



中华人民共和国新闻出版行业标准

CY/T XXX—XXXX

出版业人工智能语料加工要求

Requirements of artificial intelligence corpus processing in publication

（征求意见稿）

（本稿完成日期：XXXX-XX-XX）

XXXX—XX—XX 发布

XXXX—XX—XX 实施

国家新闻出版署 发布

目 次

前言	V
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
4 总体原则	2
4.1 合规性	2
4.2 安全性	2
4.3 时效性	2
5 语料数据构建	2
5.1 资源采集要求	2
5.2 语料预处理	3
5.2.1 语料清洗	3
5.2.2 语料脱敏	3
5.2.3 其他预处理	3
5.3 语料加工	3
5.3.1 语料切分	3
5.3.2 语料标注	3
5.3.2.1 标注原则	3
5.3.2.2 标注层次	3
5.3.3 语料对齐	4
5.3.3.1 语料对齐基础要求	4
5.3.3.2 基础语言单位对齐	4
5.3.3.3 专业语料对齐	4
5.3.3.4 模态对齐	4
5.3.3.5 版本对齐	4
5.3.4 语料输出	4
5.3.4.1 输出规格	4
5.3.4.2 命名规则	5
5.3.4.3 元数据	5
5.4 语料交付	5
5.4.1 交付内容	5
5.4.2 交付保障	5
5.4.3 交付验收	5
5.4.4 后期维护	5
6 语料质量要求	5

6.1	规范性	6
6.2	完整性	6
6.3	准确性	6
6.4	一致性	6
6.5	及时性	6
6.6	可访问性	6
6.7	稠密性	7
6.8	多样性	7
6.9	均衡性	7
6.10	相关性	7
6.11	原创性	7
6.12	可溯性	7
7	加工保障	8
7.1	加工技术	8
7.1.1	技术环境	8
7.1.2	数据存储	8
7.1.3	数据安全	8
7.1.4	加工工具	8
7.2	加工人员	8
7.3	加工过程	9
7.4	加工安全	9
附录 A	(资料性) 出版物数据标准	10
A.1	语料资源类型	10
A.1.1	资源形态分类	10
A.1.2	数据形态分类	10
A.2	语料资源数据要求	10
A.2.1	纸质出版物标准化成果数据要求	10
A.2.1.1	位图 PDF 数据要求	10
A.2.1.2	矢量图 PDF 数据要求	11
A.2.1.3	代码 PDF 数据要求	11
A.2.1.4	双层 PDF 数据要求	11
A.2.2	非纸质出版物标准化成果数据要求	12
A.2.2.1	音频标准化数据要求	12
A.2.2.2	视频标准化数据要求	12
A.3	语料资源数据结构化处理	13
A.3.1	基础数据结构化处理工作内容	13
A.3.2	基础数据结构化处理成果数据要求	13
A.3.2.1	图书结构化数据要求	13
A.3.2.2	报纸结构化数据要求	13
A.3.2.3	期刊结构化数据要求	14
A.3.3	结构化处理质量要求	14
A.4	数据质量评定方法要求	15

附录 B （资料性） 出版物 XML 数据标签使用规则	16
B.1 图书 XML 标签使用规则	16
B.1.1 图书标签分类	16
B.1.2 元数据标签使用规则	16
B.1.3 结构标签使用规则	24
B.1.4 呈现标签使用规则	28
B.1.5 样式标签使用规则	34
B.1.6 辅助标签使用规则	35
B.2 报纸 XML 标签使用规则	35
B.2.1 报纸标签分类	35
B.2.2 元数据标签使用规则	35
B.2.3 结构标签使用规则	36
B.2.4 呈现标签使用规则	37
B.2.5 样式标签使用规则	44
B.2.6 辅助标签使用规则	46
B.3 期刊 XML 标签使用规则	46
B.3.1 期刊标签分类	46
B.3.2 元数据标签使用规则	46
B.3.3 结构标签使用规则	47
B.3.4 呈现标签使用规则	48
B.3.5 样式标签使用规则	54
B.3.6 辅助标签使用规则	55
附录 C （资料性） 知识资源数据建设标准	56
C.1 知识资源数据建设工作内容	56
C.2 知识资源数据要求	56
C.2.1 关联体系数据要求	56
C.2.1.1 词库要求	56
C.2.1.2 词间关系要求	56
C.2.2 关联关系数据要求	56
C.2.2.1 知识标引位置要求	56
C.2.2.1.1 主题标引位置	56
C.2.2.1.2 扩展标引位置	57
C.2.2.1.3 关联标引位置	57
C.2.2.2 知识标引密度要求	57
C.2.2.2.1 主题标引密度	57
C.2.2.2.2 扩展标引密度	57
C.2.2.2.3 关联标引密度	57
C.3 知识资源数据质量要求	57
C.3.1 关联体系质量要求	57
C.3.1.1 词库质量要求	57
C.3.1.2 词间关系质量要求	57
C.3.2 关联关系质量要求	57

C.4 高价值语料转换	58
C.4.1 基础要求	58
C.4.2 数据输入	58
C.4.3 语料转换流程	59
C.4.3.1 资源筛选	59
C.4.3.2 筛选审核	59
C.4.3.3 人工清洗	59
C.4.3.4 清洗质检	59
C.4.3.5 人工梳理	59
C.4.3.6 语料转换	59
C.4.3.7 转换验证	59
C.4.4 数据输出	59
C.4.4.1 文件格式	59
C.4.4.2 字段定义	59
C.4.4.3 编码	60
C.4.5 质量控制	60
参考文献	61

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（排名不分先后）：广东人民出版社有限公司、暨南大学出版研究院、武汉理工数字传播工程有限公司、北京理工大学出版社有限责任公司、中国大百科全书出版社有限公司、喀什出版社、天津大学出版社有限责任公司、重庆大学电子音像出版社有限公司、扬州大学《实用临床医药杂志》、香港理工大学专业进修学院、港专学院、智荟通（重庆）数智科技有限公司、中图科信数智技术（北京）有限公司、北京今朝视界文化传媒有限公司。

本文件主要起草人：。

出版业人工智能语料加工要求

1 范围

本文件规定了出版业人工智能语料加工的总体原则、语料数据构建、语料质量要求、加工保障等全流程要求。

本文件适用于出版机构、技术提供商及相关单位开展图书、期刊、报纸、音像制品、数字出版物及其他网络出版物等多模态语料的加工活动。

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。其中，注年份的引用文件，仅该年份对应的版本适用于本文件；不注年份的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 35274—2023 数据安全技术 大数据服务安全能力要求

GB/T 38377—2019 新闻出版 知识服务 知识资源建设与服务基础术语

GB/T 38548.2—2020 内容资源数字化加工 第2部分：采集方法

GB/T 38548.3—2020 内容资源数字化加工 第3部分：加工规格

GB/T 40035—2021 双语平行语料加工服务基本要求

GB/T 45257—2025 新闻出版 知识服务 知识元提取与标引

GB/T 45652—2025 网络安全技术 生成式人工智能预训练和优化训练数据安全规范

GB/T 45674—2025 网络安全技术 生成式人工智能数据标注安全规范

ISO 639-3 Codes for the representation of names of languages—Part 3: Alpha-3 code for comprehensive coverage of languages

CY/T XXXX—202X 出版业人工智能应用安全要求

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

ZYF 001—2018 语料库通用技术规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

语料 corpus

为特定应用目的而专门收集的、有一定结构的、有代表性的、可以被计算机程序检索的、具有一定规模的语言材料或语言应用的样本。

[来源：ZYF 001—2018，3.1，有修改]

3.2

粗加工语料 rough processed corpus

经语料清洗或语料预处理后能够进行基本检索和数据提取的语料。

[来源：ZYF 001—2018，3.21，有修改]

3.3

精加工语料 fine processed corpus

根据特定语料库建设目的，采用机器或人工手段进行语料标注的语料。

[来源：ZYP 001—2018，3.22，有修改]

3.4

元数据 metadata

关于语料内容、结构、来源、质量、状况及其他特性的描述性数据。

[来源：GB/T 40035—2021，3.7，有修改]

3.5

知识元 knowledge element

在应用需求下，表达一个完整事物或概念的不必再分的独立的知识单位。

[来源：GB/T 38377—2019，2.3]

4 总体原则

4.1 合规性

应符合以下要求：

- a) 语料内容合法合规；
- b) 涉及个人信息，应取得当事人同意或者符合法律、行政法规规定的其他情形；
- c) 版权合规，不侵害版权权利人合法权益；
- d) 坚持出版业“三审制”，确保语料内容的专业性、准确性和权威性。

4.2 安全性

应严格遵守国家数据安全法律法规及行业规范，建立健全安全保障体系。

4.3 时效性

应建立语料更新与版本控制机制，保障语料的时效性。宜关注信息发展的脉络，如知识的演进过程。

5 语料数据构建

5.1 资源采集要求

资源采集要求如下：

- a) 资源应主要来源于图书、期刊、报纸、音像制品、电子出版物等，以及国家新闻出版主管部门认定的其他类型的数字化作品，游戏除外；
- b) 资源采集应坚持质量优先、价值优先原则；高价值语料建设标准宜参考附录。
- c) 应根据实际需求明确界定采集的时间范围、主题领域及内容模态；应确保语料来源的多样性，覆盖多学科领域、多语言类型、多模态形式及多角度观点；
- d) 应考虑按照应用需求选择与主样本强关联的辅助性、解释性或对抗性内容资源，提升语料多样性。
- e) 资源数据应符合规范性、完整性、准确性、一致性、时效性、可访问性的要求；采集过程应符合 GB/T 38548.2—2020 中的相关要求；
- f) 应同步采集并标注元数据信息，包括但不限于来源、采集时间、内容类型、主题分类、语种、关键词等；
- g) 应对资源的权属状态及时效性进行标注，若授权期限存在变更可能，需建立动态更新机制；
- h) 人工智能语料资源数据应采用具有稳定格式规范及技术标准的数据，版式文件宜采用 PDF 格式。

式，结构化数据及知识资源数据宜采用 XML 格式。

5.2 语料预处理

5.2.1 语料清洗

进行系统性数据清洗，包括：

- a) 对原始语料进行去重、去噪、纠错和格式统一等处理；
- b) 排查价值观、意识形态相关违法不良信息；经核验，若数据源中违法不良信息占比超过 5%，不得用于语料加工。

5.2.2 语料脱敏

对语料数据进行脱敏处理，去除语料中的身份信息和其他敏感信息，把语料转换成匿名化数据。

5.2.3 其他预处理

按需进行其他预处理，如简繁转换等。

5.3 语料加工

5.3.1 语料切分

宜按照实际标注需求将多模态语料（文本、图片、音频、视频等）分割为一系列切分单位，如按句子、段落、场景、图像关键区域、固定图像尺寸、滑动窗口或时间片段等进行分割。

5.3.2 语料标注

5.3.2.1 标注原则

语料标注原则包括：

- a) 可逆性：移除标注后可恢复语料原始状态；
- b) 独立性：标注数据可被单独提取和使用；
- c) 透明性：应公开标注原则和标注符号的意义；
- d) 责任性：应说明标注者所使用的标注方法及原则；
- e) 中立性：应采取普遍接受的中立模式；
- f) 应用性：应符合实际应用需求；
- g) 权威性：应注重依据来源的标注，便于语料应用时对于同一主题或事件的不同描述进行权威性分级。

5.3.2.2 标注层次

语料标注包含但不限于以下层次：

- a) 基本转写层标注：按照一定的层次结构和标注规则，对口语或文本进行分类和注释；
- b) 音位音素标注：对语言的音素、音节等信息进行标注；
- c) 韵律标注：对口语的重音、语调、停顿等特征进行标注；
- d) 语法标注：也叫词性标注，对单词、（汉字）词组等独立最小单元进行标注；
- e) 句法标注：对词或短语在句法结构中的语法信息进行标注；
- f) 语义标注：对词义、句义进行标注；
- g) 语用标注：对任何适用于文本中能够跨越句子边界单元或关系的信息进行标注；

- h) 专业元素标注：对科学图表、工程图纸、数据表格等非文本的专业信息进行标注；
- i) 知识标注：对事实、数值、概念、原理、技能、规则等内容进行标注；
- j) 关系标注：对知识点间的关联关系进行标注；
- k) 语种标注：对于涉及多语种的语料，应依据 ISO 639-3 标准对语料的主要语种及混合语种进行识别与标注；
- l) 文种标注：对同一语种的不同书写系统（如繁简体中文、日文汉字与假名混排）应进行对应的文种代码标注；
- m) 出版元数据标注：对出版内容资源的外部特征及内容特征等元数据信息进行标注；
- n) 版式结构标注：对出版内容资源中的视觉布局元素进行语义化标注。

5.3.3 语料对齐

5.3.3.1 语料对齐基本要求

语料对齐应根据语料类型、应用场景及技术需求选择适配的对齐级别。对齐级别包括但不限于基础语言单位对齐、专业语料对齐、模态对齐、版本对齐等。

5.3.3.2 基础语言单位对齐

- a) 词汇对齐；
- b) 语句对齐；
- c) 段落对齐；
- d) 篇章对齐。

5.3.3.3 专业语料对齐

- a) 术语对齐；
- b) 格式对齐；
- c) 语义对齐。

5.3.3.4 模态对齐

- a) 文本—音频对齐；
- b) 文本—图像对齐；
- c) 文本—视频对齐；
- d) 跨模态元数据对齐。

5.3.3.5 版本对齐

- a) 版次演进对齐；
- b) 题材演变对齐。

5.3.4 语料输出

5.3.4.1 输出规格

粗加工语料成品数据存储规格应符合 GB/T 38548.3—2020 中对文本、图片、公式、表格、音频和视频内容资源的数字化加工规格要求；精加工语料成品数据宜按实际应用需求采用 XML、JSON、JSONL、Markdown、Parquet 等格式保存。

5.3.4.2 命名规则

粗加工语料成品数据文件名应包括版本号、模态类别、来源场景、应用需求类别和采集时间段等信息；精加工语料成品数据文件名宜在粗加工成品数据文件名基础上增加标注信息。

5.3.4.3 元数据

语料元数据宜包括：

- a) 版本基本信息；
- b) 资源采集元数据；
- c) 预处理元数据；
- d) 标注元数据；
- e) 对齐元数据；
- f) 质量控制元数据；
- g) 版权权属信息。

5.4 语料交付

5.4.1 交付内容

输出内容应包括：

- a) 经授权使用的原始数据副本或其明确的可访问路径；
- b) 粗加工或精加工语料成品数据；
- c) 语料元数据及说明文档；
- d) 标注规范；
- e) 质量评价报告；
- f) 数据统计信息；
- g) 版权权属信息。

5.4.2 交付保障

应按照事先约定好的安全递交方式递交交付内容。约束内容宜包括但不限于递交数据的介质、递交数据的途径、工作数据的保存与删除原则、数据安全责任的物理或时间起始点原则等。

5.4.3 交付验收

交付验收要求包括：

- a) 应根据约定的确认验收标准，对数据标注质量进行检查与评价；
- b) 应确定是否通过数据质量验收。

5.4.4 后期维护

交付验收后应提供相关的维护服务。若数据质量未达到预期要求，应对数据进行修正。

6 语料质量要求

应采用规范性、完整性、准确性、一致性、及时性、可访问性、稠密性、多样性、均衡性、相关性、原创性、可溯性 12 类指标对语料成品数据进行质量评价，编制质量评价报告。

6.1 规范性

规范性应包括：

- a) 形式规范性：符合一定数据形式标准的程度，包括国际标准、国家标准、行业标准、行业实践经验或相关规定等；
- b) 隐私规范性：符合法律法规要求、行业标准、企业内部政策等隐私保护相关规范的程度；
- c) 安全规范性：符合人工智能模型训练安全要求，且语料内容严格遵守《出版管理条例》等国家法律法规，不包含任何禁止性内容的程度。

