

CY

中华人民共和国新闻出版行业标准

CY/T XXX.5—XXXX

中文古籍数字出版规范
第5部分：内容采集

Standard for digital publishing of Chinese ancient books—Part 5: Content extract

(点击此处添加与国际标准一致性程度的标识)

(征求意见稿)

(本草案完成时间：2025-11-04)

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX—XX—XX 发布

XXXX—XX—XX 实施

国家新闻出版署 发布

目 次

前言 II

引言 III

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 采集目标 2

5 采集范围 2

 5.1 文字采集范围 2

 5.2 图表采集范围 2

 5.3 样式采集范围 2

 5.4 结构采集范围 2

6 采集流程 2

 6.1 识别 2

 6.2 文字校对 2

7 文字采集 2

 7.1 字库编码方式 2

 7.2 集外字处理 2

 7.3 标签定义 3

 7.4 质量要求 9

8 图表采集 10

 8.1 标签定义 10

 8.2 质量要求 11

9 样式采集 12

 9.1 标签定义 12

 9.2 质量要求 14

10 结构采集 14

 10.1 标签定义 14

 10.2 质量要求 15

11 中文古籍内容采集 XML Schema 文件及 XML 样例文件 15

附录 A （规范性） 中文古籍内容采集 XML Schema 文件 16

附录 B （资料性） 中文古籍内容采集 XML 样例文件 19

参考文献 20

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。

本文件是CY/T XXX《中文古籍数字出版规范》的第5部分。CY/T XXX已经发布了以下部分：

- 第1部分：术语；
- 第2部分：元数据；
- 第3部分：长期存储；
- 第4部分：版式采集；
- 第5部分：内容采集；
- 第6部分：版式重构；
- 第7部分：古籍数字加工与应用模式；
- 第8部分：古籍数据交换；
- 第9部分：数据加工管理。

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。

本文件由全国新闻出版标准化技术委员会（SAC/TC 527）归口。

本文件起草单位：

本文件主要起草人：

引 言

随着信息技术的进步，数字阅读已经普及，现如今，数字科技不断成熟，许多先进的技术被应用于内容资源数字化中，在新闻出版领域中图书、报纸、期刊、音频、视频的数字化和出版标准都已经发布并实施，而中文古籍数字出版标准尚处于缺失状态。为顺应数字化潮流、推动新时代古籍数字化工作高质量、创新性发展，更有效解决“藏”与“用”的问题，让中文古籍文献焕发新的生命力，使更多中文古籍文献得以深层次的挖掘和现代化的呈现，制定了CY/T XXX—XXXX《中文古籍数字出版规范》。依据中文古籍数字化生产过程，拟由9个部分组成。

——第1部分：术语。目的在于规范与中文古籍数字化相关的术语，统一相关概念，避免由于概念和术语不明确而造成的交流困难、歧义和误解。

——第2部分：元数据。目的在于规范中文古籍数字出版的元数据信息，便于理解数据的含义和用途，有助于提高数据的管理、组织、质量控制、存储、共享和安全保护的效率，为中文古籍元数据应用提供依据和指导。

——第3部分：长期保存。目的在于给出中文古籍长期存储数据的类型、存储原则、存储环境和存储备份策略的相关技术要求，为中文古籍数据长期保存提供依据和指导。

——第4部分：版式采集。目的在于规定中文古籍数字化加工中版式采集对象、采集范围和采集流程并对数据规格和质量要求提出技术要求，为中文古籍数字化版式采集提供依据和指导。

——第5部分：内容采集。目的在于规定中文古籍数字化加工中内容采集目标、采集范围和采集流程，并对文字采集、样式采集、结构采集提出技术要求，为中文古籍数字化内容采集提供依据和指导。

——第6部分：版式重构。目的在于给出中文古籍数字化加工中版式重构的部件元素组成、用字要求、描述文件要求和相应的质量要求，为中文古籍数字化版式重构提供依据和指导。

——第7部分：古籍数字化加工与应用模式。目的在于给出中文古籍数字化加工成品数据类型及规格要求，并描述了长期保存、古籍电子书、古籍资源库应用所需的成品数据类型，为中文古籍数字化加工应用提供依据和指导。

——第8部分：古籍数据交换。目的在于给出中文古籍数据交换类型、数据交换的基本要求和数据交换的接口要求，为中文古籍数据交换提供指导和帮助。

——第9部分：数据加工管理。目的在于给出中文古籍数字化加工的基本流程以及人员、环境、资料、数据存储、数据备份和数据交付的管理要求，为中文古籍数字化加工管理提供指导和帮助。

中文古籍数字出版规范

第5部分：内容采集

1 范围

本部分规定了中文古籍文献内容的采集目标、采集范围、采集流程，并对文字采集、图表采集、样式采集、结构采集等提出了具体要求。

本部分适用于描述书写或印刷于1912年以前具有中国古典装帧形式的汉文书籍。

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。

ISO/IEC 10646:2020 Information technology — Universal coded character set (UCS)

