ICS 01.140
CCS A 14

WH

中华人民共和国文化行业标准

WH/T 100—2023

汉文古籍版式描述规范

The format description for Chinese ancient books

2023-09-09 发布　　2023-12-09 实施

中华人民共和国文化和旅游部　发布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由中华人民共和国文化和旅游部提出。

本文件由全国图书馆标准化技术委员会（SAC/TC 389）归口。

本文件起草单位：国家图书馆、天津图书馆、北京汉王数字科技有限公司。

本文件主要起草人：肖禹、陈红彦、张毅、董馥荣、李志峰、胡艳杰、白帆、王昭、杜立功、赵依澍、周升川、潘慧敏、谢冬荣、萨仁高娃、李国庆、江世盛、刘正珍、王晓健、王战波。

汉文古籍版式描述规范

1 范围

本文件对汉文古籍版式描述进行了规范,给出了版式描述的规范性要求。

本文件适用于对汉文古籍文本化加工结果的描述。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中,注日期的引用文件,仅该日期对应的版本适用于本文件;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。

GB/T 3792—2021 信息与文献 资源描述

GB/T 4894—2009 信息与文献 术语

GB/T 18793—2002 信息技术 可扩展置标语言(XML)1.0

GB/T 21712—2008 古籍修复技术规范与质量要求

GB/T 31219.2—2014 图书馆馆藏资源数字化加工规范 第2部分:文本资源

3 术语和定义

下列术语和定义适用于本文件。

3.1

古籍 ancient books

1911年以前(含1911年)在中国书写或印刷的书籍。

[来源:GB/T 3792—2021,3.18]

3.2

书叶 page

按文稿顺序排列的书写、印制的单张纸叶。

[来源:GB/T 21712—2008,2.9]

3.3

版框 a rectangular frame into which pages engraving

书叶(3.2)正面图文四边的围栏,一般指印刷的书。

[来源:GB/T 21712—2008,2.10]

3.4

版心 middle of page

书叶(3.2)左右对折的正中、在折叶时取作中缝标准的条状行格。雕版印刷的书籍版心通常印有书名、卷次、叶码,有的还印有一版文字总数、刊刻机构以及刻工姓氏等。

[来源:GB/T 21712—2008,2.11]

3.5

版式 format

汉文古籍的版面格式。

3.6

版式 XML 文件　format XML file

符合 GB/T 18793—2002 要求，对具有同一样式的书叶（3.2）共同拥有的版式（3.5）特点进行描述的 XML 文件。

3.7

文本　text

以字符、符号、词、短语、段落、句子、表格或者其他字符排列形成的数据，用于表达意义，其解释基本上取决于读者对于某种自然语言或者人工语言的知识。

[来源：GB/T 4894—2009，4.1.1.2.4]

3.8

图像　image

用各种观测系统以不同形式和手段观测客观世界而获得的，可以直接或间接作用于人眼进而产生视知觉的实体。

[来源：GB/T 31219.2—2014，3.3]

4　基本原则

4.1　客观描述

对汉文古籍版式特点、文本内容和位置、插图大小和位置等内容的描述客观准确。

有无版框、四周单边、四周双边、左右双边、大小字等。

4.2　描述唯一

每个书叶描述方式唯一，且每个描述数据解释方式唯一，没有歧义。拥有统一版式的多个书叶，版式相关数据描述方式唯一，包括版框位置、版心位置、文本行数、每行文字数、文字大小等。