6.2 完整性

完整性应包括：

- a) 规模完整性：满足人工智能模型应用需求最低数量的程度；
- b) 结构完整性：样本所记录信息的完整程度，无非预期的缺失值；
- c) 领域完整性：可支持人工智能任务数量的覆盖程度；
- d) 任务完整性：可支持行业领域数量的覆盖程度。

6.3 准确性

准确性应包括：

- a) 内容真实性：包含信息与其所代表的实际对象、事件或事实一致和真实的程度；
- b) 领域专业性：在特定领域的适用程度、专业深度和准确程度，且应当与特定行业的标准、术语、业务逻辑和专业知识紧密结合；
- c) 标签准确性：标注标签与语料实际内容及标注规范保持一致，无错标、漏标、偏标，标签边界清晰、判定准确。

6.4 一致性

一致性应包括：

- a) 标签一致性：对于精加工语料成品数据，不同标注者之间对相同或相似实例标注保持一致的程度；
- b) 概念一致性：使用的术语、分类标准和定义在语料内部的一致程度。

6.5 及时性

及时性应包括：

- a) 生成及时性：数据生成时间与当前时刻的接近程度；应当尽可能接近，尤其是在需要实时分析和响应的场景中；
- b) 处理及时性：数据处理时间与数据完成处理时间的接近程度；应当尽可能接近开始采集数据的时间；
- c) 发布与更新及时性：更新的频率，以及更新时间与对外发布时间的接近程度；应当尽可能及时更新，以及尽可能及时对外发布；
- d) 时序性：数据集中同一实体的数据元素之间的相对时序逻辑关系的正确程度。

6.6 可访问性

可访问性应包括：

- a) 获取难易性：是否有清晰的使用许可协议，用户是否容易理解这些许可条款，以及获取数据是

否需要特殊的权限或申请流程；

- b) 使用难易性：是否配备详尽的文档说明,包括数据字典、元数据描述(数据来源、收集方法、更新频率等)、使用指南和示例代码等；
- c) 技术支持性：是否提供官方或社区技术支持渠道，用户在使用数据过程中遇到问题时能否得到及时有效的帮助。

6.7 稠密性

稠密性应包括：

- a) 样本唯一性：样本存在重复记录的程度；
- b) 内容信息量：单位样本所涵盖的信息量。

6.8 多样性

多样性应包括：

- a) 特征多样性：包含特征种类的广泛和全面程度，覆盖描述实体不同属性和角度的程度；
- b) 类型多样性：包含类型种类的广泛和全面程度，覆盖不同格式数据的程度，包括结构化、半结构化和非结构化数据；
- c) 来源多样性：数据源头或渠道的广泛和全面程度；
- d) 任务多样性：支持人工智能任务的广泛和全面程度；
- e) 领域多样性：覆盖行业领域的广泛和全面程度；
- f) 文化多样性：包含反映不同社会群体、文化背景、观点和利益相关者多元价值观的广泛和全面程度，在符合社会主义核心价值观及相关法律法规的前提下，尽可能覆盖多元文化视角，避免系统性偏见。

6.9 均衡性

均衡性应包括：

- a) 类别均衡性：各类别数据数量的均衡程度；
- b) 来源均衡性：各来源数据数量的均衡程度。

6.10 相关性

相关性应包括：

- a) 领域相关性：与特定研究领域或应用领域的相关程度；
- b) 任务相关性：与所要完成的人工智能任务的相关程度；
- c) 逻辑相关性：包含上下文相关内容的程度。

6.11 原创性

原创性应包括：

- a) 来源原创性：是否通过专门的实验设计、定制调查、独特观测等方式首次获得；
- b) 内容原创性：内容是前所未有的组合、特定情境下的记录或是独特现象描述的程度。

6.12 可溯性

可溯性应包括：

- a) 来源可溯性：可追踪其最初来源的程度，包括数据是如何被收集的、由谁收集、在何时何地收集等信息；

- b) 链路可溯性：可追踪其从原始形态到最终处理形态整个转换过程的程度；
- c) 版本控制可溯性：可追踪其更新和迭代过程中每个版本记录的程度，包括变化历史、修改原因和责任人等。

7 加工保障

7.1 加工技术

7.1.1 技术环境

技术环境应符合以下要求：

- a) 硬件环境应满足语料加工的计算资源需求，包括处理器、内存、存储等配置要求；
- b) 软件环境应支持加工工具的正常运行，包括操作系统、依赖库、运行时环境等；
- c) 网络环境应保障数据传输的安全性、稳定性，符合相关网络安全标准要求。

7.1.2 数据存储

数据存储应符合以下要求：

- a) 应建立规范的数据存储目录结构，便于数据管理和检索；
- b) 应采用可靠的存储介质，确保数据的安全性和持久性；
- c) 应建立数据备份机制，定期进行数据备份和恢复验证；
- d) 应对存储数据进行完整性校验，防止数据损坏或丢失。

7.1.3 数据安全

数据安全能力应符合 GB/T 35274—2023 中大数据处理安全能力的相关要求。

其他数据安全相关要求应符合 CY/T XXXX《出版业人工智能应用安全要求》的相关要求。

7.1.4 加工工具

加工工具应符合以下要求：

- a) 应根据加工任务难度、数据处理规模及数据属性特征、数据安全控制层级与方式，合理选择加工工具。
- b) 加工工具应支持多模态语料的处理需求，具备数据导入、预处理、标注、对齐等功能；
- c) 工具应具备日志记录功能，能够追溯加工过程中的操作信息。

7.2 加工人员

语料加工的人员管理应符合以下要求：

- a) 加工人员应熟悉出版行业相关法律法规及标准规范，包括但不限于版权管理、数据安全等要求；
- b) 加工人员应具备相应的专业技能，如自然语言处理、音视频处理、标注规则理解等；
- c) 加工人员应参加岗前能力培训并考核合格，培训内容应包括典型安全风险场景及相关安全问题案例与识别方法等。加工人员应签署保密协议，严格遵守数据安全与隐私保护规定。对于离岗的加工人员或因出现安全风险问题被取消资格的人员，应同时撤销其平台工具和数据的访问权限；
- d) 审核人员应具备语料审核所需的出版专业水平；
- e) 加工方应建立人员能力档案，记录人员承担工作的相关内容，用于进行加工人员能力评估与加工质量追踪。

7.3 加工过程

语料加工过程应建立全流程管理机制，包括：

- a) 编制加工任务说明，制定加工计划，明确各环节的任务分工、时间节点及质量要求；
- b) 根据加工任务中的数据安全描述，确定数据的安全等级，实施相应的安全管理措施；加工过程中的大数据组织管理安全能力及大数据服务风险管理能力，应符合 GB/T 35274—2023 中的相关要求；
- c) 实施过程监控，对加工进度、语料质量进行检查，确保符合本文件及相关标准要求；
- d) 建立问题反馈与处理机制，对加工过程中出现的异常情况及时记录并解决；
- e) 建立风险预警机制，及时识别并通报加工过程中的潜在风险；
- f) 实施其他有助于按时按质完成语料加工任务的管理措施。

7.4 加工安全

语料加工安全要求的通用安全要求、预训练数据处理活动的安全要求、优化训练数据处理活动的安全要求及相应评价方法应符合 GB/T 45652—2025 的相关要求。

语料加工过程的标注工具、标注规则、标注人员、标注核验安全要求及标注安全评价方法应符合 GB/T 45674—2025 的相关要求。

附 录 A
(资料性)
出版物数据标准

A.1 语料资源类型

A.1.1 资源形态分类

用于出版业语料加工的资源，按照其出版物类型应采用图书、期刊、报纸、音像制品、电子出版物等，以及国家新闻出版主管部门认定的其他类型的数字化作品，游戏除外。

A.1.2 数据形态分类

语料资源数据按数据形态宜分为以下类别：

- a) 排版源文件（如 FBD、INDD）；
- b) 版式文件（如 PDF、OFD）；
- c) 标记语言文件（如 XML）；
- d) 富文本文件（如 RTF）；
- e) 音频格式文件（如 MP3、WAV、AAC）；
- f) 视频格式文件（如 MP4、MPEG-4、AVI）；
- g) 图像格式文件（如 JPG、PNG、TIFF）。

资源数据应符合本附录的相关要求。

A.2 语料资源数据要求

A.2.1 纸质出版物标准化成果数据要求

A.2.1.1 位图 PDF 数据要求

位图 PDF 是指纸质文档经扫描仪逐页扫描为图像后封装生成的 PDF，仅含图像层，无独立可检索文字层。位图 PDF 数据应符合以下要求：

- a) 彩色图像采用 24 位真彩色模式，灰度图像采用 8 位灰度模式，黑白图像宜采用二值模式；
- b) 面向长期保存应用的图像成品数据，采用无损压缩 TIFF 文件类型作为存储格式，分辨率应在 300dpi 或 600dpi；
- c) 面向发布应用的图像成品数据，采用有损压缩 JPEG 格式作为文件格式，分辨率应在 100dpi 以上；
- d) 内容应完整，无残破和缺失；若内容不完整，判定为不合格；
- e) 颜色准确，无失真情况；
- f) 图像差错率应低于千分之一；
- g) 无噪点、无阴影、无黑线；若存在噪点、阴影或黑线，判定为不合格；
- h) 无明显倾斜和扭曲；
- i) 元数据著录项目完整，著录信息准确；若著录项目不完整或著录信息不准确，判定为不合格；
- j) 对于集外字或现有字库中无法显示的汉字，用符号“■”表示缺字，并应建立“集外字表”，详细填写该字的描述信息；若集外字未用“■”标示或未建立“集外字表”，判定为不合格；
- k) 书签应完整，包含封面页、版权页、目录页及各章节起始页，内容及书签链接应准确；若书签缺项、内容错误或跳转失效，判定为不合格。

A. 2. 1. 2 矢量图 PDF 数据要求

矢量图PDF是指文档内线条、曲线、几何图形、文字轮廓等元素均采用数学路径参数进行描述，且不包含文本字符编码的PDF格式文件。矢量图PDF数据应符合以下要求：

- a) 内容完整，无跑版或乱序情况；若内容不完整或出现跑版、乱序情况，判定为不合格；
- b) 插图清晰，图片内容完整，颜色准确；
- c) 元数据著录项目完整，著录信息准确；若元数据著录项目缺项或信息不准确，判定为不合格；
- d) 对于集外字或现有字库中无法显示的汉字，用符号“■”表示缺字，并应建立“集外字表”，详细填写该字的描述信息；若集外字未以“■”标示或未建立集外字表，判定为不合格；
- e) 书签应完整，包含封面页、版权页、目录页及各章节起始页，内容及书签链接应准确；若书签缺项、内容错误或跳转失效，判定为不合格；
- f) PDF 数据中矢量化对象数据质量的差错率应低于万分之一。

A. 2. 1. 3 代码 PDF 数据要求

代码 PDF 是文档内线条、曲线、几何图形、文字轮廓等元素均采用数学路径矢量参数进行描述，包含可选取、可编辑、可检索文字层的 PDF 文件。代码 PDF 数据应符合以下要求：

- a) 修饰性图片、艺术字、底纹、线条、图表、公式等应以矢量图形形式嵌入，矢量化对象数据质量的差错率应低于万分之一；
- b) 文本应采用字符编码（如 UTF-8）存储，确保文本可被检索、复制和提取；
- c) 内容完整，无跑版或乱序情况；若内容不完整或出现跑版、乱序情况，判定为不合格；
- d) 图像位置、尺寸与原版一致，无颜色失真，无倾斜；若图像位置、尺寸与原版不一致或出现颜色失真、图像倾斜，判定为不合格；
- e) 文字准确度满足使用需求，字符顺序准确，字体完全或以嵌入子集方式嵌入，文字差错率应低于万分之三；
- f) 元数据著录项目完整，著录信息准确；若元数据著录项目缺项或信息不准确，判定为不合格；
- g) 对于集外字或现有字库中无法显示的汉字，用符号“■”表示缺字，并应建立“集外字表”，详细填写该字的描述信息；若集外字未以“■”标示或未建立“集外字表”，判定为不合格；
- h) 书签应完整，包含封面页、版权页、目录页及各章节起始页，内容及书签链接应准确；若书签缺项、内容错误或跳转失效，判定为不合格。

A. 2. 1. 4 双层 PDF 数据要求

双层 PDF 是包含图像层与文字层，文字采用标准字符编码存储，支持文本选取、复制、检索的 PDF 格式文件。双层 PDF 数据应符合以下要求：

- a) PDF 文件与实体书内容保持一致，无缺页、重页，页码顺序颠倒等情况；
- b) PDF 文件中文字编码正确，支持检索和复制；
- c) PDF 文件需要嵌入其引用到的所有字体，避免乱码；若 PDF 文件未嵌入其引用到的所有字体导致乱码，判定为不合格；
- d) 图像层和文字层的内容应对位准确；
- e) 彩色图像采用 24 位真彩色模式，灰度图像采用 8 位灰度模式，黑白图像宜采用二值模式；
- f) 面向长期保存应用的图像成品数据，采用无损压缩 TIFF 文件类型作为存储格式，分辨率应在 300dpi 或 600dpi；当图像分辨率为 300dpi 时，图像层和文字层对位偏差不应超过 12 个像素；当图像分辨率为 600dpi 及以上时，两者对位偏差不应超过 24 个像素；
- g) 面向发布应用的图像成品数据，采用有损压缩 JPEG 格式作为文件格式，分辨率应在 100dpi

以上；

- h) 内容完整，无残破和缺失；若内容不完整，判定为不合格；
- i) 颜色准确，无失真情况；
- j) 无噪点、无阴影、无黑线；若存在噪点、阴影或黑线，判定为不合格；
- k) 无明显倾斜和扭曲；
- l) 元数据著录项目完整，著录信息准确；若著录项目不完整或著录信息不准确，判定为不合格；
- m) 对于集外字或现有字库中无法显示的汉字，用符号“■”表示缺字，并应建立“集外字表”，详细填写该字的描述信息；若集外字未用“■”标示或未建立“集外字表”，判定为不合格；
- n) 书签应完整，包含封面页、版权页、目录页及各章节起始页，内容及书签链接应准确；若书签缺项、内容错误或跳转失效，判定为不合格；
- o) 文字差错率应低于万分之三，图像差错率应低于千分之一。

A. 2. 2 非纸质出版物标准化成果数据要求

A. 2. 2. 1 音频标准化数据要求

数字音频是用二进制编码数据表示音频信号的比特序列。音频标准化数据是经过加工后，符合质量要求，且以数字形态存在的声音文件。

音频标准化数据应符合以下要求：

- a) 内容完整，无缺失，且不存在多余的数据。
- b) 音量适中，声音清晰且无其他干扰杂音；应从音频的失真、噪声、音质及动态应用方面对其进行质量评价，评价等级应符合 CY/T 168—2019 中的相关要求；
- c) 面向长期保存级应用方向的音频标准化数据，采样率应不低于 22.05kHz，量化级不低于 24bit，通道数采用多声道、双声道或单声道，具体由原始资料特性决定。常用格式（压缩方式）为 BWF、WAV 及 APE；
- d) 面向发布服务级应用方向的音频标准化数据，采样率应不低于 22.05kHz，量化级不低于 16bit，通道数采用双声道或单声道。常用格式（压缩方式）为 MP3、AAC、WMA 及 AIFF；
- e) 可在播放器中正常打开、播放。

在音视频标准化数据质量检验中，有一项不合格即判定整个音视频标准化数据不合格。

A. 2. 2. 2 视频标准化数据要求

数字视频是以数字信息记录的视频资料。视频标准化数据是经过加工后，符合质量要求，且以数字化形态存在的视频文件。

视频标准化数据应符合以下要求：

- a) 内容完整、清晰、连贯，无缺失及丢帧情况，画面无抖动、停滞、模糊等，与原始内容的色差在合理范围内；
- b) 音量适中，声音清晰且无其他干扰杂音，音量电平应在 -20dB~0dB 之间，噪声或失真应不超过原有记录；
- c) 音频与视频内容对应准确，无明显偏差；
- d) 面向长期保存级应用方向的视频标准化数据，分辨率不低于 720×576，帧速不低于 25 帧/s 或 30 帧/s，视频速率不低于 1152kbit/s，音频速率不低于 224kbit/s，常用格式（编码方式）为 MPEG-2、AVI；
- e) 面向发布服务级应用方向的视频标准化数据，分辨率不低于 352×288，帧速不低于 15 帧/s 或 20 帧/s，视频速率不低于 640 kbit/s，音频速率不低于 128 kbit/s，常用格式（编码方式）为 MPEG-4、