GB 18030 信息技术 中文编码字符集

GB/T 38548.5—2020 内容资源数字化加工 第5部分：质量控制

CY/T XXX. 4—XXXX 中文古籍数字出版规范 第4部分：版式采集

3 术语和定义

下列术语和定义适用于本文件。

3.1

外字 character outside the set

在指定集合范围内无法认同的字符。

3.2

简洁文字串 simplified text string

对页面中版式特殊的文字内容的一种表述方式，这种内容因版式特殊而难以实现单字采集，仅录入其文字，并以字符串格式处理。

示例1：中文古籍中对注文的注释。

3.3

校勘信息 text modify information

对页面内容中校勘的描述。

3.4

段落结束标记 paragraph end symbol

添加在行尾字符后的一种标记，用于表示段落的结束。

3.5

文字段 text paragraph

一行中版式信息相同的连续字符。

注：版式信息包括文字大小、文字所在行、文字方向等。

3.6

文字行 text row

同一行的所有文字段的总和。

3.7

文字块 text

同一页面中，由若干文字行组成的一段文字。

4 采集目标

通过对中文古籍的内容部分进行数字化加工、提取和标引，将采集内容形成应用于长期保存和数字出版的数字化形态文件。

中文古籍内容采集的成果物包括采集了文字、图表、样式和结构的XML文件。

5 采集范围

5.1 文字采集范围

文本采集范围包括字符文字、空白、标点符号、简洁文字串、校勘信息、文字块、标注等。

5.2 图表采集范围

图表采集包括对古籍中的图像和表格进行采集。

图像采集是对页面中的插图、印章、集外字图等信息的采集。

表格采集范围包括：表格的单元格、文字块、图像等内容。

5.3 样式采集范围

采集页面中文字、段落等样式信息：

- a) 字体样式信息包括：字体、字号、加粗、倾斜、底色等；
- b) 段落样式信息包括：提格、缩进等。

5.4 结构采集范围

采集页面中的章、节、页、列等信息。

6 采集流程

6.1 识别

使用版式采集获得文字内容块信息文件，可通过OCR识别处理或手工录入的方式，输出具有文字内容、文字字体和结构信息的过程文件。

6.2 文字校对

应对文字进行校对，具体校对方法如下。

- a) 横校：按阅读顺序、图文对照校对文字，使与图像中的文字保持一致。
- b) 纵校：打乱阅读顺序，按字符汉字聚类进行校对。
- c) 辅助校对：使用语义智能机器校对、字形自动比对、人工易错字/词校对、OCR 易错字/词校对、地名/人名/词语校对等技术对文字校对结果进行辅助校验。

7 文字采集

7.1 字库编码方式

中文古籍字库宜采用GB 18030字符集，编码方式宜采用UTF-8。

7.2 集外字处理

7.2.1 替换

7.2.1.1 符号替换

在OCR或文字录入过程中，将集外字直接替换为“口”、“.”等特定符号。

7.2.1.2 图形替换

将集外字替换为图形，图形数据保留集外字的字形特征并添加必要的描述信息。

7.2.1.3 集内字替换

集外字替换为集内字，应满足2个基本条件：

- a) 集内字应与集外字的读音、含义和用法均相同；
- b) 对文字字形没有严格要求，允许对异体字、避讳字、讹误字等进行规范。

7.2.2 造字

7.2.2.1 集外字造字处理说明

在字符集的自定义区为集外字定义编码，编码与集外字的字形一一对应。

7.2.2.2 编码原则

集外字编码应遵循以下原则：

- a) 一致性：中间编码应符合 ISO/IEC 10646:2020 的规定，其标准编码部分同 ISO/IEC 10646:2020 的编码一致；
- b) 充足性：应保证中间编码的编码空间充足；
- c) 适应性：编码方案应方便文字的显示、存储、传输和交换等，应满足现有软件（如 XML 解析器、数据库等）对文档的要求；
- d) 可靠性：编码方案应可靠，避免因系统支持不充分而造成显示错误。

7.2.2.3 编码方式

集外字的编码方式宜采用以下方式：

- a) 采用 UTF-16 编码方案，前 16 个平面采用 UTF-16 编码方案，第 16 平面保留为扩充用途；
- b) 从第 3 平面开始，每个平面的最后 64 个字符保留不做编码使用；
- c) 新增中间编码的第一个字符从 0x30000 开始；
- d) 扩充方案采用 ISO/IEC 10646 中的 IVS 表示机制，即用已编码的字符加上 IVS 字符来表示一个扩充字符。扩充字符由两个代码组成的编码字符表示，第一个代码为第 16 平面的编码，第二个代码为 IVS 编码。扩充编码的顺序，第 16 个平面附加第 1 个 IVS 形成第 17 个平面，第 16 个平面附加第 2 个 IVS 形成第 18 个平面，以此类推。

7.3 标签定义

7.3.1 字符文字

英文标签：<Char>

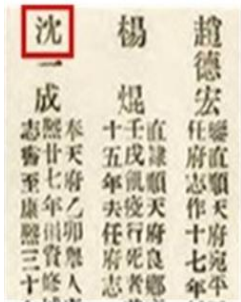
说明：古籍中的文字，如印刷体文字和手抄本文字。

属性：字符文字属性见表1。

表1 字符文字属性

中文名称	英文名称	说明	类型	必备性	可重复性
编码	ID	字符的页内 ID； 同一页面内各字符的 ID 唯一	Int	必备	不可重复
字位序号	No	字符在行中的序号，起始值设为 1	Int	必备	不可重复
位置	Position	字符在页面的坐标位置，使用矩形方式表示时， 其参数为“left, top, right”	Int	必备	不可重复
扩展属性	Extension	补充未定义属性	String	有则必备	不可重复

