的数据以文件形式存储，采用了特殊的文件结构。标文通的文件存储格式设计有两个出发点：一是便于信息检索，二是节省存储空间。然而，很多情况下，这两者之间是相互矛盾的。例如，因为 XML 数据是纯文本的树形结构，十分便于检索，人们希望标文通的文件存储格式就是一个标准的 XML 文件，然而，有些办公文档含有大量图片等多媒体数据，采用纯文本形式描述文件体积会很庞大，因此希望尽可能有效地加以压缩。标文通文件存储格式的设计兼顾了这两种需求。

针对上述第一种需求，标文通文件可以存储成标准的 XML 形式，这时，文件的结构见图 1-1 文件的结构。

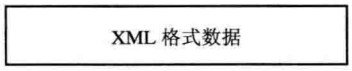

图 1-1 文件的结构

还有一种特殊情况，办公文档如果含有用户定义的逻辑内容，例如，需要对一个财务账单排版，而用户又希望财务账单能够独立于格式数据单独存取，这时标文通文件存储结构如图 1-2 所示。

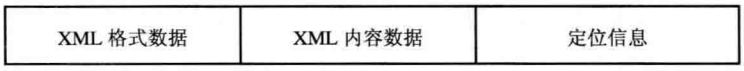

图 1-2 文件的另一种存储结构

其中，“XML 格式数据”存储标文通的格式信息，“XML 内容数据”存储用户数据。针对上述例子，“XML 内容数据”存储的是 XML 描述的用户账单数据，这些 XML 数据均由标识符属性与 XML 格式数据的排版描述相对应，例如，指定账单数据中的“品名”对应格式数据相应位置的以特定字形表现的“句”。上图中的“定位信息”指明以字节表达的“XML 格式数据”的结束和“XML 内容数据”的起始位置。这样使得一个标文通文档既可存储格式、内容混排的文档，又可存储格式、内容分离的文档。

当一个文档包含插图或其他多媒体内容时，可以有两种存放办法：一是将这些数据按 Base64 编码成 ASCII 字符，这样就可以内嵌在 XML 文件之中；二是可以将这些内容按二进制存放在单独的数据块中，也称外挂方式。这时，标文通的文件格式如图 1-3 所示。

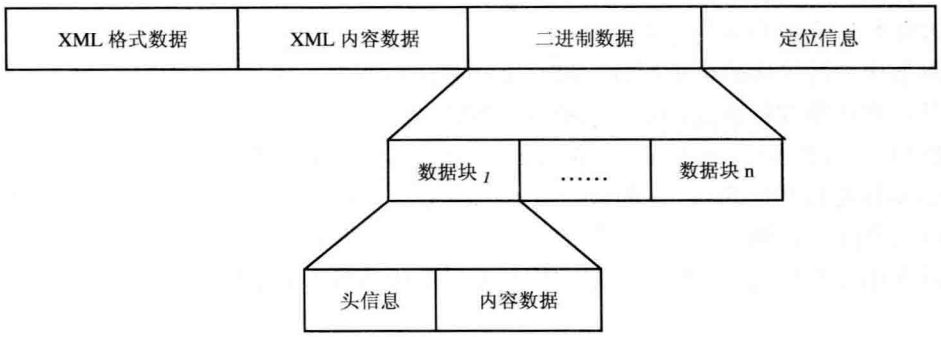

图 1-3 标文通文件的存储格式

在 XML 数据区之后，有一个二进制数据区，它由一个以上的数据块组成，每个二进制数据块均代表一个多媒体的内容。每个多媒体数据有可能含有头信息和内容数据，随媒体格式不同而不同。可以通过各个数据块的头信息定位各个数据块。XML 数据与外挂数据块通过内部名称进行关联。

一般来说，一个标文通文件包括 XML 格式数据、XML 内容数据、多媒体二进制数据以及定位信息。XML 格式数据和内容数据可以压缩，也可以不压缩，由应用程序指定。压缩 XML 数据可以大大减小文件的存储空间和传送时间，也能增加文档的安全保密性。如果不对 XML 数据进行压缩，XML 文件以它原有的结构存在，可以方便检索和快速浏览。当压缩时，XML 数据（包括内嵌的 Base64 编码数据）采用文本压缩方法。压缩过的文件，也由头信息和内容数据构成。