MP4/WMV/FLV/RM;

f) 可在播放器中正常打开、播放。

视频标准化数据的检验样本覆盖率应不低于 10%，抽样不合格率应不高于万分之三。

A.3 语料资源数据结构化处理

A.3.1 基础数据结构化处理工作内容

对标准化处理成品数据（如图书、报纸、期刊、音视频文件等），进行内容资源的结构拆分、标引、重组与聚合，将其转化为计算机可读格式（宜采用XML格式），以形成新的结构化内容体系，从而灵活支持多种检索与知识组织应用，满足长期保存、编辑与格式转换需求。

A.3.2 基础数据结构化处理成果数据要求

纸质出版物基础数据结构化处理应包含图书、报纸、期刊的基础数据结构化处理。

A.3.2.1 图书结构化数据要求

图书结构化数据应符合以下要求：

- a) 数据组成结构应包含元数据集、内容数据及关联的对象数据；
- b) 标签元素表达及编码方式：
 - 1) 不同位置应采用不同元素名；
 - 2) 元素标记应不以父元素或祖先元素路径为前缀；
 - 3) 编码方式宜采用 UTF-8。
- c) 章节层级结构应与图书原版式保持一致。各级标题的标引层级应与图书目次及正文实际层级一一对应，不应出现层级缺失、错位或跳跃；
- d) 内容关联信息应包含以下链接或引用关系：
 - 1) 目次与正文章节的链接关系；
 - 2) 脚注引用点与脚注的引用关系；
 - 3) 插图引用点与插图的引用关系；
 - 4) 表格引用点与表格的引用关系；
 - 5) 参考文献引用点与参考文献的引用关系；
 - 6) 图像引用点与图像文件的链接关系。
- e) 关联关系标引应准确，正文引用点与目标内容之间应建立唯一映射关系，命名应一一对应且具有唯一性；
- f) 图像应逐幅标引，插图、表格图、公式图、补字图（生僻字及特殊符号图片）等不同类型的图像，应采用不同的标签进行区分标引。图像应嵌入正文对应位置，确保阅读顺序正确。图像格式宜为 JPEG，分辨率宜为 300dpi，像素及原始比例应保持不变；
- g) 内容样式与格式宜与图书版式、语义保持一致。

A.3.2.2 报纸结构化数据要求

报纸结构化数据应符合以下要求：

- a) 数据组成结构应包含元数据集、内容数据及关联的对象数据；
- b) 标签元素表达及编码方式：
 - 1) 不同位置应采用不同元素名；
 - 2) 元素标记应不以父元素或祖先元素路径为前缀；

- 3) 编码方式宜采用 UTF-8。
- c) 章节层级结构应与报纸原版式保持一致。各级标题的标引层级应与报纸版面导航及正文实际层级一一对应，不应出现层级缺失、错位或跳跃；
- d) 内容关联信息应包含以下链接或引用关系：
 - 1) 导读与新闻内容的链接关系；
 - 2) “上接”和“下转”的链接关系；
 - 3) 插图引用点与插图的引用关系；
 - 4) 表格引用点与表格的引用关系；
 - 5) 图像引用点与图像文件的链接关系。
- e) 关联关系标引应准确，正文引用点与目标内容之间应建立唯一映射关系，命名应一一对应且具有唯一性；
- f) 图像应逐幅标引，插图、表格图、公式图、补字图（生僻字及特殊符号图片）等不同类型的图像，应采用不同的标签进行区分标引。图像应嵌入正文对应位置，确保阅读顺序正确。图像格式宜为 JPEG，分辨率宜为 300 dpi，像素及原始比例应保持不变；
- g) 内容样式与格式宜与报纸版式、语义保持一致。

A. 3. 2. 3 期刊结构化数据要求

期刊结构化数据应符合以下要求：

- a) 数据组成结构应包含元数据集、内容数据及关联的对象数据；
- b) 标签元素表达及编码方式：
 - 1) 不同位置应采用不同元素名；
 - 2) 元素标记应不以父元素或祖先元素路径为前缀；
 - 3) 编码方式宜采用 UTF-8。
- c) 章节层级结构应与期刊原版式保持一致。各级标题的标引层级应与期刊目次及正文实际层级一一对应，不应出现层级缺失、错位或跳跃；
- d) 内容关联信息应包含以下链接或引用关系：
 - 1) 脚注引用点与脚注的引用关系；
 - 2) 插图引用点与插图的引用关系；
 - 3) 表格引用点与表格的引用关系；
 - 4) 引文引用点与引文的引用关系；
 - 5) 图像引用点与图像文件的链接关系。
- e) 关联关系标引应准确，正文引用点与目标内容之间应建立唯一映射关系，命名应一一对应且具有唯一性；
- f) 图像应逐幅标引，插图、表格图、公式图、补字图（生僻字及特殊符号图片）等不同类型的图像，应采用不同的标签进行区分标引。图像应嵌入正文对应位置，确保阅读顺序正确。图像格式宜为 JPEG，分辨率宜为 300dpi，像素及原始比例应保持不变；
- g) 内容样式与格式宜与期刊版式、语义保持一致。

A. 3. 3 结构化处理质量要求

纸质出版物结构化（包括图书、报纸、期刊）的数据质量应符合以下要求：

- a) 结构化数据的文字差错率应低于万分之三；
- b) 章节层级与目次、正文实际层级不一致，或存在层级缺失、错位、跳跃的，判定为不合格。
- c) 引用点与目标内容之间未建立唯一映射关系，或命名不具有唯一性的，判定为不合格。

- d) 图像未逐幅标引，判定为不合格；
- e) 图像未嵌入正文对应位置，导致阅读顺序错误，判定为不合格；
- f) 图像倾斜、变形或色彩失真，判定为不合格；
- g) 图像分辨率未达到要求，判定为不合格；
- h) 图像像素或原始比例被改变，判定为不合格；

A.4 数据质量评定方法要求

出版物语料资源的标准化数据及结构化数据中相关元素的差错率计算方法应符合 CY/T 114—2015 和 CY/T 101.4—2014 中的相关要求。

附录 B

(资料性)