字符文字示例见图1。



注：右上角的红框所表示的就是一个字符文字，字体字号应基于古籍原版式。

图1 字符文字

7.3.2 空白

- 英文标签：<Blank>
说 明：古籍正文中空白的地方。
属 性：空白属性见表2。
子 元 素：残缺文字<Incomplete_char>
模糊文字<Blurred_char>

表2 空白属性

中文名称	英文名称	说明	类型	必备性	可重复性
编码	Id	空白的页内 ID； 同一页面内各空白的 ID 唯一	Int	必备	不可重复
位置	Position	空白在页面的坐标位置，使用矩形方式表示时，其 参数为“left,top,right”	Int	必备	不可重复
扩展属性	Extension	补充未定义属性	String	有则必备	不可重复

- 采集空白的数据规格，应遵循：
- a) 字间的空白按一个字符处理，描述为一个全角空格（UNICODE 编码为 U+3000）；
 - b) 行首空白按缩进处理；
 - c) 空白行与段末的空白不采集。
- 空白示例见图2。

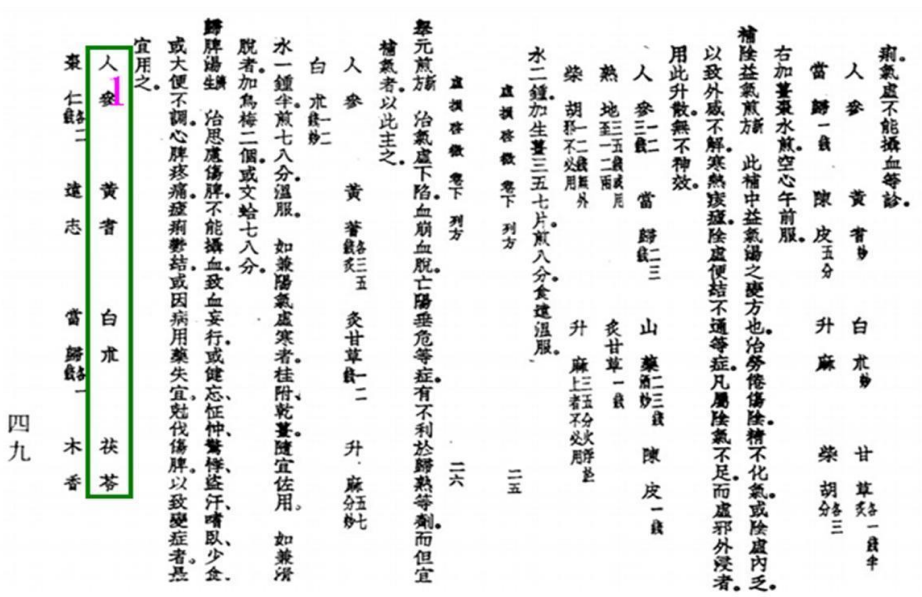


图2 空白示例

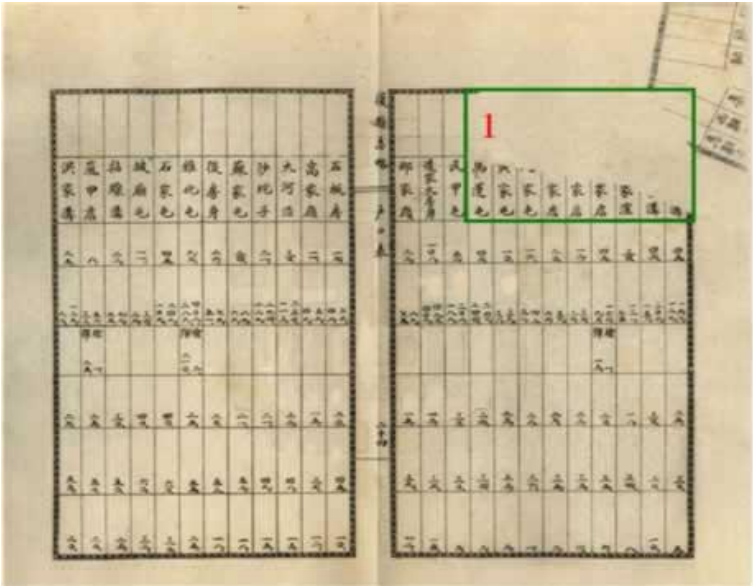
7.3.3 残缺文字

英文标签：<Incomplete_char>
说明：古籍中没有显示的文字。
属性：残缺文字属性见表3。

表3 残缺文字属性

中文名称	英文名称	说明	类型	必备性	可重复性
编码	Id	残缺文字在页内的 ID； 同一页面内各残缺文字的 ID 唯一	Int	必备	不可重复
位置	Position	残缺文字在页面的坐标位置	Int	必备	不可重复
扩展属性	Extension	补充未定义属性	String	有则必备	不可重复
数量	Number	页面中残缺文字的数量	Int	必备	不可重复