4.3　易实现

版式描述形式简单，使 XML 文件容易加工和解释，利于汉文古籍文本后续更深层次加工使用。

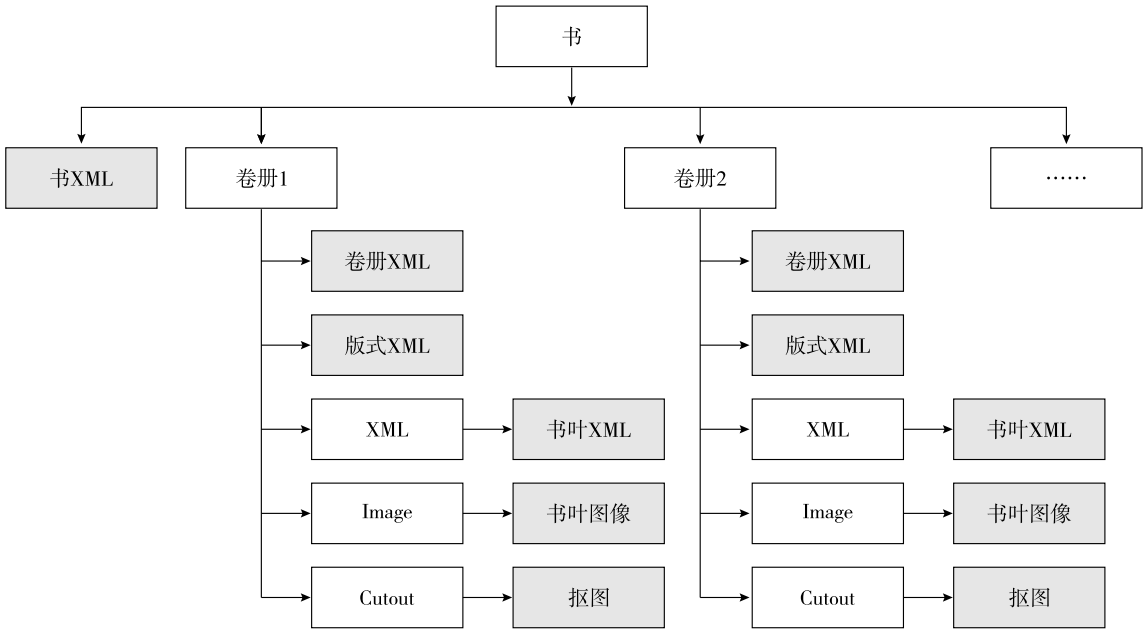
4.4　可扩展

可对 XML 进行扩展以适应更多的汉文古籍样式，例如新的字符修饰样式等。扩展部分是新增加的内容，不使用已有的内容代替，不与已有内容发生逻辑冲突。

5　汉文古籍版式描述

5.1　概述

汉文古籍版式描述存放目录可分为 3 个层级：第一层级为书文件夹；第二层级为卷册文件夹；第三层级为书叶 XML 文件夹、书叶图像文件夹及抠图文件夹。其中，书文件夹和卷册文件夹的命名可自定义，书叶 XML 文件夹名称应为“XML”，书叶图像文件夹名称应为“Image”，抠图文件夹名称应为“Cutout”。存放目录结构见图 1：



注：白色背景的为文件夹，灰色背景的为文件。

图 1　汉文古籍版式描述数据存储目录结构示意图

汉文古籍版式描述存放目录中内容应符合以下要求：

a）书文件夹存放一部书的所有数据。书文件夹下包括一个书 XML 文件和一个或者多个卷册文件夹。

b）卷册文件夹包括一个卷册 XML 文件、一个版式 XML 文件、一个 XML 文件夹、一个 Image 文件夹和一个 Cutout 文件夹。

c）卷册 XML 文件描述卷册包含的书叶和书叶的顺序。

d）版式 XML 文件描述卷册的版式信息。

e）XML 文件夹存放卷册中的书叶 XML 文件。

f）Image 文件夹存放卷册中的书叶原始图像。

g）Cutout 文件夹存放卷册中的所有插图和集外字、模糊字的抠图。

h）汉文古籍版式描述数据保存在版式 XML 和书叶 XML 中。

5.2　基于 XML 的版式描述

5.2.1　版式 XML 文件

版式 XML 文件用来统一描述卷册中书叶的版式信息，有助于书叶样式严格统一，其命名规则为：Format.xml，版式 XML 文件的标签及其属性见表 1：

表 1　版式 XML 文件标签

XML 标签名	注解	说明	属性			样例
			属性名称	注解	说明	
xml	文档类型定义	定义文档版本编码	version	版本	XML 文件的版本	<?xml version="1.0" encoding="utf-8"?>
			encoding	编码	文字编码	
root	根节点	根节点	version	版本	版式 XML 文件的版本	<root version="1.0">
formats	版式列表	定义一组版式	无	无	无	<formats>