出版物 XML 数据标签使用规则

XML 数据标签应按照图书、报纸、期刊分别设计使用规则。

B.1 图书 XML 标签使用规则

B.1.1 图书标签分类

图书标签按用途分为元数据标签、结构标签、呈现标签、样式标签和辅助标签。

B.1.2 元数据标签使用规则

用于标引图书版权元数据信息。根标签为<版权元数据>或<copyright-meta>。

表 B.1 图书元数据标签使用规则

序号	标签中文名称	标签英文名称	标签释义	标签用法 (示例)
1	分类号	classification-code	标引图书内容的学科属性或其他特征,并用于检索图书的分类代码。注:图书在版编目数据以《中国图书馆分类法》作为标引依据。	<版权元数据> ...<分类号>I218.65</分类号>... </版权元数据>
2	标识符	identifier	数字资源在一定体系中的唯一标识。	<版权元数据> ...<标识符>000013020230001</标识符>... </版权元数据>
3	目录名	directory-name	图书文件存放的目录名称,由书名、书号、分册信息、版次信息组成	<版权元数据> ...<目录名>当代岭南文化名家_97872181273300000_ (1-1)_第1版</目录名>... </版权元数据>
4	题名	title	赋予图书内容的正式名称。正题名以及题名说明文字著录于此。	<版权元数据> ...<题名>当代岭南文化名家</题名>.. </版权元数据>
5	作者	author	创作作品的自然人,也可以是法人或其他组织。	<版权元数据> ...<作者>梁某某,王某某</作者>... </版权元数据>
6	作者简介	author-biography	对主要责任者(作者)的生平背景、学术成就、代表作品、专业领域等的简要介绍,用于辅助读者了解创作主体背景。	<版权元数据> ...<作者简介>梁某某,19××年×月生于广东省××县××镇,著名某剧表演艺术家.....</作者简介>... </版权元数据>
7	ISBN	isbn	国际标准书号。一个出版单位出版的一部作品的一个版本的唯一识别代码。以 ISBN 为标识符,包含一位校验码在内的 13 位数字。	<版权元数据>... <ISBN>978-7-218-12733-0</ISBN>... </版权元数据>
8	统一书号	china-unified-book-number	全国统一书号的简称,是我国 1956 年启用、ISBN 体系推行后仍在限定范围使用的法定图书编号,与 ISBN 为两套完全独立的标识体系,专用于不适用 ISBN 的出版物,编号结构为“出版社代号·图书分类号·出版序号”。	<版权元数据> ...<统一书号>15114·12342</统一书号>... </版权元数据>
9	其他书号	other-book-number	用于标识未采用国际标准书号 (ISBN) 或中国统一书号的内部发行、内部使用图书的编号,适用于内部资料、内部交流用书等非公开出版的图书。	<版权元数据> ...<其他书号>R323</其他书号>... </版权元数据>
10	分册名	volume-title	一种图书包含若干分册时,这些分册的名称即为分册书名。	<版权元数据> ...<分册名>西方的衰落</分册名>... </版权元数据>
11	分册号	volume-number	多卷书、丛书、分册连续出版的图书中,单册图书的顺序编号,用于区分同一总题名项下的不同分册/分卷资源,实现多册资源的结构化关联。	<版权元数据> ...<分册号>0001</分册号>... </版权元数据>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
12	分册信息	volume-information	与分册号对应的分册信息,用于完整标识分册资源的内容边界与归属。	<版权元数据> ...<分册信息>第一册</分册信息>... </版权元数据>
13	责任编辑	responsible-editor	负责稿件初审、编辑加工和付印样的通读工作,对编辑、设计、排版、校对、印刷等出版环节的质量负有监督责任的人员。	<版权元数据> ...<责任编辑>毛某某</责任编辑>... </版权元数据>
14	美术编辑	art-editor	负责出版物美术设计的编辑人员。	<版权元数据> ...<美术编辑>毛某某</美术编辑>... </版权元数据>
15	技术编辑	copy-editor	负责出版物技术设计的编辑人员。	<版权元数据> ...<技术编辑>毛某某</技术编辑>... </版权元数据>
16	责任校对	responsible-corrector	对一部书稿负有主要校对责任的工作人员。	<版权元数据> ...<责任校对>曾某某</责任校对>... </版权元数据>
17	责任印制	responsible-printing	承担出版物的印刷和制作工作,包括选择印刷厂、监督印刷过程、检查印刷质量等。他们的工作直接影响到出版物的最终呈现效果。	<版权元数据> ...<责任印制>毛某某</责任印制>... </版权元数据>
18	发行者	distributor	从事出版物发行的机构。 注:发行单位包括出版社、发行所、批销中心、书店等。	<版权元数据> ...<发行者>中国人民大学出版社</发行者>... </版权元数据>
19	印张	printed-sheets	印刷用纸的计量单位,一印张等于全张纸的一半(即一面),图书的总印张数可以通过计算书页数和开本大小得出。	<版权元数据> ...<印张>26.25</印张>... </版权元数据>
20	印数	number-of-copies	同一版本出版物印刷的数量。	<版权元数据> ...<印数>10</印数>... </版权元数据>
21	字数	character-count	图书中的文字数量,通常以千字为单位进行统计。	<版权元数据> ...<字数>62.5千字</字数>... </版权元数据>
22	定价	price	由出版者印制在出版物上的价格。	<版权元数据> ...<定价>CNY 10.00</定价>... </版权元数据>
23	出版时间	publication-date	出版物首次正式出版发行的法定日期,是出版物版本认定、版权期限计算的核心时间项。 对于古籍,指与古籍资源本身生命周期中的一个事件相关的时间。此项著录古籍原件书写刻印的年份。年号纪年以中国朝代、年号、纪年的顺序著录。	示例一: <版权元数据> ...<出版时间>清光緒元年(1875)</出版时间>... </版权元数据> 示例二: <版权元数据> ...<出版时间>日期(年号纪年): 宋紹定元年(刻版), 日期(公元纪年): 1228(刻版)</出版时间>... </版权元数据>
24	印刷时间	printing-date	古籍印刷时间(现代出版物印刷时间一般在印次体现),指将古籍资源印制在纸张等介质上的时间。此项说明古籍资源与出版日期不同的印刷时间。著录的印刷日期应当晚于出版日期。	<版权元数据> ...<印刷时间>清光緒六年(1880)</印刷时间>... </版权元数据>
25	版次	edition-info	图书排版的次数。	<版权元数据> ...<版次>2023年10月第1版 </版次>... </版权元数据>
26	印次	impression-number	同一版本的图书印刷的次数。	<版权元数据> ...<印次>2023年11月第1次印刷</印次>... </版权元数据>
27	版权说明	copyright-statement	经作者或版权所有者授权出版的作品,可标注版权符号©,并注明版权所有者的姓名及首次出版年份。排印在版本记录页的上部位置。	<版权元数据> ...<版权说明>©×××出版社 2026年</版权说明>... </版权元数据>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
28	标准号	standard-number	标准化文件的唯一法定身份编号,用于精准标识国家标准(GB)、行业标准(CY或CY/T)、地方标准、团体标准等正式标准文本,编号通常由标准代号+顺序号+发布年份组成,是标准类数字资源检索、归类、版本区分的核心元数据。	<版权元数据> …<标准号>GB/T 1.1-2002</标准号>… </版权元数据>
30	标准发布单位	standard-issuing-authority	发布法定标准文件的主管机构或批准单位,仅适用于标准类出版物的元数据著录,明确标准的法定发布主体。	<版权元数据> …<标准发布单位>国家市场监督管理总局</标准发布单位>… </版权元数据>
31	标准发布时间	standard-issue-date	标准文件正式发布的法定日期,是标准版本效力、生效周期认定的核心时间项,仅适用于标准类出版物的元数据著录。	<版权元数据> …<标准发布时间>2003-01-31</标准发布时间>… </版权元数据>
32	标准实施单位	standard-implementing-authority	负责标准文件的落地执行、监督管理、宣贯实施的主管机构或责任单位,仅适用于标准类出版物的元数据著录。	<版权元数据> …<标准实施单位>×××出版社</标准实施单位>… </版权元数据>
33	团体责任者	corporate-body	对出版物内容创作、编辑、审定承担主要责任的法人/非法人团体机构,是区别于个人作者的集体责任主体。	<版权元数据> …<团体责任者>新华编辑部</团体责任者>… </版权元数据>
34	次要团体责任者	secondary-corporate-creator	对出版物内容承担辅助性、次要性责任的团体机构,不承担核心创作责任。	<版权元数据> …<次要团体责任者>新华编辑部</次要团体责任者>… </版权元数据>
35	资源类型	resource-type	标识数字出版资源的内容形态与产品类型,用于资源分类、检索与结构化管理。	<版权元数据> …<资源类型>代码 PDF</资源类型>… </版权元数据>
36	源文件	source-file	数字出版资源的原始制作母版文件,即用于生成最终发布版本的原生排版工程文件、高清素材母版、原始编校文件等,是资源版本溯源、重制与版权管理的核心载体。	<版权元数据> …<源文件>Q0647 澳门海洋文化的发展与影响.zip</源文件>… </版权元数据>
37	资料来源	source	数字出版资源及其中引用内容、原始素材的来源出处,包含原始出版物、授权机构、采集渠道、版权归属方等信息,用于版权合规核查与内容溯源。	<版权元数据> …<资料来源>广东人民出版社</资料来源>…</版权元数据>
38	资料类型	source-format	标识资源所引用或包含的素材资料的细分类型,用于资源内容的结构化拆解与精细化管理。	<版权元数据> …<资料类型>排版文件(方正)</资料类型>… </版权元数据>
39	文件质量说明	file-quality-desc	对数字资源文件的技术参数、质量等级、合规情况、瑕疵问题等的详细文字描述,是资源质量管控的核心记录项。	<版权元数据> …<文件质量说明>P42 缺图;缺版权页</文件质量说明>… </版权元数据>
40	文件质量	file-quality-level	按照行业标准对数字资源文件设定的标准化质量等级标识,用于资源入库审核、质量管控与分级应用。	<版权元数据> …<文件质量>半成品</文件质量>… </版权元数据>
41	加工记录标识号	processing-record-id	古籍数字资源的加工唯一标识号,表示古籍元数据和对象数据的关联。	<版权元数据> …<题名>資治通鑑綱目: 五十九卷, 首一卷</题名>… …<加工记录标识号>160113020230001</加工记录标识号>… </版权元数据>
42	国家珍贵古籍名录号	national-rare-book-id	古籍在《国家珍贵古籍名录》中被赋予的编号。	<版权元数据> …<国家珍贵古籍名录号>00233</国家珍贵古籍名录号>… </版权元数据>
43	省级珍贵古籍名录号	provincial-rare-book-id	古籍在各地“省级珍贵古籍名录”中被赋予的编号。	<版权元数据> …<省级珍贵古籍名录号>GD-00123</省级珍贵古籍名录号>… </版权元数据>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
44	古籍普查登记号	ancient-book-census-id	古籍在收藏单位《古籍普查登记目录》中被赋予的普查登记编号。	<版权元数据> ...<古籍普查登记号>110000-0101-0008981</古籍普查登记号>... </版权元数据>
45	中国古籍善本书目号	rare-book-catalog-id	古籍在《中国古籍善本书目》中被赋予的编号。	<版权元数据> ...<中国古籍善本书目号>110000-0101-0008981</中国古籍善本书目号>... </版权元数据>
46	书目记录标识号	record-identifier	古籍在该收藏单位书目系统或联机公共检索目录（OPAC）中的系统号。	<版权元数据> ...<书目记录标识号>002839940</书目记录标识号>... </版权元数据>
47	并列题名	parallel-title	出版物另外一个语种的题名。	<版权元数据> ...<并列题名>The universe or the infinitely great and the infinitely little</并列题名>... </版权元数据>
48	其他题名	other-title	所著录古籍或其他文献中所题的与正题名等不同的题名。	<版权元数据> ...<其他题名>漢魏六朝一百三家集</其他题名>... </版权元数据>
49	主要责任者	creator	对创建古籍负主要责任的实体。	<版权元数据> ...<主要责任者>盧某某</主要责任者>... </版权元数据>
50	主要责任者说明	statement-of-responsible-creator	古籍主要责任者、其他责任者所属的朝代或国别。	<版权元数据> ...<主要责任者说明>宋</主要责任者说明>... </版权元数据>
51	主要责任方式	responsibility-of-creator	古籍形成过程中责任者对古籍负有的责任类型。古籍常见责任方式包括：撰、纂、修、著、注、编、辑、译、校、释文、整理等。	<版权元数据> ...<主要责任方式>撰</主要责任方式>... </版权元数据>
52	其他责任者	contributor	对古籍资源的创建有贡献的其他责任实体。其他责任者的实体包括个人或团体、机构。通常用其他责任者的名称来标识这一条目。	<版权元数据> ...<其他责任者>楊鶴</其他责任者>... </版权元数据>
53	其他责任者说明	statement-of-responsible-contributor	古籍其他责任者所属的朝代或国别。	<版权元数据> ...<其他责任者说明>宋</其他责任者说明>... </版权元数据>
54	其他责任方式	responsibility-of-contributor	古籍形成过程中其他责任者对古籍负有的责任类型。古籍常见责任方式包括：撰、纂、修、著、注、编、辑、译、校、释文、整理等。	<版权元数据> ...<其他责任方式>撰</其他责任方式>... </版权元数据>
55	版本类型	edition	古籍因制作方式的不同而产生的不同种类名称。此项著录古籍原物的版本类型，包括稿本、写本、抄本、绘本、刻本、铃印本、活字印本、铅印本、磁版印本、铜版印本、拓本、其他，以及附加说明。	<版权元数据> ...<版本类型>刻本</版本类型>... </版权元数据>
56	出版人	publisher-person	出版物的出版单位法定代表人/主要负责人，对出版物的内容质量、出版合规性承担最终责任。	<版权元数据> ...<出版人>张三</出版人>... </版权元数据>
57	出版者	publisher	从事出版活动的专业机构。 注：出版单位包括报社、期刊社、图书出版社、音像出版社和电子出版物出版社，网络出版单位以及不设立报社、期刊社的报纸编辑部和期刊编辑部。	<版权元数据> ...<出版者>人民出版社</出版者>... </版权元数据>
58	出版者电话	publisher-phone	出版者的联系电话号码。	<版权元数据> ...<出版者电话>020-88888888</出版者电话>... </版权元数据>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
59	出版者电子邮箱	publisher-email	出版者的电子邮箱。	<版权元数据> …<出版者电子邮箱>12345@gdpph.com</出版者电子邮箱>… </版权元数据>
60	出版者网址	publisher-website	出版者的网站地址。	<版权元数据> …<出版者网址>www.gdpph.com</出版者网址>… </版权元数据>
61	出版者地址	publisher-address	出版社的详细地址，可能包括街道、城市、省份或国家等信息。	<版权元数据> …<出版者地址>北京市中关村大街 31 号</出版者地址>… </版权元数据>
62	出版方式	publishing-method	古籍书写或刻印的方式。如“寫稿”“抄寫”“刻版”“刻版印刷”“修版”“修版重印”等。	<版权元数据> …<出版方式>修版</出版方式>… </版权元数据>
63	印刷者	printer	从事印刷、复制活动的专业机构。	<版权元数据> …<印刷者>北京昌联印刷有限公司</印刷者>… </版权元数据>
64	印刷地址	place-of-printing	印刷者使用工具批量制作古籍资源复本的地点。此项说明与出版者地址不同的印制古籍资源的地点。	<版权元数据> …<印刷地址>廣東</印刷地址>… </版权元数据>
65	印刷方式	printing-method	古籍原件印刷的方式。此项著录古籍的印刷方式、色彩、修版、补版、初印、后印、纸张等信息。	<版权元数据> …<印刷方式>公文紙印</印刷方式>… </版权元数据>
66	日期	date	与古籍资源本身生命周期中的一个事件相关的时间。此项著录古籍原件书写刻印的年份。年号纪年以中国朝代、年号、纪年的顺序著录。	<版权元数据> …<日期><年号纪年>日期（年号纪年） ： 宋紹定元年（刻版）</年号纪年><公元纪年>日期（公元纪年）： 1228（刻版）</公元纪年></日期>…</版权元数据>
67	年号纪年	regnal-date	以帝王在位期间的年号为纪年标识，辅以阴阳历纪月纪日的中国古代纪年法。中国封建王朝年号前应加朝代名。可参照文物出版社的《中国历史年代简表》。	<版权元数据> …<日期><年号纪年>日期（年号纪年） ： 宋紹定元年（刻版）</年号纪年>… </版权元数据>
68	公元纪年	western-date	与古籍资源本身生命周期中的一个事件相关的公元纪年时间。此项著录古籍原件书写刻印时间对应的公元纪年年份。	<版权元数据> …<日期><公元纪年>日期（公元纪年） ： 1228（刻版）</公元纪年></日期>… </版权元数据>
69	图表	graphic-note	对古籍资源内容中图像及表格方面的说明。	<版权元数据> …<图表>…</图表>… </版权元数据>
70	装帧形式	binding	将古籍加工为现有物理状态的方法，此项著录古籍原件的装订方式，如线装、经折装、卷轴装、蝴蝶装、包背装等。	<版权元数据> …<装帧形式>线装</装帧形式>… </版权元数据>
71	数量	quantity	古籍的单位计量统计结果。此项著录古籍原件的数量。量词通常用册、函表述。	<版权元数据> …<数量>6 册（1 函）</数量>… </版权元数据>
72	开本尺寸	dimension	出版物单页幅面大小的称谓。对古籍物理载体规格大小的测量记录。	<版权元数据> …<开本尺寸>： 23.4cm×15.3cm</开本尺寸>… </版权元数据>
73	附件	accompanying-material	内容与出版物主体部分有直接联系，在装订或其他组装方式上与主体部分相分离的附属资料。古籍主体以外的附加资料或物品。	<版权元数据> …<附件>附光盘一张</附件>… </版权元数据>
74	版本描述	edition-description	此项完整著录古籍原件的版本信息，包括出版日期、出版者、出版地、印刷日期、印刷者、印刷地、版本类型等。若古籍原件残缺而以其他刻本或抄本配补者，应以括注的形式说明配补之卷次及其版本信息。	<版权元数据> …<版本描述>宋乾道七年（1171）蔡夢弼東塾刻本（卷七至九、一百二十四至一百三十配宋淳熙三年張枅桐川郡齋刻八年耿秉重修本，卷一至六、十九至二十一、二十三至二十六、三十一至三十三、三十九至六十、七十一至八十、九十至一百一十六配宋刻十四行本，卷十至十五、二十七至三十、六十二至七十配另一宋刻本）</版本描述>… </版权元数据>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
75	底本附注	original-version-description	与本资源相关的原作版本的说明，如校刊、影印、刊印古籍时作为依据的底本。抄本所依据的原本也可著录在此。	<版权元数据> …<底本附注>據清康熙十八年刻本影印<底本附注>… </版权元数据>
76	残存附注	imperfect-description	古籍原件实际存/缺卷的内容和数量。使用括号括注具体存/缺的卷次，卷数、卷次用中文表示，不连续的卷次之间用顿号间隔。	<版权元数据> …<残存附注>存四十七卷(樂城集三十五至四十三、四十七至五十，後集全，三集全)</残存附注>… </版权元数据>
77	缺字附注	missing-characters-description	记录和描述现有字库中缺少的文字。	<版权元数据> …<缺字附注>𠄎=[澄(ㄉㄨㄥˋ)](cheng)</缺字附注>… </版权元数据>
78	丛书附注	series-description	此处著录丛书事项，如丛书题名、丛书子目所处丛书内部序列位置等。	<版权元数据> …<丛书附注>知不足齋叢書</丛书附注>… </版权元数据>
79	合订附注	bound-description	说明多种或多册古籍合订的信息。不同版本的合函/合订古籍，每个版本做一条记录。	<版权元数据> …<合订附注>與“續資治通鑑綱目：二十七卷”合訂，40册，本書为册1-30。</合订附注>… </版权元数据>
80	版式	layout	古籍的行款字数、界格、书口、边栏等信息，此项著录古籍原件每叶或每半叶的行数和每行的大小字数，以及书口、版框形式、鱼尾等情况。行数、字数、双行小字字数用中文表示。	<版权元数据> …<版式>四周双栏版框，双花鱼尾，大黑口，朱丝栏，半葉七行，行十五字，小字雙行同，細黑口，左右雙邊。</版式>… </版权元数据>
81	版框尺寸	frame-size	对古籍原件版框的规格大小的测量记录。著录古籍原件版框的高度、宽度尺寸。高度、宽度之间以“×”相连，单位“cm”。	<版权元数据> …<版框尺寸>20.5cm×13.8cm</版框尺寸>… </版权元数据>
82	收藏历史	provenance	古籍的递传源流以及相关内容。此项著录古籍原件的收藏沿革和在流传过程中产生的各种特征。	<版权元数据>…<收藏历史>鈐“一九四九年武強賀孔才捐贈北平圖書館之圖書”印。</收藏历史>… </版权元数据>
83	批校题跋者	inscription-writer	在古籍上书写有关本书品评、考订、记事等文字的责任者。此项著录古籍原件上与本书内容及收藏流传有关的批校题跋的责任者。	<版权元数据> …<批校题跋者>黄某某，顧某某</批校题跋者>… </版权元数据>
84	批校题跋者说明	writer-stat	古籍批校题跋者所属的朝代或国别。	<版权元数据> …<批校题跋者说明>清</批校题跋者说明>… </版权元数据>
85	批校题跋方式	inscription-role	责任者对古籍批校题跋负有的责任类型。古籍常见批校题跋责任方式包括：序、跋等。	<版权元数据> …<批校题跋方式>校并跋</批校题跋方式>… </版权元数据>
86	文物级别	cultural-relics-level	被著录古籍根据珍贵程度所划分的级别。	<版权元数据> …<文物级别>一级古籍</文物级别>… </版权元数据>
87	破损等级	damage-level	被著录古籍根据破损程度所划分的级别。	<版权元数据> …<破损等级>五级破损</破损等级>… </版权元数据>
88	收藏单位	collection-unit	收藏古籍原件的单位名称。	<版权元数据> …<收藏单位>国家图书馆</收藏单位>… </版权元数据>
89	索书号	call-number	古籍收藏单位为了检索和排架的需要而给予每个古籍资源的一个特定号码。	<版权元数据> …<索书号>9527</索书号>… </版权元数据>
90	丛书题名	series-title	在内容和形式上，或者典藏方式上具有一定联系，并具有独立子目题名和总题名的古籍集合。此处为古籍所属的丛书记录的题名名称。	<版权元数据> …<丛书题名>新刊五子書：二十卷</丛书题名>… </版权元数据>
91	丛书编号	series-number	丛书具有的表示次第的文字及编号。	<版权元数据> …<丛书编号>第3辑</丛书编号>… </版权元数据>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
92	分丛书名	subseries-title	一套丛书包括有若干套隶属于它的“分丛书”时，“分丛书”的名称即为该套丛书的分丛书名。注：分丛书可能有从属于丛书名的书名，也可能没有。分丛书可能有编号，也可能没有。	<版权元数据> …<分丛书名>经济学系列</分丛书名>… </版权元数据>
93	分丛书编号	subseries-number	一套丛书包括有若干套隶属于它的“分丛书”时，“分丛书”的编号即为该套丛书的分丛书编号。注：分丛书可能有从属于丛书名的书名，也可能没有。分丛书可能有编号，也可能没有。	<版权元数据> …<分丛书编号>第II编</分丛书编号>… </版权元数据>
94	丛书责任说明	series-statement-of-responsibility	丛书的主编者、编辑者及其责任方式的表述。	<版权元数据> …<丛书责任说明>陈某某主编</丛书责任说明>… </版权元数据>
95	丛书链接	series-link	子目所属丛书记录的加工记录标识号。	<版权元数据> …<丛书链接>000013020230014</丛书链接>… </版权元数据>
96	子目	sub-series	组成古籍丛编的单种古籍，此处为古籍丛书所属的子目题名。	<版权元数据> …<子目>書經注：十二卷；資治通鑑釋文：三十卷；注陸宣公奏議：十五卷；史載之方：二卷；海藏老人陰證略例：一卷；本草衍義：二十卷；東萊呂紫微師友雜誌：一卷；東萊呂紫微雜說：一卷；可書：一卷；東原錄：一卷；地理葬書集注：九卷，附葬書問對一卷；醫經正本書：一卷；人倫大統賦：二卷；乙巳占：十卷；太上老子道德經集解：二卷；夷堅甲志：二十卷，乙志二十卷，丙志二十卷，丁志二十卷。</子目>… </版权元数据>
97	子目链接	sub-series-link	丛书所辖子目记录的加工记录标识号。填写丛书所辖每个子目所对应的加工记录标识号，有多个子目的重复本字段，分别著录。	<版权元数据> …<子目链接>000013020230014</子目链接>… </版权元数据>
98	合订题名	bound-with	与著录对象装订在一起的古籍的题名。此处为与著录古籍原件合订的古籍题名。	<版权元数据> …<合订题名>涪州石魚題名記</合订题名>… </版权元数据>
99	合订链接	bound-with-link	与本部古籍合订的古籍的加工记录标识号。如果有多个合订记录，重复本字段，分别著录。	<版权元数据> …<合订链接>1601130202300020002</合订链接>… </版权元数据>
100	中国分类主题词表	subject-heading-cct	依据《中国分类主题词表》对古籍资源进行标引的规范主题词。	<版权元数据> …<中国分类主题词表>古籍-分类-中国</中国分类主题词表>… </版权元数据>
101	四部分类法	four-division-classification	依照中国传统的四部分类法对古籍资源进行标引的类名。每级分类之间以“汉字空格”间隔。	<版权元数据> …<四部分类法>史部 地理類 雜誌之屬</四部分类法>… </版权元数据>
102	册数	total-number-of-volumes	古籍的册数。	<版权元数据>…<册数>具体册数</册数>… </版权元数据>
103	总叶数	total-number-of-leaves	古籍的总叶数。	<版权元数据> …<总叶数>具体总叶数</总叶数>… </版权元数据>
104	透字	translucent-characters	古籍是否存在透字情况（有/无）。	<版权元数据> …<透字>有</透字>… </版权元数据>
105	夹框	framed	古籍是否存在夹框情况（有/无）。	<版权元数据> …<夹框>有</夹框>… </版权元数据>
106	夹字	inserted-characters	古籍是否存在夹字情况（有/无）。	<版权元数据> …<夹字>有</夹字>… </版权元数据>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
107	褶皱	wrinkles	古籍褶皱的总数量及位置。	<版权元数据> …<褶皱>具体数量及位置</褶皱>… </版权元数据>
108	缺残叶	missing-or-damaged-leaves	古籍缺残叶的总数量及位置。	<版权元数据> …<缺残叶>具体数量及位置</缺残叶>… </版权元数据>
109	重叶	duplicate-leaves	古籍重叶的总数量及位置。	<版权元数据> …<重叶>具体数量及位置</重叶>… </版权元数据>
110	签条	labels	古籍签条的总数量及位置。	<版权元数据> …<签条>具体数量及位置</签条>… </版权元数据>
111	夹纸	inserted-papers	古籍夹纸的总数量及位置。	<版权元数据> …<夹纸>具体数量及位置</夹纸>… </版权元数据>
112	登记人员	registration-staff	古籍登记人员及其所在单位。	<版权元数据> …<登记人员>具体人员及单位</登记人员>… </版权元数据>
113	登记日期	registration-date	古籍登记的日期。	<版权元数据> …<登记日期>具体日期</登记日期>… </版权元数据>
114	刊刻者	woodblock-carver	又称刊工、雕工、镌工，指古代从事雕版印刷时，负责将文字、图像镌刻于书版之上的手工艺人，是古籍刊刻的直接执行者。	<版权元数据> …<刊刻者>张三</刊刻者>… </版权元数据>
115	地名	spatial	古籍资源内容所涉及的地域范围。	<版权元数据> …<地名>岭南地区</地名>… </版权元数据>
116	年代	temporal	古籍资源内容所涉及的时间范围。	<版权元数据> …<年代>清</年代>… </版权元数据>
117	责任者附注	creator-description	责任者的姓名、字号、生平等方面需要说明的情况，相同责任者在古籍中使用的名称与责任者项著录的名称出现差异时，可在此说明。	<版权元数据> …<责任者附注>王某某，原名某某，本书有“原名某某字某某”印。</责任者附注>… </版权元数据>
118	子目附注	sub-series-description	对本资源所包含子目的说明，此处著录没有单独书目记录的从编子目事项，如子目题名、责任者以及子目序列等。	<版权元数据> …<子目附注>1. 四庫全書敘：一卷；2. 姚某某觀書例：一卷；3. 田隴初觀書後例：一卷；4. 四川省城尊經書院記：一卷；</子目附注>… </版权元数据>
119	附录附注	appendix-description	籍正文之后的附加性内容，包括附刻，处著录没有单独书目记录的附录信息。	<版权元数据> …<附录附注>附讀春秋偶筆：一卷</附录附注>… </版权元数据>
120	获得方式	availability	古籍的获得来源、购买价格等，此项著录古籍原物出处的相关事项。	<版权元数据> …<获得方式>1989 年购买于北京中国书店</获得方式>… </版权元数据>
121	其他编号	other-identifier-number	古籍收藏单位给予某个古籍资源除典藏号之外的其他特定编号。	<版权元数据> …<其他编号>9527</其他编号>… </版权元数据>
122	备注	memo	书籍相关的补充说明信息。	<版权元数据> …<备注>须重新登记。</备注>… </版权元数据>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
123	提要	abstract	对出版物主要内容的概述性介绍。包括古籍资源内容、形式的要点。	<p>示例一：<版权元数据> …<提要>随着铁路智能化进一步推动高速铁路的快速发展,列车的节能优化与自动控制技术成为高速铁路智能化的关键。本书基于高速列车运行线路重复性、模型结构不变性等特征。介绍满足正点下列车全局节能的列车运行优化方法,并针对不同情况分别阐述具有自主学习能力的迭代学习运行控制策略,以实现高速列车位移和速度的精确跟踪保证列车运行安全性。</提要>… </版权元数据></p> <p>示例二：<版权元数据> …<提要>白樸代表作,記載唐明皇與楊貴妃愛情故事。唐邊將安祿山戰敗當斬,明皇因喜其趨奉赦免之,令他在宮中侍候。楊貴妃與他有私,楊國忠奏請明皇,調他為漁陽節度使。……正當歡娛之際,安祿山叛變,奪取潼關,直逼長安。文武大臣均無計禦敵,勸明皇逃往四川。龍武將軍陳玄禮護駕至馬嵬坡,六軍不發,殺死楊國忠,又要求殺楊貴妃以謝天下。明皇不得已令貴妃自盡。安祿山叛亂平定,明皇重回長安,終日思念貴妃,於秋雨之夜,聞雨打梧桐之聲,倍極傷情,引為終身之憾。</提要>… </版权元数据></p>
124	语言种类	language	此项著录古籍内容所使用的主体语种,偶尔出现的其他语种不必著录。	<版权元数据> …<语言种类>汉语</语言种类>… </版权元数据>
125	权限	rights	权限管理元素一般包括一个对资源的权限管理的声明,或者是对提供这一信息的服务的参照。权限管理一般包括知识产权(IPR), 版权和其他的产权。	<版权元数据> …<权限>僅限國家圖書館善本特藏閱覽室內閱覽</权限>… </版权元数据>
126	文献类型	type	按著录对象的类型著录来自于受控词表中的值。如果要描述数字资源的文件格式、物理媒体或尺寸规格,应在“格式”元素中著录。	<版权元数据> …<文献类型>汉文古籍</文献类型>… </版权元数据>