采集残缺文字的版式，每一个残缺文字均处理为一个空心五角星（“☆”），UNICODE编码为U+2606。
残缺文字示例见图3。



注：编号1为页面残缺部分，包含4个残留文字。

图3 残缺文字示例

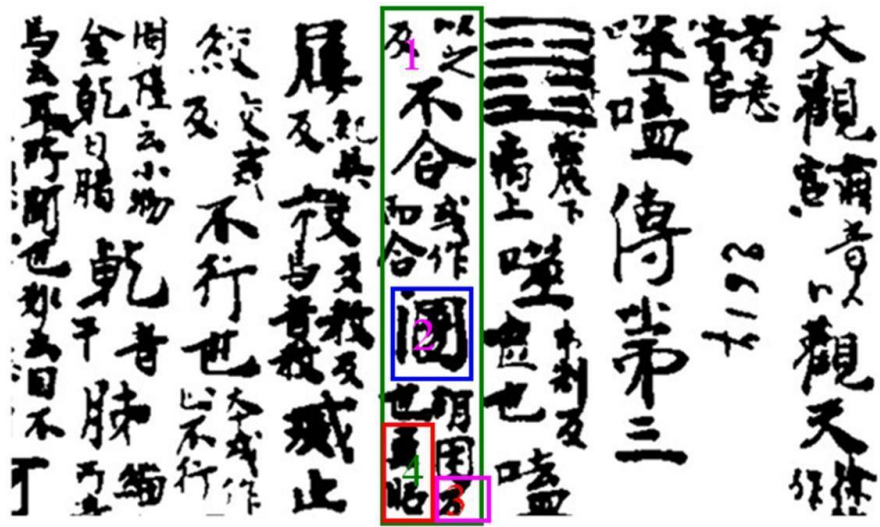
7.3.4 模糊文字

英文标签：<Blurred_char>
说明：古籍中显示不清晰的文字。
属性：模糊文字属性见表4。

表4 模糊文字属性

中文名称	英文名称	说明	类型	必备性	可重复性
编码	Id	模糊文字在页内的 ID； 同一页面内各模糊文字的 ID 唯一	Int	必备	不可重复
位置	Position	模糊文字在页面的坐标位置	Int	必备	不可重复
扩展属性	Extension	补充未定义属性	String	有则必备	不可重复

采集模糊文字的版式，每一个模糊文字均处理为符号“★”（UNICODE编码U+2605）。
模糊文字示例见图4。



注：编号1的部分为一个文字行，包含3个模糊文字（编号2为1个，编号4为2个）和1个残缺文字（编号3）。

图4 模糊文字示例

7.3.5 标点符号

英文标签：<Puncture>
说 明：古籍文中的标点符号。
属 性：无。
不占位标点的页面示例见图5。



图5 不占位标点示例

7.3.6 简洁文字串

英文标签：<Text_simplify>
说 明：古籍中的一串文字。
属 性：简洁文字串属性见表5。

表5 简洁文字串属性

中文名称	英文名称	说明	类型	必备性	可重复性
位置	Position	在页面的坐标位置	Int	必备	不可重复
编号	Id	在页面中的编号，取值不能与 char 的 id 相同	Int	必备	不可重复
类型	Type	文字串的类型	String	必备	不可重复

一个简洁文字串的页面示例见图6。



图6 简洁文字串示例

7.3.7 校勘信息

英文标签: <Text_modify>
说明: 古籍的修正信息。
属性: 校勘信息属性见表6。

表6 校勘信息属性

中文名称	英文名称	说明	类型	必备性	可重复性
位置	Position	在页面的坐标位置	Int	必备	不可重复
编号	Id	在页面中的编号, 取值不能与 char 的 id 相同	Int	必备	不可重复
类型	Type	校勘的类型, 取值“增加”、“删除”、“修改”	String	必备	不可重复
原始文本	Source	文本修改前的内容, 如内容不可识读, 应用“★”代替	String	必备	不可重复