表 1　版式 XML 文件标签（续）

XML 标签名	注解	说明	属性			样例
			属性名称	注解	说明	
format	版式	formats 的子节点，定义一种版式	name	版式的名称	用户对版式的命名	<format name="[光绪]顺天府志" dpi="72" page_width="861.59" page_height="770.40" page_frame="50.40,96.23,812.63,733.91">
			dpi	版式的基准 DPI	根据版式的基准 DPI，可将版式中的像素值转换为毫米、厘米等物理长度值	
			page_width	版式的书叶宽	版式书叶宽度的像素值	
			page_height	版式的书叶高	版式书叶高度的像素值	
			page_frame	版式的书叶版框位置	版式书叶版框的像素位置，以“,”分隔开的 4 个数值，依次代表左上右下的像素值	
using_page	版式作用于哪些书叶	确定哪些书叶使用该版式	page_id_range	书叶 id 范围	使用该版式的书叶 id 的范围，连续的书叶 id 使用“-”连接起始和结束叶的 id 值，不连续时使用“,”连接	<using_page page_id_range="2-23,25" odd_even="0" />
			odd_even	奇偶性	0：所有叶码 1：奇数叶码 2：偶数叶吗	
text_formats	文本版式列表	定义一组文本版式	无	无	无	<text_formats>
text_format	文本版式	text_formats 的子节点，定义一个文本版式	region	文本的区域位置	文本的矩形区域位置，属性值为矩形的左上右下 4 边的像素值以符号“,”连接	<text_format region="421.07,114.00,442.20,218.44" font_id="3" para_style_id ="1" alignment="0" direction="1"/>
			font_id	字体 ID	字体列表中的一个字体 ID	
			para_style_id	段落样式 ID	段落样式列表中的一个段落样式 ID	
			alignment	对齐方式	0：头部对齐（横排左对齐，竖排上对齐） 1：居中对齐 2：尾部对齐（横排右对齐，竖排下对齐）	
			direction	文字方向	0：横排 1：竖排	
images	图像列表	定义一组图像	无	无	无	<images>
image	图像	images 的子节点，定义一个图像	name	版式图像的文件名	版式所需图像的文件名，图像存放于 Cutout 文件夹中	<image name="上鱼尾.jpg" region="419.63,238.13,443.40,266.94" />
			region	图像的区域位置	图像的矩形区域位置，属性值为矩形的左上右下 4 边的像素值以符号“,”连接	
lines	线段列表	定义一组线段	无	无	无	<lines>
line	线段	lines 的子节点，定义一条线段	start_point	起始点	起始点坐标的像素值	<line start_point="287.63,282.15" end_point="307.12,282.15" weight="0.96" />
			end_point	结束点	结束点坐标的像素值	
			weight	线宽	像素数线宽	