B.1.3 结构标签使用规则

结构标签用于标注书籍的逻辑结构组成部分。主要体现电子图书中保留下来的物理结构和组成图书完整内容的分块。可分为装帧、版权页、文前内容、正文内容和文后内容。

表 B.2 图书结构标签使用规则

序号	中文标签名称	英文标签名称	标签释义	标签用法（示例）
1	图书	book	用文字或图画、符号记录知识于纸张等载体,并具有相当篇幅的非连续性出版物。	图书 XML 文件根节点。 <图书> <装帧>…</装帧> <版权页>…</版权页> <文前内容>…</文前内容> … </图书>
2	装帧	binding	在书刊出版前,从工艺、技术和艺术等方面对其形态进行的整体规划。	<装帧> <封面>…</封面> <腰封>…</腰封> … </装帧>

序号	中文标签名称	英文标签名称	标签释义	标签用法（示例）
3	封面	cover	书刊的外层，指图书的封一。	<封面> <书页图 页码="1" 宽度="2480" 高度="3508"> <图片链接>/img/封面.jpg</图片链接> </书页图> </封面>
4	腰封	belt-cover	勒在封面腰部的纸带，可印制与本书相关的文字。	<腰封> <书页图 页码="2" 宽度="2000" 高度="300"> <图片链接>/img/腰封.jpg</图片链接> </书页图> </腰封>
5	前勒口	front-flap	封面或护封在翻口处向里折转的延长部分。	<前勒口> <书页图 页码="5" 宽度="2480" 高度="3508"> <图片链接>/img/前勒口.jpg</图片链接> </书页图> </前勒口>
6	后勒口	back-flap	封底或护封在翻口处向里折转的延长部分。	<后勒口> <书页图 页码="5" 宽度="2480" 高度="3508"> <图片链接>/img/后勒口.jpg</图片链接> </书页图> </后勒口>
7	书脊	spine	书背两侧的凸起部分。	<书脊> <书页图 页码="1" 宽度="200" 高度="3508"> <图片链接>/img/书脊.jpg</图片链接> </书页图> </书脊>
8	封底	back-cover	书刊外表的背面部分。又称“封四”。	<封底> <书页图 页码="200" 宽度="2480" 高度="3508"> <图片链接>/img/封底.jpg</图片链接> </书页图> </封底>
9	版权页	edition-recording-page	提供图书的版权说明、图书在版编目数据和版本记录的书页。	<版权页> <版权元数据>…</版权元数据> <书页图 页码="6" 宽度="2480" 高度="3508"> <图片链接>/img/版权页.jpg</图片链接> </书页图> </版权页>
10	书名页	title-leaf	图书正文之前载有完整书名信息的书页。	<书名页> <书页图 页码="7" 宽度="2480" 高度="3508"> <图片链接>/img/书名页.jpg</图片链接> </书页图> </书名页>
11	附书名页	half-title-leaf	图书正文之前补充的图书信息，包括编委会信息等。	<附书名页> <书页图 页码="8" 宽度="2480" 高度="3508"> <图片链接>/img/附书名页.jpg</图片链接> </书页图> <段落>《潮州文化丛书》编纂委员会</段落> <段落>主 任：何晓军 何广延</段落> </附书名页>
12	牌记	colophon	又称书牌、木记、刊记，指古籍中专门标记刊刻信息的独立版面单元，多位于卷首、卷末或目录之后，常规刊刻朝代、刊刻年月、刊刻机构、刊刻者、坊肆堂号、版权声明等内容，是古籍版本鉴定、刊刻源流考证的核心直接依据。	…刻书题记、刊记、坊肆信息，多在卷首 / 卷末。<牌记> <书页图 页码="1" 宽度="2480" 高度="3508"> <图片链接>/img/牌记.jpg</图片链接> </书页图> 万历二十三年金陵唐氏刊本 </牌记>…
13	序	preface	置于正文之前的有关本书的独立文章，由他人或作者撰写。	<序> <一级标题>序</一级标题> <分段>这是一段文本内容</分段> <插图><图片链接>/img/插图1.jpg</图片链接></插图> … </序>

序号	中文标签名称	英文标签名称	标签释义	标签用法（示例）
14	前言	foreword	置于正文之前，由作者撰写的有关本书的说明文字。	<图书> <前言> <一级标题>前言 </一级标题> <分段>这是一段文本内容</分段> <插图><图片链接>/img/插图 1. jpg</图片链接></插图> ... </前言> </图书>
15	凡例	explanatory- notes	关于图书内容、编纂体例及使用方法的文字说明。一般置于正文或目次之前。	<凡例> <一级标题>凡例</一级标题> <分段>这是一段文本内容</分段> <插图><图片链接>/img/插图 1. jpg</图片链接></插图> ... </凡例>
16	目录	catalog	按一定次序编排以供查考的名目。	<目录> <一级标题>目录</一级标题> <分段>第一章 绪论.....001</分段> <分段>第二章002</分段> ... </目录>
17	卷册目录	volume catalog	古籍卷册目次，按一定次序编排以供查考的名目。	<卷册目录> <一级标题>目录</一级标题> <分段>标题一.....001</分段> <分段>标题二.....005</分段> ... </卷册目录>
18	正文内容	body	图书或文章的主要部分，包含作者想要传达的核心信息、观点、故事或研究数据等。	<正文内容> <一级标题>...</一级标题> <分段>...</分段> <二级标题>...</二级标题> <分段>...</分段>... </正文内容>
19	一级标题	main-title	正文中划分的第一层级标题，在古籍 XML 标引中也可称为部，用于概括和引领该部分正文内容的核心主题。在层级标题体系中，一级标题位于最上方，以下可以有二级、三级等更低层次的标题。	<正文内容> <一级标题>...</一级标题> <分段>...</分段> <二级标题>...</二级标题> <分段>...</分段>... </正文内容>
20	二级标题	secondary-title	正文中划分的第二层级标题，在古籍 XML 标引中也可称为编，用于进一步细分章节或段落的内容。在层级标题体系中，二级标题隶属于一级标题，位于三级标题之上。	<正文内容> <一级标题>...</一级标题> <分段>...</分段> <二级标题>...</二级标题> <分段>...</分段>... </正文内容>
21	三级标题	tertiary-title	正文中划分的第三层级标题，在古籍 XML 标引中也可称为篇、卷，用于进一步细分章节或段落的内容。在层级标题体系中，三级标题隶属于二级标题，位于四级标题之上。	<正文内容> ...<三级标题>...</三级标题> <分段>...</分段>... </正文内容>
22	四级标题	quaternary-title	正文中划分的第四层级标题，在古籍 XML 标引中也可称为章，用于进一步细分三级标题下的内容。在层级标题体系中，四级标题隶属于三级标题，位于五级标题之上。	<正文内容> ...<四级标题>...</四级标题> <分段>...</分段>... </正文内容>
23	五级标题	quinary-title	正文中划分的第五层级标题，在古籍 XML 标引中也可称为节，用于进一步细分四级标题下的内容。在层级标题体系中，五级标题隶属于四级标题，位于六级标题之上。	<正文内容> ...<五级标题>...</五级标题> <分段>...</分段>... </正文内容>

序号	中文标签名称	英文标签名称	标签释义	标签用法（示例）
24	六级标题	senary-title	正文中划分的第六层级标题，在古籍XML标引中也可称为子节，用于进一步细分五级标题下的内容。在层级标题体系中，六级标题隶属于五级标题，位于七级标题之上。	<正文内容> ...<六级标题>...</六级标题> <分段>...</分段>... </正文内容>
25	七级标题	septenary-title	正文中划分的第七层级标题，在古籍XML标引中也可称为子节、回，用于进一步细分六级标题下的内容。在层级标题体系中，七级标题隶属于六级标题，位于八级标题之上。	<正文内容> ...<七级标题>...</七级标题> <分段>...</分段>... </正文内容>
26	八级标题	octonary-title	正文中划分的第八层级标题，用于进一步细分七级标题下的内容。在层级标题体系中，八级标题隶属于七级标题。	<正文内容> ...<八级标题>...</八级标题> <分段>...</分段>... </正文内容>
27	后记	afterword	置于书末的有关本书的说明文字，与序、前言有所呼应和补充。	<后记> <一级标题>后记</一级标题> <分段>...</分段> <插图><图片链接>/img/插图 1. jpg</图片链接></插图> ... </后记>
28	附录	appendix	附在正文后面的有关文章、图片、资料。	<附录> <一级标题>附录一</一级标题> <分段>...</分段> <插图><图片链接>/img/插图 1. jpg</图片链接></插图> <一级标题>附录二</一级标题> <表格图><图片链接>/img/表格图 1. jpg</图片链接></表格图> ... </附录>
29	参考文献	bibliography	在全书正文之后或各部分之后一一列出的参考、引用资料的名单。	<参考文献> <一级标题>参考文献</一级标题> <分段>[序号]作者. 书名[M]. 出版地：出版社，出版年份：起止页码.</分段>... </参考文献>
30	索引	index	汇集书刊中包含的字词、语句、名词、事件、编号等主题，以适当方式编排，指引读者查找的检索工具。	<索引> <一级标题>索引</一级标题> <表格图><图片链接>/img/表格图. jpg</图片链接></表格图> </索引>
31	勘误表	errata	附在书刊中，更正文字错误的表格。	<勘误表> <表格图><图片链接>/img/表格图. jpg</图片链接></表格图> </勘误表>
32	作者简介	author-introduction	对著作者的身份、著述情况的简要介绍。	<作者简介> <分段>...</分段> <插图><图片链接>/img/插图 1. jpg</图片链接></插图> ... </作者简介>
33	校勘记	collation	记录古籍在整理或校对过程中，不同版本之间文字差异（如讹、脱、衍、倒），以及据以改正依据的文字说明。通常附于篇末、卷末或作为独立卷册。	<校勘记><分段>校勘记：諸侯王女曰翁主</分段><分段>“翁主”，原作“公主”，據毛利本、景祐本、紹興本改。按：《漢書》卷一下《高帝紀下》“女子公主”顏師古注引如淳作“翁主”，師古曰：“天子不親主婚，故謂之公主。諸王即自主婚，故其女曰翁主。翁者，父也，言父主其婚也。”</分段></校勘记>