二进制数据块的内容组成见表 1-2。

表 1-2 二进制数据块内容组成

名称	字节数	说明
HEADER_ CRC	2	循环冗余码校验块（从 BLOCK_SIZE 到 COMMENT_DATA）
BLOCK_SIZE	2	块的大小（从 BLOCK_SIZE 到 COMMENT_DATA）
BLOCK_FLAGS	1	块标志。第 1 位为 0，第 2 位如果为 1，表示包里包含 1 个注释块，否则为 0。其他位为 0。
COMP_CODE	1	取值含义如下： 0　无压缩 1　对文本使用无损压缩算法-PPMII 2　对图像使用无损压缩算法-PNG 3　对图像使用有损压缩算法-JPEG2000 4　对声音使用有损压缩算法-VORBIS 5　预留 …　预留 254　预留 255　对其他文件使用无损压缩算法-DEFLATE
DATA_CRC32	4	文件数据的 32 位循环冗余码校验
COMP_SIZE	4	压缩后文件大小
ORIG_SIZE	4	未压缩文件大小
NAME_LEN	2	文件名长度
FILE_NAME	变长	文件名
COMMENT_SIZE	2	注释块大小，只有当 BLOCK_FLAGS 的第 1 位置 1 时才存在
COMMENT_DATA	变长	注释块，只有当 BLOCK_FLAGS 的第 1 位置 1 时才存在
COMPRESSED_DATA	变长	压缩数据块，随 COMP_FLAGS 的值不同，该块结构也不同

二进制文件打包时自动针对不同的信息类型和用户的选择，由打包压缩模块自动采用最适合的压缩算法进行压缩和打包处理。

定位信息描述 XML 文件的文件大小，快速定位各文件块的位置。定位信息放在标文通文档的最后，这基于两种考虑：1）标文通文档一开始就是 XML 文件，可以方便读取文档起始部分的信息，而这些信息对于文本检索时提取 XML 文件内容摘要是非常有用的；2）方便程序直接找到定位信息后，快速定位 XML 文件之后的各个文件块。当文档不存在 XML 内容数据和外挂文件且 XML 数据不进行压缩时，不需要文件定位信息。

文件定位信息的组成见表 1-3。

表 1-3 文件定位信息的组成

名称	字节数	说明
XML_USR_FILE_SIZE	4	XML 内容数据块的大小，当该部分不存在时为 0
XML_FILE_SIZE	4	全部数据构成的文件大小之和，当 XML 内容数据和外挂数据不存在时，该数值就是 XML 格式数据的大小
XML_FILE_FLAG	4	文件标记。第 1 个字节预留； 第 2 个字节（X）设置如下： ● 如果 X&0x01 非 0 表示压缩 XML 格式数据，否则不压缩； ● 如果 X&0x02 非 0 表示压缩 XML 内容数据，否则不压缩； 最后 2 个字节置 0，用来判断是否是一个未压缩的纯 XML 文件包。如果一个文档的最后两个字节是"/〉"，表示是一个未压缩的纯 XML 文件包，否则不是

标文通文件存储应该能够支持表 1-4 中常用媒体数据。

表 1-4 常用媒体数据

文本	txt / xml / htm / html
声音	wav / mid / rmi / au / mp3 / rm / ra / snd
图像（含图形）	png / jpg / bmp / pbm / ras / gif / svg
动态视频	avi / mpg / qt / rm / asf / animation-gif
其他	

标文通支持的数据压缩方法见表 1-5。

表 1-5 标文通支持的数据压缩方法

数据格式		压缩算法
纯文本	XML	PPMII
图像	BMP	PNG JPEG2000
	TIF	
	GIF	
	PNG	
	JPG	
音频、视频	AVI	AVS
通用无损压缩		DEFLATE

对于其他格式的数据如果需要压缩，一般先转换成上述格式后再进行压缩，这意味着经过中文办公软件的处理，可能会改变原先的数据格式。但这对于办公文档来说并不重要。

例如，假设 MyFile.uof 由以下部分组成，见表 1-6。

表 1-6 MyFile 的内容

类别	名称	大小（字节）	备注
XML 格式数据	配音图片.uof	9972	不压缩
XML 内容数据	多媒体数据.xml	103	不压缩
外挂图片	BlueHills.jpg	28521	
外挂音频	Town.mid	22097	