表 1　版式 XML 文件标签（续）

XML 标签名	注解	说明	属性			样例
			属性名称	注解	说明	
rectangles	矩形框列表	定义一组矩形框	无	无	无	<rectangles>
rectangle	矩形框	rectangles 子节点，定义一个矩形框	region	矩形框的区域位置	属性值为矩形框的左上右下 4 边的像素值以符号“,”连接	<rectangle region="731.07,189.00,1231.20,689.44" weight="1.05" />
			weight	线宽	像素数线宽	
box_and_line	边框栏线信息	节点信息为空时，表示无边框栏线	middle_area_width	版心宽度	像素数	<box_and_line middle_area_width="25.93" box_space="4.80,4.80,4.80,4.80" left_column_num="10" right_column_num="10" show_column_line="0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0,0" column_line_weight="0.96" out_box_weight="5.27" inner_box_weight="0.96" />
			box_space	内外边框距离的像素数	内外边框距离，用“,”分隔的 4 个值依次表示左上右下 4 个方向上内外框线之间的距离，如果为 0 则表示没有内框线	
			left_column_num	版心左侧行数	版心左侧行数	
			right_column_num	版心右侧行数	版心右侧行数	
			show_column_line	是否显示栏线	从右向左按顺序描述，用“,”分隔 0：不显示 1：显示 允许值为空字符串，值为空时显示所有栏线	
			column_line_weight	栏线宽	栏线像素宽度	
			out_box_weight	外边框宽	外边框像素宽度	
			inner_box_weight	内边框宽	内边框像素宽度	
fonts	字体列表	定义一组字体	无	无	无	<fonts>
font	字体	fonts 的子节点，定义一种字体	id	字体 ID	id 值从 1 开始且唯一，用于区分字体，在使用该字体的地方引用此 id 值	<font id="1" name="大字" face="@宋体" size="19.29" width_stretch_ratio="1.20" char_space="0.00" location_type="0" style="7"/>
			name	字体的名称	用户对字体的命名	
			face	字体类型	字体在字体文件中的名称，加字符“@”表示竖排	
			size	字号	字体大小，以像素为单词，也是字体的高度值	
			width_stretch_ratio	字体宽度缩放比例	字体宽=size*width_stretch_ratio	
			char_space	字间距	字与前一字之间的像素距离	
			location_type	位置类型	0：字位于版框内 1：字位于版框外 如果处于上文定义的版框区域以外，则不用来进行高起计算	

表 1　版式 XML 文件标签（续）

XML 标签名	注解	说明	属性			样例
			属性名称	注解	说明	
font	字体	fonts 的子节点，定义一种字体	style	字体风格	0：正常 1：加粗 2：斜体 4：加下划线 8：阳文 16：阴文 风格可以叠加，如值为 1+2+4，表示文字加粗、斜体，并且加下划线	
para_styles	段落样式列表	定义一组段落样式	无	无	无	<para_styles>
para_style	段落样式	para_styles 的子节点，定义一种段落样式	id	段落样式 ID	id 值从 1 开始且唯一，用于区分段落样式，在使用该段落样式的地方引用此 id 值	<para_style id="1" name="段落样式 1" line_space="0.00" head_space="18.08" tail_space="18.08"/>
			name	段落样式的名称	用户对段落样式的命名	
			line_space	行间距	行与前一行之间的像素距离	
			head_space	首字距版框位置	像素数，计算高起边框时用到	
			tail_space	尾字距版框位置	像素数	

5.2.2　书叶 XML 文件

书叶 XML 文件用来描述汉文古籍每一个书叶的具体信息，具体 XML 标签和属性见表 2：

表 2　书叶 XML 文件标签

XML 标签名	注解	说明	属性			样例
			属性名称	注解	说明	
xml	文档类型定义	定义文档版本编码	version	版本	XML 文件的版本	<?xml version="1.0" encoding="utf-8"?>
			encoding	编码	文字编码	
root	根节点	根节点	version	版本	书叶 XML 文件的版本	<root version="1.0">
page	书叶	描述书叶内容	page_id	书叶的 id 号	书叶的 id 号，从 1 开始并且在全书中具有唯一性	<page page_id="1" dpi="300" page_width="3590.00" page_height="3210.00" page_frame="226.00,401.00,3374.00,3068.00" image_name="001.jpg">
			dpi	书叶的基准 DPI	根据书叶的基准 DPI，可将书叶中的像素值转换为毫米、厘米等物理长度值	