B.1.4 呈现标签使用规则

呈现标签用于标注图书内包含的实质性文字和图片元素。文字和图片是最细粒度的元素，不具备业务逻辑性，所以将它向上抽象，结合业务应用实际和标注难易度，分为图片、段落、公式、注释、诗词等。

表 B.3 图书呈现标签使用规则

序号	中文标签名称	英文标签名称	标签释义	标签用法（示例）
1	注释编号	annotation-number	对图书的某些内容或文字所作的说明的编号。	<分段> 时令已快到惊蛰<注释编号>⑥</注释编号>,雪当然再不会存留,往往还没等落地,就已经消失得无踪无影了。 </分段>
2	注释内容	annotation-content	对图书的某些内容或文字所作的说明。	<注释内容><注释编号>⑥</注释编号>惊蛰,是二十四节气中的第三个节气。斗指丁,太阳到达黄经 345°,于公历 3 月 5—6 日交节。 </注释内容>
3	化学公式	chemical-formula	用元素符号和数字的组合表示物质组成的式子。	<化学公式> $2\text{H}<\text{下标}>2</\text{下标}>\rightarrow 2\text{H}<\text{下标}>2</\text{下标}>+02$ </化学公式>
4	数学公式	mathematical-formula	用数学符号表示几个量之间关系的式子。	<数学公式> $ a-b \geq a - b $ </数学公式>
5	其他公式	other-formula	除了化学和数学公式外的其他公式。	<其他公式> 流通中所需要的货币量 = (待售商品的数量 × 物价水平) / 单位货币流通速度 </其他公式>
6	补字图	glyph	对计算机无法输出的字符用图片形式展示。	<分段> 惊 <补字图 页码="1" 宽度="120" 高度="80"><图片链接>/img/蛰.jpg</图片链接></补字图>,是二十四节气中的第三个节气。斗指丁,太阳到达黄经 345°,于公历 3 月 5—6 日交节。 </分段>
7	表格图	table-fig	表格内容以图片形式展示。用于标注文中出现的表格,已表格图方式进行标注。	<附录> <分段>附录二</分段> <表格图><图片链接>/img/表格图 1.jpg</图片链接></表格图> ... </附录>
8	公式图	graphic-formula	在图形中嵌入数学公式或表达式的图表。对于无法用文字进行标注的公式,用公式图进行标注。	<附录> <分段>附录二</分段> <公式图><图片链接>/img/公式图.jpg</图片链接></公式图> ... </附录>
9	插图	figure	插在书刊文字中间用于说明内容的图画。	<作者简介> <分段>时令已快到惊蛰,雪当然再不会存留,往往还没等落地,就已经消失得无踪无影了。</分段> <插图><图片链接>/img/插图.jpg</图片链接></插图> ... </作者简介>
10	落款图	signature-picture	常见于前言、序等地方,章节内容编写者无法用简单文字表达的落款署名,用于标注文中内容的落款图片。标签中页码属性表示该图出现的页码。	<分段> <落款图 页码="15"><图片链接>imageCut/img00015001.jpg</图片链接></落款图> ... </分段>
11	背景图	background-picture	用于页面、底部的图像,它作为背景存在,不直接参与内容的展示,但可以为整个页面或文档提供视觉上的背景支持或氛围营造。	<分段> <背景图 页码="15" 宽度="1920" 高度="1080"><图片链接>imageCut/img00015001.jpg</图片链接></背景图> ... </分段>

序号	中文标签名称	英文标签名称	标签释义	标签用法（示例）
12	书页图	page-picture	图书中每一页的图像表示。部分页面以图片形式进行整页截图保存时，用书页图标注，例如书名页、版权页等在标注同时会整张页面以图片形式标注一次。	<书名页> <书页图 页码="4" 宽度="2480" 高度="3508"><图片链接>/img/书名页.jpg</图片链接></书页图> </书名页>
13	图题	picture-caption	图片的标题。通常出现在图片下方。	<附录> <分段>附录二</分段> <插图 页码 ="15"> <图片链接>/img/插图.jpg</图片链接> <图题>麦克斯韦方程组</图题> </插图> ... </附录>
14	图注	picture-footnote	图片的注释。用于标注图片补充性注释、图例等，通常出现在图题下方。	<附录> <分段>附录二</分段> <插图 页码 ="15"><图片链接>/img/插图.jpg</图片链接> <图题>麦克斯韦方程组</图题> <图注>使用高斯单位制</图注> </插图> ... </附录>
15	有线表	border-table	具有明显的横竖线条来界定单元格的表格。	<有线表> <表题>苗语句典表</表题> ... </有线表>
16	无线表	table	没有明显的边框线的表格。	<无线表> <表题>苗语句典表</表题> ... </无线表>
17	表题	caption	表格题名。	<有线表> <表题>苗语句典表</表题> ... </有线表>
18	表头	table-head	表格表头，用于标注表格顶部第一行的单元格区域，相当于数据表的字段名称集合。	<有线表> <表题>语言句典表</表题> <表头> <单元格头 行范围="2">条目</单元格头> <单元格头 列范围="3" 行范围="1">普通话</单元格头> <单元格头 行范围="2">苗语</单元格头> <单元格头 行范围="2">分词情况</单元格头> <单元格头 行范围="1">北京话</单元格头> <单元格头 行范围="1">广东话</单元格头> <单元格头 行范围="1">上海话</单元格头> </表头> </有线表>
19	表正文	table-body	表格数据，用于标注表格内容数据。	<有线表> ... <表正文> <行> <单元格>条目 1</单元格> <单元格>第一行第一列</单元格> <单元格>第一行第二列</单元格> <单元格>第一行第三列</单元格> </行> <行> <单元格>条目 2</单元格> <单元格>第二行第一列</单元格> <单元格>第二行第二列</单元格> <单元格>第二行第三列</单元格> </行> </表正文>... </有线表>

序号	中文标签名称	英文标签名称	标签释义	标签用法（示例）
20	表尾	table-foot	表格总结、汇总等，通常出现在表内结尾。	用于标注底部区域，是表格内容结束后的部分，如：汇总、结论、合计等。 <有线表> ... <表尾> <行> <单元格>合计</单元格> <单元格>50.00</单元格> <单元格>70.00</单元格> <单元格>100.00</单元格> </行> </表尾>... </有线表>
21	行	table-row	以行为单位，为表格内的单元格分组。	用于标注表格中单元格隔行。 <有线表> ... <表正文> <行> <单元格>条目 1</单元格> <单元格>第一行第一列</单元格> <单元格>第一行第二列</单元格> <单元格>第一行第三列</单元格> </行> ...</表正文> </有线表>
22	单元格头	table-cell-header	表格内单元格头。	用于标注表格中最小单元格，跨行或跨列的单元格需设置属性（跨行数/列数）。 <有线表> ...<表头> <单元格头 行范围="2">条目</单元格头> <单元格头 列范围="3" 行范围="1">普通话</单元格头> <单元格头 行范围="2">苗语</单元格头> <单元格头 行范围="2">分词情况</单元格头> <单元格头 行范围="1">北京话</单元格头> <单元格头 行范围="1">广东话</单元格头> <单元格头 行范围="1">上海话</单元格头> </表头>... </有线表>
23	单元格	table-cell	表格内单元格。	用于标注表格中最小单元格，跨行或跨列的单元格需设置属性（跨行数/列数）。 <有线表> ... <表正文> <行> <单元格>条目 1</单元格> <单元格>第一行第一列</单元格> <单元格>第一行第二列</单元格> <单元格>第一行第三列</单元格> </行> <行> <单元格>条目 2</单元格> <单元格>第二行第一列</单元格> <单元格>第二行第二列</单元格> <单元格>第二行第三列</单元格> </行> </表正文>... </有线表>
24	表注	table-footnote	表格注释。通常位于表格下方。	<有线表> <表题>苗语句典表</表题> <表正文> </表正文> </有线表> <表注>特指 2025 年数据</表注>

序号	中文标签名称	英文标签名称	标签释义	标签用法（示例）
25	书信	letter	用于交流的具备特殊格式的文本内容。	用于标注书信内容。 <书信> <书信标题>手连手心连心</书信标题> <开头称谓>亲爱的唐冉</开头称谓> <问候语>你好！</问候语> <书信内容> 在这寒假之前，我们学校举行了一次“手拉手”的活动…… </书信内容> <结束语>身体健康，学习进步！</结束语> <书信落款>你的朋友陈研</书信落款> <书信落款>2018年2月26日</书信落款> </书信>
26	书信标题	letter-title	书信的标题。	<书信> <书信标题>手连手心连心</书信标题> <书信称谓>亲爱的唐冉</书信称谓> <问候语>你好！</问候语> <书信内容> 在这寒假之前，我们学校举行了一次“手拉手”的活动…… </书信内容> <结束语>身体健康，学习进步！</结束语> <书信落款>你的朋友陈研</书信落款> <书信落款>2018年2月26日</书信落款> </书信>
27	开头称谓	salutation	书信开头的称谓、提称语。	<书信> <书信标题>手连手心连心</书信标题> <开头称谓>亲爱的唐冉</开头称谓> <问候语>你好！</问候语> …… </书信>
28	问候语	greeting	书信的开头语、启事敬辞。	用于标注书信中内容开始前的问候语。 <书信> …… <开头称谓>亲爱的唐冉</开头称谓> <问候语>你好！</问候语> <书信内容> 在这寒假之前，我们学校举行了一次“手拉手”的活动…… …… </书信内容>… </书信>
29	书信内容	letter-body	书信的主体内容。	<书信> …… <问候语>你好！</问候语> <书信内容> 在这寒假之前，我们学校举行了一次“手拉手”的活动…… </书信内容> <结束语>身体健康，学习进步！</结束语>… </书信>
30	结束语	closing	书信的结束语、祝颂语。	用于标注书信内容结束时的结束语。 <书信> …… <书信内容> 在这寒假之前，我们学校举行了一次“手拉手”的活动…… </书信内容> <结束语>身体健康，学习进步！</结束语> <书信落款>你的朋友陈研</书信落款> <书信落款>2018年2月26日</书信落款> </书信>
31	书信落款	letter-signature	书信的落款署名、时间等。	用于标注书信落款（写信人、时间、地点等）。 <书信> …… <结束语>身体健康，学习进步！</结束语> <书信落款>你的朋友陈研</书信落款> <书信落款>2018年2月26日</书信落款> </书信>
32	诗词	poem	文学创作的具备特殊格式的文本内容。单句出现在正文中则不需要标引为诗词。	<诗词> <诗词标题>短歌行</诗词标题> <诗词作者>曹操</诗词作者> <诗词内容> 对酒当歌，人生几何！ 譬如朝露，去日苦多。 </诗词内容>

序号	中文标签名称	英文标签名称	标签释义	标签用法（示例）
				</诗词>
33	诗词标题	poem-title	诗词的标题。	<诗词> <诗词标题>短歌行</诗词标题>… </诗词>
34	诗词内容	poem-body	诗词的主体内容。	<诗词> … <诗词内容> 对酒当歌，人生几何！ 譬如朝露，去日苦多。 </诗词内容> </诗词>
35	诗词作者	poet	诗词的作者。	<诗词> …<诗词作者>曹操</诗词作者>… </诗词>
36	文字块	text	古籍中无明确层级标题的连续文本块。	<文字块><竖排> <文字行 indent="…" no="…">…… </文字行> ……<文字行 indent="…" no="…"><文字段><字符文字 位置="…" id="…" no="0" extension="…">何</字符文字> ……</文字段></文字行> ……</竖排></文字块>
37	文字行	text-row	古籍中的文字行，由文字段和段落结束标记组成。	用于标注古籍中的文字行，由文字段和段落结束标记组成。 <文字块>… <文字行 indent="…" no="…">……</文字行>…</文字块>
38	文字段	chunk	将行内文本划分为具有特定语言或逻辑意义的文本片段（如一个词组、一句诗或一个校勘单元）。	用于标注将行内文本划分为具有特定语言或逻辑意义的文本片段（如一个词组、一句诗或一个校勘单元）。 <文字块><竖排> <文字行 indent="…" no="…">…… </文字行> ……<文字行 indent="…" no="…"><文字段><字符文字 位置="…" id="…" no="0" extension="…">何</字符文字> ……</文字段></文字行> ……</竖排></文字块>
39	字符文字	char	单个汉字及其字形属性。	用于标注单个汉字及其字形属性。 <文字段><字符文字 位置="…" id="…" no="0" extension="…">何</字符文字> ……</文字段>
40	标点符号	puncture	明确标识文本中的现代标点或句读符号，将其与普通字符文字在结构上区分开来。	用于明确标识文本中的现代标点或句读符号，以将其与普通字符文字在结构上区分开来。 <文字段><字符文字 位置="…" id="…" no="…" extension="…">荔</字符文字><字符文字 位置="…" id="…" no="…" extension="…">枝</字符文字><标点符号>、</标点符号><字符文字 位置="…" id="…" no="…" extension="…">楊</字符文字><字符文字 位置="…" id="…" no="…" extension="…">梅</字符文字> …….</文字段>
41	简洁字符串	text-simplify	分离了校勘信息和标点后的纯净检索文本。	<简洁字符串 位置="…" id="…" type="…">此八論題目十行</简洁字符串>
42	校勘信息	text-modify	不同版本间的文字差异、补遗或校对注释。	<校勘信息 位置="…" id="…" type="修改（modify）" source="祐">佑</校勘信息>
43	牌记	colophon	又称书牌、木记、刊记，指古籍中专门标记刊刻信息的独立版面单元，多位于卷首、卷末或目录之后，常规刊刻刊刻朝代、年月、刊刻机构、刊刻者、坊肆堂号、版权声明等内容，是古籍版本鉴定、刊刻源流考证的核心直接依据。	刻书题记、刊记、坊肆信息，多在卷首/卷末。 <牌记>万历二十三年金陵唐氏刊本</牌记>…