示例: 校勘信息如图 7 中框内所示。

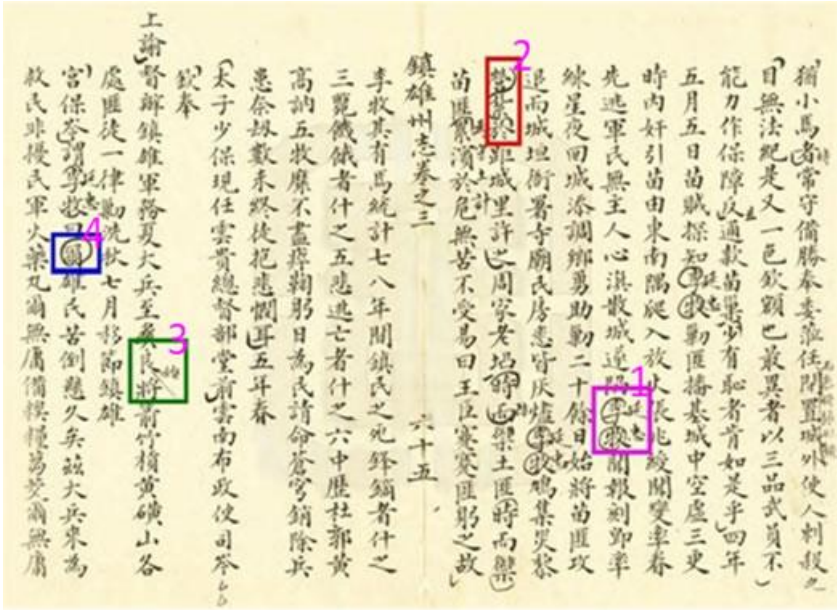


图7 校勘信息示例

7.3.8 文字块

英文标签: <Text>
说明: 定义一个页面的文字内容。
属性: 文字块属性见表7。
子元素: 文字行 <Text_row>

表7 文字块属性

中文名称	英文名称	说明	类型	必备性	可重复性
文字块方向	Direction	横排文字块或纵排文字块	String	必备	可重复

示例: 文字块如图8所示。



图8 文字块示例

7.3.9 文字行

英文标签: <Text_row>
说明: 定义页面一行的内容。
属性: 无
子元素: 无

7.3.10 标注

英文标签: <Chunk_semantic>
说明: 古籍中对正文的标注信息
属性: 标注属性见表8。

子 元 素：批注<Notes>
释音<Phonetic>
释义<Paraphrase>
释形<Form_notes>

表8 标注属性

中文名称	英文名称	说明	类型	必备性	可重复性
位置	Position	标注在页面中的坐标位置	Int	必备	不可重复
字体	Face	标注字体字形	String	必备	可重复
字号	Size	标注字体的大小	Int	必备	可重复

7.3.11 批注

英文标签：<Notes>
说 明：古籍中对正文的批注信息
属 性：批注属性同标注[7.3.10]属性。

7.3.12 释音

英文标签：<Phonetic>
说 明：古籍中对正文的释音信息
属 性：示音属性同标注[7.3.10]属性。

7.3.13 释义

英文标签：<Paraphrase>
说 明：古籍中对正文的释义信息
属 性：释义属性同标注[7.3.10]属性。

7.3.14 释形

英文标签：<Form_notes>
说 明：古籍中对正文的释形信息
属 性：释形属性同标注[7.3.10]属性。

7.4 质量要求

7.4.1 完整性

内容采集的文字应完整，不应出现缺漏和错误。

7.4.2 规范性

内容采集的内容结构化文档采用XML1.0及以上版本，结构化规范描述文件采用XSD1.0及以上版本。

7.4.3 有效性

内容采集后的成品数据文件应能通过相关软件及系统读出，不允许出现数据损坏、异常报错、无法打开等错误。读出的数据应完整，不允许出现编码混乱等无法使用的错误。

7.4.4 准确性

文字准确性评定标准单位为10000个字符，文字差错率小于万分之三，差错率计算公式：
差错率=差错数/检验文字总字符数

7.4.5 文字准确性检测

中文古籍采集文字准确性检测应遵循GB/T 38548.5—2020的规定。

8 图表采集

8.1 标签定义

8.1.1 图像采集

英文标签: <Image>
说 明: 古籍中出现的插图、印章或集外字图等。
属 性: 图像元素的属性见表9。

表9 图像属性

中文名称	英文名称	说明	类型	必备性	可重复性
位置	Position	在页面的坐标位置	Int	必备	不可重复
类型	Type	图像类型如文字图像、表格图像、图片图像	String	必备	可重复
编码	Id	图像在文中的编号	Int	必备	不可重复

切分插图时，应先切出标题、说明等文本，再切出插图。
示例：插图页面示例见图 9，编号 1 和 2 分别为图的标题和注释。

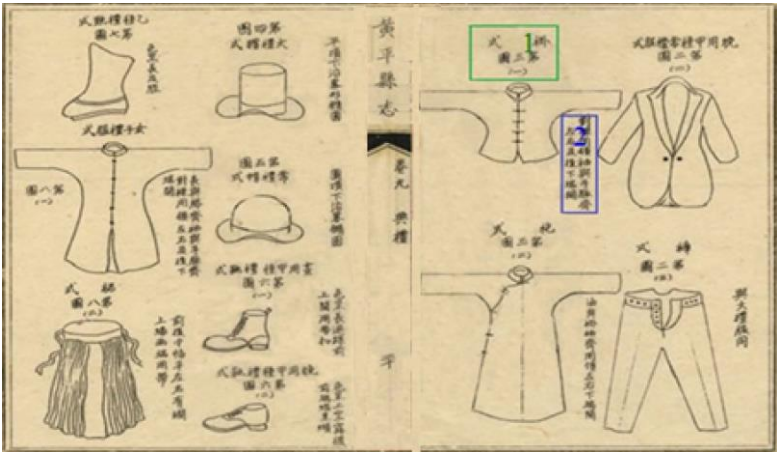


图9 插图页面示例

8.1.2 表格

英文标签: <Table>
说 明: 古籍中的表格。
属 性: 表格属性见表10。
子 元 素: 单元格<Tcell>

表10 表格属性

中文名称	英文名称	说明	类型	必备性	可重复性
表格方向	Direction	表格的纵横方向	String	必备	不可重复
表格行号	Row	行的高度	Int	必备	可重复
表格列号	Col	列的宽度	Int	必备	可重复
表格接续方向	Open	表格左右接续或上下接续	String	可选	不可重复
表格结合点	Join	表格结合的相对位置点	String	可选	不可重复

8.1.3 单元格

英文标签: <tcell>
说 明: 定义表格中的一个单元格。

属性：单元格元素的属性见表11。

表11 单元格属性

中文名称	英文名称	说明	类型	必备性	可重复性
单元格行号	Row	单元格行号	Int	必备	不可重复
单元格列数	Colspan	单元格所在列	Int	可选	可重复
单元格行数	Rowspan	单元格所在行	Int	可选	可重复
单元格接续方向	Continue	单元格左右接续或上下接续	String	必备	可重复

表格内容为文字时，完整描述文字块信息；表格内容为插图、内嵌表格时，将插图、内嵌表格以图像形式存储。

示例1：表格页面示例一如图 10 所示，图中的页面应描述的内容为两个表格和一个文本块（编号 1）：表格部分处理为两个表格。

Figure 10 shows a historical Chinese map page with a grid of place names and distances. A red box highlights a vertical column of text, and a blue box highlights a horizontal row of text. A green box highlights a small text block at the top right.