打包后的文件 MyFile.uof 的情况如表 1-7。

表 1-7 MyFile 打包后的文件定位信息

名称	字节数	实际值	说明
XML_USR_FILE_SIZE	4	0x00 0x00 0x00 0x67	表示 XML 内容数据（即“多媒体数据.xml”）大小为 0x67，十进制为 103 字节。
XML_FILE_SIZE	4	0x00 0x00 0x27 0x5B	表示文件的大小为 0x275B，十进制为 10075，是 XML 格式数据（“配音图片.uof”）和 XML 内容数据（“多媒体数据.xml”）之和。
XML_FILE_FLAG	4	0x00 0x00 0x00 0x00	第 1 字节 0x00（预留）； 第 2 字节 0x00，由于&0x01、&0x02 均为 0，表示不压缩 XML 格式数据，亦不压缩 XML 内容数据； 第 3 字节 0x00、第 4 字节 0x00，由于不是“/>”，表示不是一个未压缩的纯 XML 文件包。

据此，我们可以推断出，XML 格式数据（“配音图片.uof”）的大小为 9972 字节（10075-103）。从文件开始，截取 9972 字节，便可获得“配音图片.uof”的数据内容。

从第 9972 字节（从 0 开始计数）开始，取 103 字节，便可获得“多媒体数据.xml”的数据内容。

从 10075 字节开始，便是第 1 个二进制数据块的开始，见表 1-8。

表 1-8 MyFile 的第 1 个外挂数据块

名称	字节数	实际值	说明
HEADER_ CRC	2	0x04 0x60	从 BLOCK_SIZE 到 COMMENT_DATA 的循环冗余码校验块
BLOCK_SIZE	2	0x00 0x1C	从 BLOCK_SIZE 到 COMMENT_DATA 的块大小，即表示头部有 28 个字节
BLOCK_FLAGS	1	0x00	表示没有注释块
COMP_CODE	1	0x05	预留字节
DATA_CRC32	4	0xE8 0xE6 0xA3 0x13	文件数据的 32 位循环冗余码校验
COMP_SIZE	4	0x00 0x00 0x16 0x93	压缩后文件大小为 5779
ORIG_SIZE	4	0x00 0x00 0x56 0x51	未压缩文件大小为 22097 字节
NAME_LEN	2	0x00 0x08	文件名长度为 8 个字节
FILE_NAME	变长	0x54 0x6F 0x77 0x6E 0x2E 0x6D 0x69 0x64	文件名为“Town.mid”
COMMENT_SIZE	2	不存在	注释块大小
COMMENT_DATA	变长	不存在	注释块
COMPRESSED_DATA	变长	0x01 0x00 0x78 0xDA…	压缩数据块，总计 5779 个字节

从 15882（10075+28+5779）字节开始，便是第 2 个二进制数据块的开始，见表 1-9。

表 1-9 MyFile 的第 2 个外挂数据块

名称	字节数	实际值	说明
HEADER_ CRC	2	0x7B 0xED	从 BLOCK_SIZE 到 COMMENT_DATA 的循环冗余码校验块
BLOCK_SIZE	2	0x00 0x21	从 BLOCK_SIZE 到 COMMENT_DATA 的块大小，即表示头部有 33 个字节
BLOCK_FLAGS	1	0x00	表示没有注释块
COMP_CODE	1	0x05	预留字节
DATA_CRC32	4	0x6B 0x16 0x19 0x58	文件数据的 32 位循环冗余码校验
COMP_SIZE	4	0x00 0x00 0x68 0x47	压缩后文件大小为 26695 字节
ORIG_SIZE	4	0x00 0x00 0x6F 0x69	未压缩文件大小为 28521 字节
NAME_LEN	2	0x00 0x0D	文件名长度为 13 个字节
FILE_NAME	变长	0x42 0x6C 0x75 0x65 0x48 0x69 0x6C 0x6C 0x73 0x2E 0x6A 0x70 0x67	文件名为“BlueHills.jpg”
COMPRESSED_DATA	变长	0x01 0x00 0x78 0xDA…	压缩数据块，总计 26695 个字节

《中文办公软件应用程序开发接口规范》提供了一套 API 用于标文通的文件存储，中文办公软件的生产商应该提供对该 API 的支持，中文办公软件的二次开发商应该采用该 API 实现对标文通文件的访问。

标文通文件的后缀可以是：

- .uof：表示一般的标文通文件，可以是任意的文字处理、电子表格或演示文稿文档。
- .uot：表示标文通格式的文字文档。某些用户可能需要从文件后缀上能够区别出是文字处理文档。
- .uos：表示标文通格式的电子表格文档。某些用户可能需要从文件后缀上能够区别出是电子表格文档。
- .uop：表示标文通格式的演示文稿文档。某些用户可能需要从文件后缀上能够区别出是演示文稿文档。