序号	中文标签名称	英文标签名称	标签释义	标签用法（示例）
44	印章	seal	又称铃印，指古籍书页上铃盖的印章印记，是古籍流传、收藏、鉴定的核心实物依据，按功能可分为藏书印、鉴藏印、刊刻印、闲章、官印、御玺等类别，标注时需标记其位置、类型、形制等核心属性。	<印章 位置="3,3" 宽度="100" 高度="100"></图片链接>images/seal.png</印章>
45	墨钉	ink-nail	又称墨等、黑钉，指古籍雕版版面中刊刻的黑色实心方形占位符号，一个墨钉对应一个汉字的位置，核心功能是标记刻版时缺文待补、文字未定、避讳占位的位置，是古籍刻本中常见的校勘占位符号，标注时需标记占位原因、对应原字（如可考证）等属性。	<墨钉/>
46	空围	empty-enclosure	又称白匡、方空、白丁，指古籍版面中以空心方框（□）或实心白色方块呈现的占位符号，一个空围对应一个汉字的位置，核心功能是标记脱文、漫漶无法辨识的文字、避讳空字、失传古字的位置，是古籍抄本、刻本中通用的缺文标记符号，标注时需标记占位原因、对应原字（可考时）等属性。	<空围/>
47	题辞	inscription	他人题写的诗词或短文，通常位于卷首。	<题辞><诗词>...</诗词></题辞>
48	地图	map	带有地理方位信息的特殊图片类型。	<地图> <插图 页码 ="15"> <图片链接>/img/插图 1. jpg</图片链接> <图题>春秋势力图</图题> <图注>...</图注> </插图></地图>
49	符咒	charm	宗教、民俗文献中的神秘图案或非标准汉字组合。	<符咒> <插图 页码 ="15"> <图片链接>/img/插图 1. jpg</图片链接></插图></符咒>
50	落款	signature	常见于前言、序等地方，章节内容编写者的落款署名、时间等。	<落款>编者</落款> <落款>2022 年 1 月 31 日</落款>
51	引文	reference	来自其他图书的内容。	用于标注引用其他图书的内容，一般字体不同于正文文字。 <引文>我走着，只觉得全身空虚，轻飘飘的，有时若不倚着东西，就怕会向前扑下去。遇到这样的情形时，我就倚着电杆，暂时不动，等好了一点才走。这种感觉起先只是在近午时才有，后来就时时有了，甚至于倚着电杆，亦觉得身在半空似的，四下的土地都在移动颠簸</引文>
52	强调	hi	用于标注仅有排版样式差异、无明确语义的文本。	用于标注与正文字体不一致的内容，通常为楷体。 <强调>我走着，只觉得全身空虚，轻飘飘的，有时若不倚着东西，就怕会向前扑下去。遇到这样的情形时，我就倚着电杆，暂时不动，等好了一点才走。这种感觉起先只是在近午时才有，后来就时时有了，甚至于倚着电杆，亦觉得身在半空似的，四下的土地都在移动颠簸</强调>
53	内容页	content-page	记录内容的图片页。	用于标注记录内容的图片页。嵌套书页图进行标注。 <序> <内容页> <书页图> <图片链接>imageCut/书页图 1. jpg</图片链接> </书页图> </内容页> </序>

B.1.5 样式标签使用规则

样式标签用于标注文本内容中的特殊展现样式。样式标签无法实现对书籍的完整标注，仅体现文本的样式，包括斜体、粗体、居左、居中、居右等，在实际标注过程中需配合结构标签和呈现标签使用，在结构标签和呈现标签标注的内容中，嵌套样式标签使用。

序号	中文标签名称	英文标签名称	标签释义	标签用法（示例）
1	分段	p	构成文章的基本单位，通常由几句话或一组句子组成，用于表达一个相对完整的思想或意义。段落之间通常有明显的分隔标记，如换行或缩进等。	用于标注文中出现的文段内容。 <分段> 时令已快到惊蛰,雪当然再不会存留,往往还没等落地,就已经消失得无踪无影了。</分段>
2	粗体	bold	通过加粗文字的字形来强调或突出某些内容。粗体常用于标题、关键词或需要特别强调的文本。	<粗体>时令已快到<斜体>惊蛰</斜体>,雪当然再不会存留,往往还没等落地,就已经消失得无踪无影了。</粗体>
3	斜体	italic	通过将文字倾斜来呈现不同的视觉效果。斜体常用于表示书名、人名、引文或需要稍微强调但不至于过于突兀的文本。	<粗体>时令已快到<斜体>惊蛰</斜体>,雪当然再不会存留,往往还没等落地,就已经消失得无踪无影了。</粗体>
4	下划线	underline	在文字下方添加的一条直线，用于强调、标注或链接文本。下划线常用于表示超链接、拼写错误或需要特别注意的文本。	用于标注文中带下划线的文字。若整段带下划线的文字不需要再有分段标签。 <粗体>时令已快到<下划线>惊蛰</下划线>,雪当然再不会存留,往往还没等落地,就已经消失得无踪无影了。</粗体>
5	上着重	super-emphasis	对文本中的某些内容进行特别强调或突出。除了使用粗体、斜体或下划线等排版方式外，还可以通过添加特殊标记来实现着重效果。	用于标注文中带着上重号的文字。若整段为上重重的话不需要再有分段标签。 <粗体>时令已快到<上着重>惊蛰</上着重>,雪当然再不会存留,往往还没等落地,就已经消失得无踪无影了。</粗体>
6	下着重	sub-emphasis	对文本中的某些内容进行特别强调或突出。除了使用粗体、斜体或下划线等排版方式外，还可以通过添加特殊标记来实现着重效果。	用于标注文中带着下重号的文字。若整段为下重重的话不需要再有分段标签。 <粗体>时令已快到<下着重>惊蛰</下着重>,雪当然再不会存留,往往还没等落地,就已经消失得无踪无影了。</粗体>
7	上标	superscript	比同一行中其他文字稍高的文字，通常作为一种上角标志的符号。主要用途包括：脚注或引用标记和科学表达、特殊意义的扩展代号。	<数学公式> a<上标>2</上标>-b<上标>2</上标>=(a+b)(a-b) </数学公式>
8	下标	subscript	出现在正常字体下边的数字、字母或其他标志，常用于图书中的公式、数学表达式或化学复合物的描述。	<化学公式> 2H<下标>2</下标>O→2H<下标>2</下标>+O2 </化学公式>
9	居左	left-align	将元素向页面的左边靠齐，从而在视觉上形成整齐有序的布局。	用于标注文中居左展示的文字。 <居左>你好</居左>
10	居中	center-align	将元素放置在页面的中心位置，使左右两侧的空间相等。	用于标注文中居中展示的文字。 <居中>你好</居中>
11	居右	right-align	在设计中将元素放在画面的右侧，使得页面看起来更加平衡。	用于标注文中居右展示的文字。 <居右>你好</居右>
12	删除线	strike-through	在文字上画一条线，用于表示该文字已被删除或不再适用。在文档编辑中，删除线常用于标记需要删除的文本内容。	用于标注文中带删除线的文字。若整段带删除线的文字不需要再有分段标签。 <粗体>时令已快到<删除线>惊蛰</删除线>,雪当然再不会存留,往往还没等落地,就已经消失得无踪无影了。</粗体>
13	横排	horizontal-format	字符由左至右横向顺序排列成行的排版格式。	用于标注文中横向排版的文字内容。 <横排>时令已快到<删除线>惊蛰</删除线>,雪当然再不会存留,往往还没等落地,就已经消失得无踪无影了。</横排>
14	竖排	vertical-format	字符由上而下竖向排列成行的排版格式。	用于标注文中竖向排版的文字内容。 <竖排>时令已快到<删除线>惊蛰</删除线>,雪当然再不会存留,往往还没等落地,就已经消失得无踪无影了。</竖排>
15	文字边框	border	带文字边框的段落或内容。	<文字边框> 语句（1）（2）中含有变量…… </文字边框>

16	段落底色	bg-color	带底色的段落或内容。	<分段> <段落底色>语句（1）（2）中含有变量……</段落底色> 在无法判断它们的真假 …… </分段>
17	正文小字	small-style	字体小于周围正常文字的内容。	<分段> <正文小字>在</正文小字> 这寒假之前…… </分段>
18	正文大字	big-style	字体大于周围正常文字的内容。	<问候语> <正文大字>亲爱</正文大字> 的唐冉： </问候语>
19	朱色	red-color	红色印刷或手写的文字。	<分段> ……<朱色>何以解忧</朱色>… </分段>
20	墨色	black-color	黑色印刷或手写的文字，未标引时默认为墨色。	<分段> ……<墨色>何以解忧</墨色>… </分段>

表 B.4 图书样式标签使用规则

B.1.6 辅助标签使用规则

辅助标签中设计语种标签，用于标引语言的种类，如中文、英文、法文等。语言是人类交流思想的工具；每种语言都有其独特的词汇、语法和表达方式。使用<语种>或<languages>标签进行标引，用于标注文中出现的特殊语种内容。并使用属性 lang（xs: language）区分语种类型：1. 英语；2. 法语；3. 西班牙语；4. 日语；5. 韩语；6. 德语；7. 阿拉伯语；8. 俄语。

示例：

用于标注文中出现的特殊语种内容。

<语种 lang="fr"> Un ami est l'une des plus belles choses que l'on puisse avoir, et l'une des meilleures choses que l'on puisse être.</语种>

B.2 报纸 XML 标签使用规则

B.2.1 报纸标签分类

报纸标签按用途分为元数据标签、结构标签、呈现标签、样式标签和辅助标签。

B.2.2 元数据标签使用规则

用于标引报纸版权的元数据信息。根标签为<版权元数据>或<copyright-meta>。

表 B.5 报纸元数据标签使用规则

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
1	报纸唯一标识符	identifier	在特定上下文中，给予报纸资源的一个明确的标识。一般采用字符串或数字代码。建议采用符合正式标识体系的字符串进行标识。	<报纸唯一标识符>××××</报纸唯一标识符>
2	正报名	title	报纸在新闻出版管理部门登记的正式名称。	<正报名>人民日报</正报名>
3	CN	cn-number	中国报刊管理部门为统计管理编制的代号体系，以“CN”为国别标识，由地区号、序号及分类号组成，1988 年正式实施。	<统一刊号>CN11-0065</统一刊号>
4	ISSN	issn	国际标准连续出版物号。	<ISSN>0253-1795</ISSN>
5	期	volume-number	报纸在本年度的连续出版顺序编号。	<期>2026 年第 12 期</期>
6	总期号	total-issue-	从创刊号开始累计的跨年度总序号。	<总期号>32456</总期号>

		number		
7	总版数	total-pages	本期报纸总版面数（如：16版）。	<总版数>16</总版数>
8	出版日期	publication-date	本期报纸正式出版发行的具体年月日。	<出版日期>2023-10-25</出版日期>
9	主管者	in-charge	出版单位创办时的申请者以及该出版单位的主办单位的上级领导部门。	<主管者>中共中央宣传部</主管者>
10	出版者	publisher	从事出版活动的专业机构。	<出版者>人民日报社</出版者>
11	印刷者	printer	从事印制、复制活动的专业机构。	<印刷者>北京日报印务有限责任公司</印刷者>
12	发行者	delivery	从事出版物发行的机构。	<发行者>人民日报社 </发行者>
13	主编	chief-editor	作品编纂工作或出版物编辑工作的主要负责人。	<主编>张××</主编>
14	出版频率	publication-frequency	报纸的出版频率，如日报、周报、旬报、半月报。	<出版频率>日报</出版频率>
15	幅面尺寸	format	报纸的单幅面尺寸规格。	<幅面尺寸>对开</幅面尺寸>
16	单价	unit-price	单份报纸的零售价格。	<单价>CNY 2.00</单价>
17	年价	whole-year-price	报纸全年的订阅总费用。	<年价>CNY 720.00</年价>
18	语言种类	language	报纸正文使用的主要语言文字类别。	<语言种类>中文</语言种类>
19	网址	web-site	报纸出版单位的官方网站地址。	<网址>http: //www. examplepaper. com</网址>
20	地址	address	报纸出版单位所在的物理通讯地址。	<地址>北京市朝阳区金台西路2号</地址>
21	邮编	zip-code	报纸出版单位所在地的邮政编码。	<邮编>100733</邮编>
22	办公电话	office-number	报社用于日常行政与业务联系的电话号码。	<办公电话>010-6536****</办公电话>
23	办公传真	fax-number	报社用于文件传输的传真号码。	<办公传真>010-6536****</办公传真>

B.2.3 结构标签使用规则

报纸结构标签用于标注报纸的逻辑结构组成部分，主要体现报纸中保留下来的物理结构和组成报纸完整内容的分块，可分为版面、报头等。

表 B.6 报纸结构标签使用规则

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
1	报纸	newspaper	以国内外社会、政治、经济、文化等新闻为主要内容，定期出版发行的散页出版物。	<报纸> ...<版面> ... <报头>×××</报头> ...</版面> ...<报纸>
2	版面	page	报纸上每一页的整面。	<版面> ... <报头>×××</报头> ...</版面>
3	报头	masthead	报纸刊登报名的地方，一般都在第一版的上端。	<版面> ... <报头>×××</报头> ...</版面>
4	刊登报名	published-masthead	刊登报纸名称。	<版面> ... <刊登报名>《中国青年报》</刊登报名> ...</版面>
5	英文报名	english-registration	刊登报纸英文命名。	<版面> ... <英文报名>China Daily、Beijing Review</英文报名> ...</版面>
6	报眉	newspaper-eyebrow	报纸版面上方眉线区域所印的文字部分。	<版面> ... <报眉>2026年4月21日 星期二 第12876期 总第23456版</报眉> ...</版面>
7	报眼	newspaper-eye	横排报纸报头旁边的版面。	<版面> ... <报眼>今日要闻摘要</报眼> ...</版面>

8	内容提要	summary	对某一文献或资料内容进行高度概括和总结的一种形式。	<版面> ... <内容提要>本文分析了 2026 年第一季度经济数据, 指出消费市场回暖趋势明显。</内容提要> ...</版面>
9	头条	headline	在新闻报道中最重要、最具吸引力的内容。	<版面> ... <头条>×××</头条> ...</版面>
10	倒头条	behind-page	对一个事件或社会话题的总结。	<版面> ... <倒头条>×××</倒头条> ...</版面>
11	导读	navigation	引导或辅导阅读行为的词语。	<版面> ... <导读>×××</导读> ...</版面>
12	栏目	column	某一类内容的标题或版面。	<版面> ... <栏目>《理论纵横》</栏目> ...</版面>