图10 表格页面示例一

示例2：表格页面示例二如图 11 所示，表格内的插图名称（编号 1）以文本形式标记；编号为 2 的框线标记的部分，以图像形式存储。

Figure 11 shows a historical Chinese page with a grid of illustrations and text. A red box highlights a vertical column of text, and a blue box highlights a horizontal row of text. A green box highlights a small text block at the top right.

图11 表格页面示例二

8.2 质量要求

对图表采集的质量要求应符合CY/T XXX. 4-XXXX中第8章的规定。

9 样式采集

9.1 标签定义

9.1.1 行文顺序

英文标签：<Order>
说 明：古籍文字的排列顺序。
属 性：行文顺序属性见表12。

表12 行文顺序属性

中文名称	英文名称	说明	类型	必备性	可重复性
顺序	Sequence	文字顺序与页面阅读顺序一致， 如为从右到左、从上到下	String	必备	可重复

示例：古籍行文顺序如图 12 所示。



图12 行文顺序示例

9.1.2 提格缩进

英文标签：<Text_row_indent>
说 明：古籍每行的第一个文本与文本区域顶部的空格。
属 性：提格缩进属性见表13。

表13 提格缩进属性

中文名称	英文名称	说明	类型	必备性	可重复性
编号	Number	初值设为 1，按阅读顺序依次编号	Int	必备	不可重复
提格缩进	Indent	indent="0"时，属性不出现。indent<0， 表示提格；indent>0，表示缩进	Int	必备	不可重复

示例：提格缩进示例见图 13，图中，1、2、4、6、8 号框内的行均为缩进行，3、5、7 号框内的数据均为提格行，
剩余两行为顶格行。

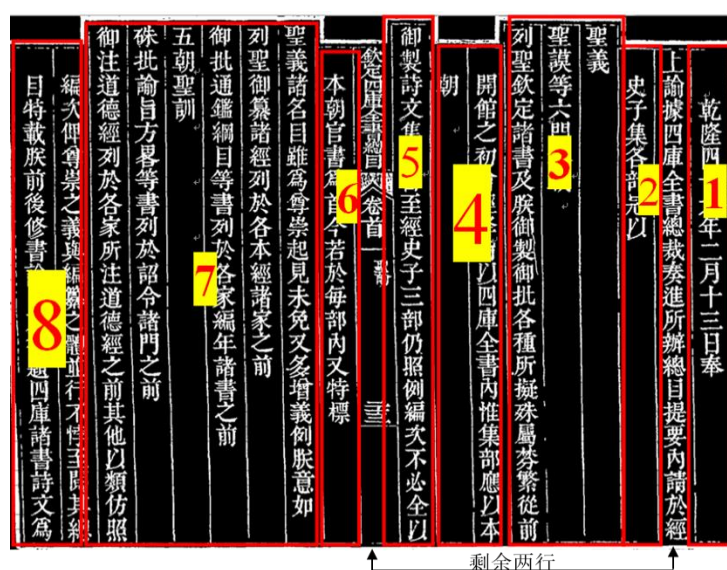


图13 提格缩进示例

9.1.3 段落

英文标签: <Paragraph>

说明：古籍的段落，具有换行标记。

属性：段落属性见表14。

表14 段落属性

中文名称	英文名称	说明	类型	必备性	可重复性
旋转角度	Rotation	文字段的旋转角度，常见值： 0：无旋转； 90：顺时针旋转 90 度； 180：顺时针旋转 180 度； 270：顺时针旋转 270 度； 若属性不出现，默认rotation="0"。	Int	必备	不可重复