除上述后缀之外，标文通还有一套为模板文件保留的后缀，分别是：

- .uott：表示文字处理文档模板。
- .uost：表示电子表格文档模板。
- .uopt：表示演示文稿文档模板。

对于上述后缀的文档，是否经过压缩打包，则要根据文件定位信息判断，不能简单地从文档名称后缀上判断。

1.3 标文通的命名空间

XML 的命名空间是为了解决命名冲突，为元素和属性命名而引入的逻辑空间。标文通采用 7 个命名空间来描述文档格式，见表 1-10。

表 1-10 标文通的命名空间

名称	前缀	命名空间 URI	说明
uof.xsd	uof	http://schemas.uof.org/cn/2003/uof	标文通的最高层，包括了文字处理、电子表格和演示文档三种办公应用共同具有的基本元素
字.xsd	字	http://schemas.uof.org/cn/2003/uof-wordproc	包括文字处理使用的基本元素，同时也提供给其他应用使用
表.xsd	表	http://schemas.uof.org/cn/2003/uof-spreadsheet	包括电子表格使用的基本元素，同时也提供给其他应用使用
演.xsd	演	http://schemas.uof.org/cn/2003/uof-slideshow	包括演示文档使用的基本元素，同时也提供给其他应用使用
图.xsd	图	http://schemas.uof.org/cn/2003/graph	包括绘图相关的主要元素
数.xsd	数	http://www.w3c.org/1998/Math/MathML	包括与数学符号/公式相关的主要元素。这里直接采用 MathML 标准
svg.xsd	svg	http://www.w3.org/2003/svg	包括图形相关的主要元素。这里直接使用 SVG 标准

标文通将一如既往尽可能采用已制定的标准。

1.4 认识简单的文字处理文档

下面从一个简单的例子（图 1-4）开始，了解一下文字处理文档的基本结构。

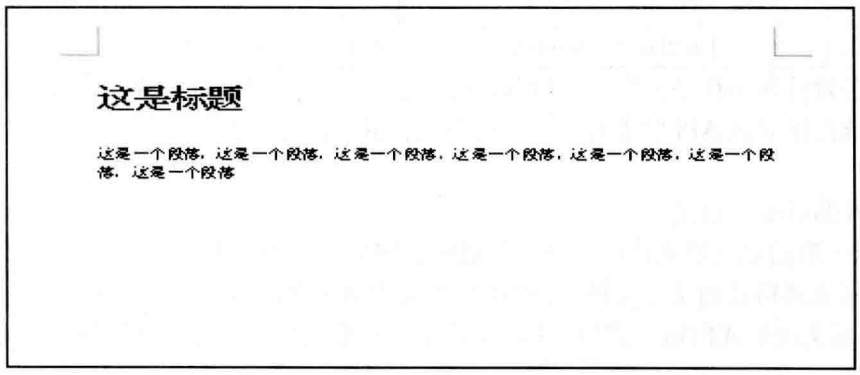

图 1-4 一个简单的文字处理文档

对于该例所示的文档，其相应内容见代码 1-1。

代码 1-1 一个简单的文字处理文档

```
1   ?xml version="1.0" encoding="UTF-8"?>
2   <uof:UOF xmlns:uof="http://schemas.uof.org/cn/2003/uof" xmlns:图="http://schemas.uof.org/cn/2003/graph" xmlns:字
    ="http://schemas.uof.org/cn/2003/uof-wordproc" xmlns:表="http://schemas.uof.org/cn/2003/uof-spreadsheet" xmlns:演
    ="http://schemas.uof.org/cn/2003/uof-slideshow" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://schemas.uof.org/cn/2003/uof D:\UOF\uof_schema\uof.xsd" uof:language="cn" uof:version="1.1"
    uof:mimetype="vnd.uof.text" uof:locID="u0000">
3   <uof:元数据 uof:locID="u0001">
4   <uof:作者 uof:locID="u0005">ZYH</uof:作者>
5   <uof:创建日期 uof:locID="u0008">2010-01-11T11:11:11</uof:创建日期>
6   <uof:编辑次数 uof:locID="u0009">1</uof:编辑次数>
7   <uof:编辑时间 uof:locID="u0010">P0Y0M0DT0H3M10S</uof:编辑时间>
8   <uof:创建应用程序 uof:locID="u0011">EIOffice 2009</uof:创建应用程序>
9   <uof:公司名称 uof:locID="u0018">bistu</uof:公司名称>
```