表 2　书叶 XML 文件标签（续）

XML 标签名	注解	说明	属性			样例
			属性名称	注解	说明	
page	书叶	描述书叶内容	page_width	书叶宽	书叶宽度的像素值	
			page_height	书叶高	书叶高度的像素值	
			page _frame	书叶版框位置	书叶版框的像素位置，以“,”分隔开的 4 个数值，依次代表左上右下的像素值	
			image_name	书叶图像的名称	书叶文件对应的原图的名称	
format_texts	文本版式对应的文字列表	定义一组文本版式对应的文字	无	无	无	<format_texts>
format_text	文本版式对应的文字	format_texts 的子节点，该文字的坐标及字体效果等从版式文件中取得	无	无	无	<format_text> [光绪]顺天府志</format_text>
blocks	图文区域列表	定义一组图文区域	无	无	无	<blocks>
image_block	插图图像区域	blocks 的子节点，定义一个插图图像区域	region	插图区域位置	插图的矩形区域位置，属性值为矩形的左上右下 4 边的像素值以符号“,”连接	<image_block region="244.00,416.00,1748.00,3028.00" image_name="017-KT-001.jpg"/>
			image_name	插图的图像文件名	插图图像保存在 Cutout 文件夹中	
text_block	文本区域	blocks 的子节点，定义一个文本区域	region	文本区域位置	文本的矩形区域位置，属性值为矩形的左上右下 4 边的像素值以符号“,”连接	<text_block region="3228.00,526.00,3331.00,3017.00">
text_line	文本行	text_block 的子节点，定义一个文本行	region	文本行的区域位置	文本行的矩形区域位置，属性值为矩形的左上右下 4 边的像素值以符号“,”连接	<text_line region="3228.00,526.00,3331.00,3017.00" column_index="0" direction="1" para_style_id ="1" bussiness_type="0">
			column_index	栏的索引值	文本行所属栏的索引，从 0 开始，属性信息为空时表示没有分栏	
			direction	文字方向	0：横排 1：竖排	
			para_style_id	段落样式 ID	段落样式定义见版式文件	
			bussiness_type	大小字	0：大字 1：小字	

表 2　书叶 XML 文件标签（续）

XML 标签名	注解	说明	属性			样例
			属性名称	注解	说明	
char	文本字符	text_line 的子节点，定义一个文本字符	region	文本字符的区域位置	文本字符的矩形区域位置，属性值为矩形的左上右下 4 边的像素值以符号“,”连接	<char region="2478.00,2221.00,2581.00,2319.00" font_id="1" rotation="0">通</char>
			font_id	字体 ID	字体定义见版式文件	
			rotation	角度	单个字符的旋转角度	
			ids	表意文字描述字符串	表意文字描述字符串	
blur	模糊字	text_line 或 format_text 的子节点，定义一个模糊字	region	模糊字的区域位置	模糊字的矩形区域位置，属性值为矩形的左上右下 4 边的像素值以符号“,”连接	<blur region="1469.00,2820.00,1572.00,2917.00" image_name="006-BL-001.jpg"/>
			image_name	模糊字抠图文件名	模糊字抠图的图像文件名，图像保存在 Cutout 文件夹中	
bracket	括号	text_line 或 format_text 的子节点，char、gaiji 或 blur 的父节点，定义一对括号	style	括号风格	0：加框 1：加中括号 2：加八边形	<bracket style="2" type="0"><char region="1727.00,323.00,1795.00,393.00" font_id="1" rotation="0">通</char><char region="1727.00,394.00,1795.00,464.00" font_id="1" rotation="0">州</char></bracket>
			type	括号类型	0：完整的一对括号 1：头括号 2：尾括号 通过定义头括号和尾括号的类型，支持一对括号跨行、跨页的情况	
lines	线段列表	定义一组线段	无	无	无	<lines>
line	线段	lines 的子节点，定义一条线段	start_point	起始点	起始点坐标的像素值	<line start_point="287.00,282.00" end_point="307.00,282.00" weight="1.00" />
			end_point	结束点	结束点坐标的像素值	
			weight	线宽	像素数线宽	
rectangles	矩形框列表	定义一组矩形框	无	无	无	<rectangles>
rectangle	矩形框	rectangles 的子节点，定义一个矩形框	region	矩形框位置	属性值为矩形框的左上右下 4 边的像素值以符号“,”连接	<rectangle region="731.00,189.00,1231.00,689.00" weight="1.00" />
			weight	线宽	像素数线宽	