示例：段落示例如图 14 所示。

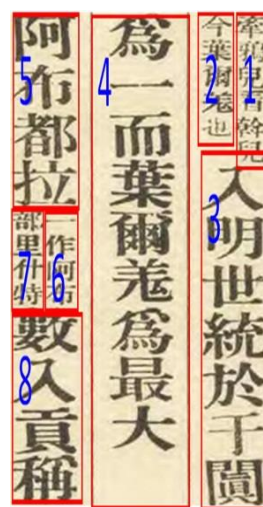


图14 段落示例

9.1.4 段落结束标记

英文标签: <p/>
说 明: 段落结束标记。
属 性: 无。

示例: 段落结束标记示例如图 15 所示。图中, 编号 1 和 2 表示一个段落的结束。

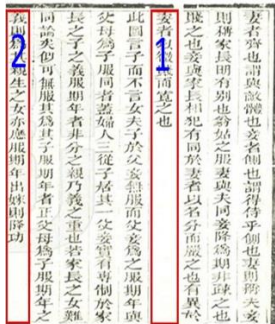


图15 段落结束标记示例

9.2 质量要求

9.2.1 样式采集准确性要求

样式采集准确性评定标准单位为1000个采集图像页, 差错率要求在千分之三以下。

9.2.2 错误统计方法

采集信息与原版内容样式不一致的错误, 每出现一处按1个差错计数, 差错主要为文字、段落格式采集错误: 字体、字号、加粗、倾斜、颜色、底色等信息采集错误。

10 结构采集

10.1 标签定义

10.1.1 章

英文标签: <Chapter>
说 明: 文章的组成部分, 一章又分为若干节。
属 性: 无。
子 元 素: 节<Part>

示例: 古籍章目录示例如图 16 所示。

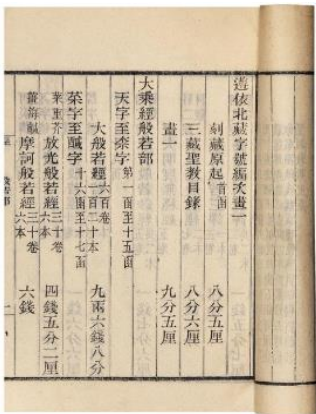


图16 章目录示例

10.1.2 节

英文标签: <Part>
说明: 章的一部分。
属性: 无。
示例: 古籍节目录如图 17 所示。

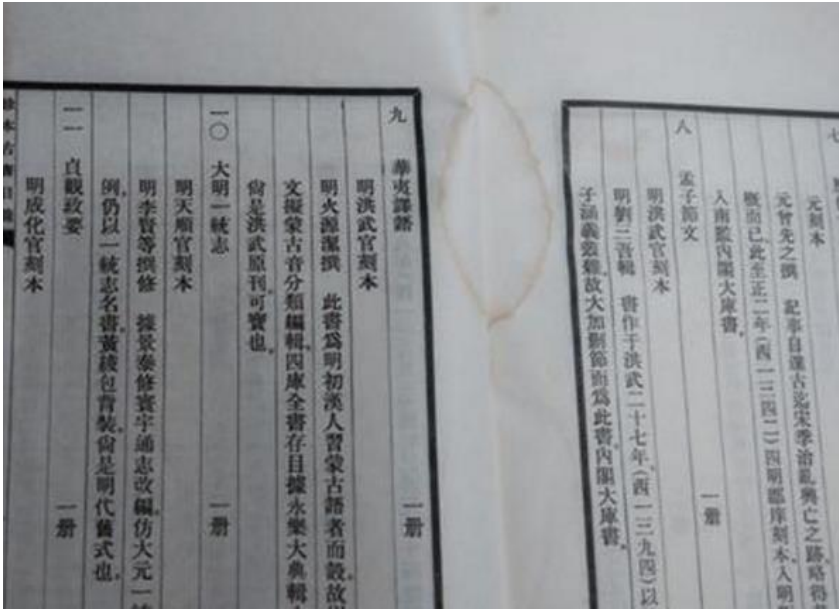


图17 节目录示例

10.1.3 页

英文标签: <Page>
说明: 定义一个页面的完整属性和内容。
属性: 页属性见表15。

表15 页属性

中文名称	英文名称	说明	类型	必备性	可重复性
页标识	ID	页文件的唯一标识符	Int	必备	不可重复
页文件大小	Size	页文件的文件大小	Int	必备	不可重复

10.2 质量要求

10.2.1 结构采集准确性要求

结构采集质量评定标准单位为1000个采集图像页，差错率要求在千分之三以下。

10.2.2 错误统计方法

结构未标引、标引错误、结构化层级错误等每处按1个差错计数。

11 中文古籍内容采集 XML Schema 文件及 XML 样例文件

中文古籍内容采集XML Schema文件见附录A，XML样例文件见附录B。

附 录 A
(规范性)
中文古籍内容采集 XML Schema 文件

```

<?xml version="1.0" encoding="UTF-8"?>
<!-- edited with XMLSpy v2013 (http://www.altova.com) by () -->
<xs:schema                                xmlns:xs="http://www.w3.org/2001/XMLSchema"
xmlns:mml="http://www.w3.org/1998/Math/MathML"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xlink="http://www.w3.org/1999/xlink" elementFormDefault="qualified">
  <xs:element name="Page">
    <xs:complexType>
      <xs:sequence>
        <xs:element name="Char">
          <xs:complexType>
            <xs:attribute name="ID"/>
            <xs:attribute name="No"/>
            <xs:attribute name="Position"/>
            <xs:attribute name="Extension"/>
          </xs:complexType>
        </xs:element>
        <xs:element name="Blank">
          <xs:complexType>
            <xs:sequence maxOccurs="unbounded">
              <xs:element name="Incomplete_char">
                <xs:complexType>
                  <xs:attribute name="Id"/>
                  <xs:attribute name="Position"/>
                  <xs:attribute name="Extension"/>
                  <xs:attribute name="Number"/>
                </xs:complexType>
              </xs:element>
              <xs:element name="Blurred_char">
                <xs:complexType>
                  <xs:attribute name="ID"/>
                  <xs:attribute name="Position"/>
                  <xs:attribute name="Extension"/>
                </xs:complexType>
              </xs:element>
            </xs:sequence>
            <xs:attribute name="ID"/>
            <xs:attribute name="Position"/>
            <xs:attribute name="Extension"/>
          </xs:complexType>
        </xs:element>
        <xs:element name="Puncture"/>
        <xs:element name="Text_simplify">
          <xs:complexType>

```

```

        <xs:attribute name="Position"/>
        <xs:attribute name="ID"/>
        <xs:attribute name="Type"/>
    </xs:complexType>
</xs:element>
<xs:element name="Text_modify">
    <xs:complexType>
        <xs:attribute name="Position"/>
        <xs:attribute name="ID"/>
        <xs:attribute name="Type"/>
        <xs:attribute name="Source"/>
    </xs:complexType>
</xs:element>
<xs:element name="Text">
    <xs:complexType>
        <xs:sequence>
            <xs:element name="Text_row"/>
        </xs:sequence>
        <xs:attribute name="Direction"/>
    </xs:complexType>
</xs:element>
<xs:element name="Chunk_semantic">
    <xs:complexType>
        <xs:sequence maxOccurs="unbounded">
            <xs:element name="Notes"/>
            <xs:element name="Phonetic"/>
            <xs:element name="Paraphrase"/>
            <xs:element name="Form_notes"/>
        </xs:sequence>
        <xs:attribute name="Position"/>
        <xs:attribute name="Face"/>
        <xs:attribute name="Size"/>
    </xs:complexType>
</xs:element>
<xs:element name="Image">
    <xs:complexType>
        <xs:attribute name="Position"/>
        <xs:attribute name="Type"/>
        <xs:attribute name="ID"/>
    </xs:complexType>
</xs:element>
<xs:element name="Table">
    <xs:complexType>
        <xs:sequence maxOccurs="unbounded">
            <xs:element name="Tcell">
                <xs:complexType>
                    <xs:attribute name="Row"/>
                    <xs:attribute name="Colspan"/>
                    <xs:attribute name="Rowspan"/>

```

```

        <xs:attribute name="Continue"/>
      </xs:complexType>
    </xs:element>
    <xs:element name="Image"/>
  </xs:sequence>
  <xs:attribute name="Direction"/>
  <xs:attribute name="Row"/>
  <xs:attribute name="Col"/>
  <xs:attribute name="Open"/>
  <xs:attribute name="Join"/>
</xs:complexType>
</xs:element>
<xs:element name="Order">
  <xs:complexType>
    <xs:attribute name="Sequence"/>
  </xs:complexType>
</xs:element>
<xs:element name="Text_row_indent">
  <xs:complexType>
    <xs:attribute name="Number"/>
    <xs:attribute name="Indent"/>
  </xs:complexType>
</xs:element>
<xs:element name="Paragraph">
  <xs:complexType>
    <xs:attribute name="Rotation"/>
  </xs:complexType>
</xs:element>
<xs:element name="p"/>
<xs:element name="Chapter">
  <xs:complexType>
    <xs:sequence>
      <xs:element name="Part"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="Page">
  <xs:complexType>
    <xs:attribute name="ID"/>
    <xs:attribute name="Size"/>
  </xs:complexType>
</xs:element>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:schema>

```

附 录 B
(资料性)
中文古籍内容采集 XML 样例文件

```

<?xml version="1.0" encoding="UTF-8"?>
<!--Sample XML file generated by XMLSpy v2013 (http://www.altova.com)-->
<Page
xsi:noNamespaceSchemaLocation="%e7%ac%ac5%e9%83%a8%e5%88%86%ef%bc%9a%e5%86%85%e5%ae%b9%e9
%87%87%e9%9b%86%20schema.xsd" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <Char ID="编码" No="字位序号" Position="位置" Extension="扩展属性"/>
    <Blank ID="编码" Position="位置" Extension="扩展属性">
        <Incomplete_char Id="编码" Position="位置" Extension="扩展属性" Number="数量
"/>
        <Blurred_char ID="编码" Position="位置" Extension="扩展属性"/>

    </Blank>
    <Puncture>, </Puncture>
    <Text_simplify Position="位置" ID="编码" Type="类型"/>
    <Text_modify Position="位置" ID="编码" Type="类型" Source="原始文本"/>
    <Text Direction="方向">
        <Text_row>文本行</Text_row>
    </Text>
    <Chunk_semantic Position="位置" Face="字体" Size="字号">
        <Notes>批注</Notes>
        <Phonetic>释音</Phonetic>
        <Paraphrase>释义</Paraphrase>
        <Form_notes>释形</Form_notes>
    </Chunk_semantic>
    <Image Position="位置" Type="类型" ID="编码"/>
    <Table Direction="方向" Row="行号" Col="列号" Open="接续方向" Join="结合点">
        <Tcell Row="行号" Colspan="列数" Rowspan="行数" Continue="接续方向"/>
        <Image>图像</Image>
    </Table>
    <Order Sequence="顺序"/>
    <Text_row_indent Number="编号" Indent="提格缩进"/>
    <Paragraph Rotation="旋转角度"/>
    <p>段落结束标记</p>
    <Chapter>
        <Part>章</Part>
    </Chapter>
    <Page ID="页标识" Size="页文件大小"/>
</Page>

```

参 考 文 献

- [1] CY/T 101.4—2014 新闻出版内容资源加工规范 第4部分：数据加工质量
-