B.2.4 呈现标签使用规则

呈现标签用于标注报纸包含的实质性文字和图片元素, 结合业务应用实际和标注难易度, 分为图片、段落、公式、注释、诗词等。

表 B.7 报纸呈现标签使用规则

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
1	新闻	news	对新近发生或正在发生的具有社会意义的事实的报道。	<版面> ... <新闻>《我国成功发射遥感卫星三十一号》</新闻> ... </版面>
2	新闻标题	news-title	新闻的核心标题, 用于概括新闻主要内容, 吸引读者注意力, 通常位于文章最显眼位置, 字体较大、醒目。	<新闻标题>科技巨头联手推动人工智能发展 产业变革加速到来</新闻标题>
3	新闻引题	news-pre-title	又称“肩题”或“眉题”, 位于主标题之上或之前, 用于交代背景、烘托气氛、引出主标题, 起引导或铺垫作用。	<新闻引题>全球科技竞争进入新阶段——</新闻引题>
4	新闻副题	news-sub-title	又称“子题”, 位于主标题之后或下方, 用于补充说明新闻内容、点明意义、范围或结果, 对主标题进行细化或延伸。	<新闻副题>——多领域应用落地, 伦理挑战仍待破解</新闻副题>
5	作者	news-creator	标明新闻的撰写者或责任人的署名, 体现新闻的来源和权威性, 可能包括姓名、职务或所属部门。	<作者>记者 李铭 实习生 王芳</作者>
6	来源	news-source	指新闻的发布机构或转载的出处, 如报社名称、通讯社、网站等, 用于说明新闻的原始来源和版权信息。	<来源>人民日报社 2026 年 4 月 21 日</来源>
7	文章区	article-block	报纸上一组关联文本、图片等元素所在的容器。	<版面> ... <文章区><文章标题>...</文章标题><分段>...</分段></文章区> ...</版面>
8	文章标题	article-title	标明文章、作品等内容的简短语句。	<版面> ... <文章标题>《乡村振兴战略实施五周年成效显著》</文章标题> ...</版面>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
9	副标题	subtitle	在主标题之后的短语或句子，通常用于补充说明主标题的内容。	<版面> ... <副标题>——聚焦产业融合与人才振兴</副标题> ...</版面>
10	关键词	keyword	从文献的題目、正文和摘要中抽选出来的，能够表达主题内容特征的重要词汇。	<版面> ... <关键词>乡村振兴、产业融合、人才振兴、农村电商</关键词> ...</版面>
11	记者	reporter	从事信息采集和新闻报道的专业人员。	<版面> ... <记者>新华社记者 王晓</记者> ...</版面>
12	分段	p	段落是构成文章的基本单位，通常由几句话或一组句子组成，用于表达一个相对完整的思想或意义。段落之间通常有明显的分隔标记，如换行或缩进等。	用于标注文中出现的文段内容。 <分段> 时令已快到惊蛰,雪当然再不会存留,往往还没等落地,就已经消失得无踪无影了。</分段>
13	诗词	poem	文学创作的具备特殊格式的文本内容。	用于标注报纸中出现的诗词，单句出现在正文中的不需要标诗词。 <诗词> <诗词标题>短歌行</诗词标题> <诗词作者>曹操</诗词作者> <诗词内容> 对酒当歌，人生几何！ 譬如朝露，去日苦多。 </诗词内容> </诗词>
14	诗词标题	poem-title	诗词的标题。	用于标注诗词标题。 <诗词> <诗词标题>短歌行</诗词标题> <诗词作者>曹操</诗词作者> <诗词内容> 对酒当歌，人生几何！ 譬如朝露，去日苦多。 </诗词内容> </诗词>
15	诗词内容	poem-body	诗词的主体内容。	用于标注诗词内容。 <诗词> <诗词标题>短歌行</诗词标题> <诗词作者>曹操</诗词作者> <诗词内容> 对酒当歌，人生几何！ 譬如朝露，去日苦多。 </诗词内容> </诗词>
16	诗词作者	poet	诗词的作者。	用于标注诗词作者。 <诗词> <诗词标题>短歌行</诗词标题> <诗词作者>曹操</诗词作者> <诗词内容> 对酒当歌，人生几何！ 譬如朝露，去日苦多。 </诗词内容> </诗词>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
17	书信	letter	用于交流的具备特殊格式的文本内容。	用于标注书信内容。 <书信> <书信标题>手连手心连心</书信标题> <开头称谓>亲爱的唐冉</开头称谓> <问候语>你好！</问候语> <书信内容> 在这寒假之前，我们学校举行了一次“手拉手”的活动…… …… </书信内容> <结束语>身体健康，学习进步！</结束语> <书信落款>你的朋友陈研</书信落款> <书信落款>2018年2月26日</书信落款> </书信>
18	开头称谓	salutation	书信开头的称谓、提称语。	用于标注书信开头称谓。 <书信> <书信标题>手连手心连心</书信标题> <开头称谓>亲爱的唐冉</开头称谓> <问候语>你好！</问候语> <书信内容> 在这寒假之前，我们学校举行了一次“手拉手”的活动…… …… </书信内容> <结束语>身体健康，学习进步！</结束语> <书信落款>你的朋友陈研</书信落款> <书信落款>2018年2月26日</书信落款> </书信>
19	问候语	greeting	书信的开头语、启事敬辞。	用于标注书信中内容开始前的问候语。 <书信> <书信标题>手连手心连心</书信标题> <开头称谓>亲爱的唐冉</开头称谓> <问候语>你好！</问候语> <书信内容> 在这寒假之前，我们学校举行了一次“手拉手”的活动…… …… </书信内容> <结束语>身体健康，学习进步！</结束语> <书信落款>你的朋友陈研</书信落款> <书信落款>2018年2月26日</书信落款> </书信>
20	书信内容	letter-body	书信的主体内容。	用于标注书信内容。 <书信> <书信标题>手连手心连心</书信标题> <开头称谓>亲爱的唐冉</开头称谓> <问候语>你好！</问候语> <书信内容> 在这寒假之前，我们学校举行了一次“手拉手”的活动…… …… </书信内容> <结束语>身体健康，学习进步！</结束语> <书信落款>你的朋友陈研</书信落款> <书信落款>2018年2月26日</书信落款> </书信>
21	书信落款	letter-signature	书信落款的署名、时间等。	用于标注书信落款（写信人+时间）。 <书信> <书信标题>手连手心连心</书信标题> <开头称谓>亲爱的唐冉</开头称谓> <问候语>你好！</问候语> <书信内容> 在这寒假之前，我们学校举行了一次“手拉手”的活动…… …… </书信内容> <结束语>身体健康，学习进步！</结束语> <书信落款>你的朋友陈研</书信落款> <书信落款>2018年2月26日</书信落款> </书信>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
22	补字图	glyph	对计算机无法输出的字符用图片形式展示。	用于标注计算机无法输出的字符。 <分段> 惊 <补字图><图片链接>/img/蛰.jpg</图片链接></补字图>，是二十四节气中的第三个节气。斗指丁，太阳到达黄经345°，于公历3月5—6日交节。 </分段>
23	表格图	table-fig	表格内容以图片形式展示。	用于标注文中出现的表格，以表格图方式进行标注。 <附录> <分段>附录二</分段> <表格图 页码="10" 宽度="800" 高度="600"><图片链接>/img/表格图1.jpg</图片链接></表格图> ... </附录>
24	公式图	graphic-formula	在图形中嵌入数学公式或表达式的图表。	对于无法用文字进行标注的公式，用公式图进行标注。 <附录> <分段>附录二</分段> <公式图 pageCode="11" 宽度="400" 高度="100"><图片链接>/img/公式图.jpg</图片链接></公式图> ... </附录>
25	插图	figure	插在书刊文字中间用于说明内容的图画。	用于标注插入在文中的图片。 <作者简介> <分段> 时令已快到惊蛰,雪当然再不会存留,往往还没等落地,就已经消失得无影无踪了。</分段> <插图 页码="12" 宽度="1024" 高度="768" > <图片链接>/img/插图.jpg</图片链接> </插图> ... </作者简介>
26	落款图	signature-picture	常见于前言、序等地方，章节内容编写者的无法用简单文字表达的落款。	用于标注文中内容的落款图片。标签中页码属性表示该图出现的页码。 <分段> <落款图 页码="15"><图片链接>imageCut/img00015001.jpg</图片链接></落款图> ... </分段>
27	背景图	background-picture	用于页面、底部的图像，它作为背景存在，不直接参与内容的展示，但可以为整个页面或文档提供视觉上的背景支持或氛围营造。	用于标注文中内容的背景图片。 <分段> <背景图 页码="15"><图片链接>imageCut\img00015001.jpg</图片链接></背景图> ... </分段>
28	商标图	trademark-image	用于识别和区分商品或服务的来源的视觉符号。	<版面> ...<商标图> <图片链接>/img/商标图.jpg</图片链接> </商标图> ...</版面>
29	广告	advertisement	通过有偿的方式向公众传播产品、服务或观念的一种宣传形式。	<版面> ... <广告> <图片链接>/img/广告图.jpg</图片链接> </广告>...</版面>
30	勘误表	errata	附在书刊中，更正文字错误的表格。	用于标注报纸中的勘误表部分内容，主要是关键信息对应的页面页码。 <勘误表> <表格图><图片链接>/img/表格图.jpg</图片链接></表格图> </勘误表>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
31	图题	picture-caption	图片的标题。	用于标注图片题目，通常出现在图片下方。 <附录> <分段>附录二</分段> <插图 页码 = "15"><图片链接>/img/插图.jpg</图片链接> <图题>麦克斯韦方程组</图题> </插图> ... </附录>
32	图注	picture-footnote	图片的注释。	用于标注图片补充性注释、图例等，通常出现在图题下方。 <附录> <分段>附录二</分段> <插图 页码 = "15"> <图片链接>/img/插图.jpg</图片链接> <图题>麦克斯韦方程组</图题> <图注>使用高斯单位制</图注></插图> ... </附录>
33	有线表	border-table	具有明显的横竖线条来界定单元格的表格。	<有线表> <表题>苗语句典表</表题> ... </有线表>
34	无线表	table	没有明显的边框线的表格。	<无线表> <表题>苗语句典表</表题> ... </无线表>
35	表题	caption	表格题名。	<有线表> <表题>苗语句典表</表题> ... </有线表>
36	表头	table-head	表格表头，用于标注表格顶部第一行的单元格区域。	<有线表> <表题>语言句典表</表题> <表头> <单元格 行范围="2">条目</单元格> <单元格 列范围="3" 行范围="1">普通话</单元格> <单元格 行范围="2">苗语</单元格> <单元格 行范围="2">分词情况</单元格> <单元格 行范围="1">北京话</单元格> <单元格 行范围="1">广东话</单元格> <单元格 行范围="1">上海话</单元格> </表头> </有线表>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
37	表正文	table-body	表格数据，用于标注表格内容数据。	<有线表> <表题>苗语句典表</表题> <表头> <单元格 行范围="2">条目</单元格> <单元格 列范围="3" 行范围="1">普通话</单元格> <单元格 行范围="2">苗语</单元格> <单元格 行范围="2">分词情况</单元格> <单元格 行范围="1">北京话</单元格> <单元格 行范围="1">广东话</单元格> <单元格 行范围="1">上海话</单元格> </表头> <表正文> <行> <单元格>条目 1</单元格> <单元格>第一行第一列</单元格> <单元格>第一行第二列</单元格> <单元格>第一行第三列</单元格> </行> <行> <单元格>条目 2</单元格> <单元格>第二行第一列</单元格> <单元格>第二行第二列</单元格> <单元格>第二行第三列</单元格> </行> </表正文> </有线表>
38	表尾	table-foot	表格总结、汇总等，通常出现在表内结尾。	用于标注底部区域，是表格内容结束后的部分，如：汇总、结论、合计等。 <有线表> <表正文> ... </表正文> <表尾> <行> <单元格>条目 2</单元格> <单元格>第二行第一列</单元格> <单元格>第二行第二列</单元格> <单元格>第二行第三列</单元格> </行> </表尾> </有线表>
39	行	table-row	表格内用于以行为单位为单元格分组。	用于标注表格中单元格隔行。 <有线表> ... <表正文> <行> <单元格>条目 1</单元格> <单元格>第一行第一列</单元格> <单元格>第一行第二列</单元格> <单元格>第一行第三列</单元格> </行> <行> <单元格>条目 2</单元格> <单元格>第二行第一列</单元格> <单元格>第二行第二列</单元格> <单元格>第二行第三列</单元格> </行> </表正文>... </有线表>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
40	单元格头	table-cell-header	表格内单元格头。	<p>用于标注表格中最小单元格，跨行或跨列的单元格需设置属性（跨行数/列数）。</p> <p><有线表></p> <p><表题>苗语句典表</表题></p> <p><表头></p> <p><单元格头 行范围="2">条目</单元格头></p> <p><单元格头 列范围="3" 行范围="1">普通话</单元格头></p> <p><单元格头 行范围="2">苗语</单元格头></p> <p><单元格头 行范围="2">分词情况</单元格头></p> <p><单元格头 行范围="1">北京话</单元格头></p> <p><单元格头 行范围="1">广东话</单元格头></p> <p><单元格头 行范围="1">上海话</单元格头></p> <p></表头></p> <p><表正文></p> <p>...</p> <p></表正文></p> <p></有线表></p>
41	单元格	table-cell	表格内单元格。	<p>用于标注表格中最小单元格，跨行或跨列的单元格需设置属性（跨行数/列数）。</p> <p><有线表></p> <p><表题>苗语句典表</表题></p> <p><表头></p> <p><单元格 行范围="2">条目</单元格></p> <p><单元格 列范围="3" 行范围="1">普通话</单元格></p> <p><单元格 行范围="2">苗语</单元格></p> <p><单元格 行范围="2">分词情况</单元格></p> <p><单元格 行范围="1">北京话</单元格></p> <p><单元格 行范围="1">广东话</单元格></p> <p><单元格 行范围="1">上海话</单元格></p> <p></表头></p> <p><表正文></p> <p><行></p> <p><单元格>条目 1</单元格></p> <p><单元格>第一行第一列</单元格></p> <p><单元格>第一行第二列</单元格></p> <p><单元格>第一行第三列</单元格></p> <p></行></p> <p><行></p> <p><单元格>条目 2</单元格></p> <p><单元格>第二行第一列</单元格></p> <p><单元格>第二行第二列</单元格></p> <p><单元格>第二行第三列</单元格></p> <p></行></p> <p></表正文></p> <p></有线表></p>
42	表注	table-footnote	表格注释。通常位于表格下方。	<p><有线表></p> <p><表题>苗语句典表</表题></p> <p><表正文></p> <p></表正文></p> <p></有线表></p> <p><表注>特指 2025 年数据</表注></p>

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
43	引文	reference	来自其他报纸的内容	用于标注引用其他报纸的内容，一般字体不同于正常文字的内容。 <引文>我走着，只觉得全身空虚，轻飘飘的，有时若不倚着东西，就怕会向前扑下去。遇到这样的情形时，我就倚着电杆，暂时不动，等好了一点才走。这种感觉起先只是在近午时才有，后来就时时有了，甚至于倚着电杆，亦觉得身在半空似的，四下的土地都在移动颠簸</引文>
44	注释编号	annotation-number	对报纸的某些内容或文字所作的说明的编号。	用于标注文中出现的注释编号。 <分段> 时令已快到惊蛰<注释编号>⑥</注释编号>，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。 </分段>
45	注释内容	annotation-content	对报纸的某些内容或文字所作的说明。	用于标注对于某些词句的解释说明相关内容。 <注释内容> <注号>⑥</注号>惊蛰，是二十四节气中的第三个节气。斗指丁，太阳到达黄经 345°，于公历 3 月 5-6 日交节。 </注释内容>

B. 2. 5 样式标签使用规则

样式标签用于标注报纸文本内容中的特殊展现样式，包括斜体、粗体、居左、居中、居右等，在实际标注过程中需配合结构标签和呈现标签使用，在结构标签和呈现标签标注的内容中，嵌套样式标签使用。

表 B. 8 报纸样式标签使用规则

序号	标签中文名称	标签英文名称	标签释义	标签用法（示例）
1	上标	superscript	上标指的是比同一行中其他文字稍高的文字，通常作为一种上角标志的符号。它在报纸中主要有以下几种用途：脚注或引用标记、数学和科学表达、特殊意义的扩展代号。	用于标注文中展示为上标的文字。 〈数学公式〉 a<上标>2</上标>-b<上标>2</上标>=(a+b) (a-b) 〈/数学公式〉
2	下标	subscript	下标是指出现在正常字体下边的数字、字母或其他标志，常用于报纸中的公式、数学表达式或化学复合物的描述。	用于标注文中展示为下标的文字。 〈化学公式〉 2H<下标>2</下标>O→2H<下标>2</下标>+O2 〈/化学公式〉
3	上着重	super-emphasis	着重是指对文本中的某些内容进行特别强调或突出。除了使用粗体、斜体或下划线等排版方式外，还可以通过添加特殊标记来实现着重效果。	用于标注文中带着上重号的文字。若整段为上着重，则不需要再有分段标签。 〈加粗〉时令已快到<上着重>惊蛰</上着重>，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。〈/加粗〉
4	下着重	sub-emphasis	着重是指对文本中的某些内容进行特别强调或突出。除了使用粗体、斜体或下划线等排版方式外，还可以通过添加特殊标记来实现着重效果。	用于标注文中带着下重号的文字。若整段为下着重，则不需要再有分段标签。 〈加粗〉时令已快到<下着重>惊蛰</下着重>，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。〈/加粗〉
5	下划线	underline	下划线是在文字下方添加的一条直线，用于强调、标注或链接文本。在文档中，下划线常用于表示超链接、拼写错误或需要特别注意的文本。	用于标注文中带下划线的文字。若整段带下划线，则不需要再有分段标签。 〈加粗〉时令已快到<下划线>惊蛰</下划线>，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。〈/加粗〉
6	删除线	strike-through	删除线是在文字上划一条线，用于表示该文字已被删除或不再适用。在文档编辑中，删除线常用于标记需要删除的文本内容。	用于标注文中带删除线的文字。若整段带删除线，则不需要再有分段标签。 〈加粗〉时令已快到<删除线>惊蛰</删除线>，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。〈/加粗〉
7	特殊符号	special-id	指在书写和交流中具有特定意义或用途的符号	〈版面〉 …〈特殊符号〉××××〈/特殊符号〉…〈/版面〉
8	居左	left-align	居左是指将元素向页面的左边靠齐，从而在视觉上形成整齐有序的布局。	用于标注文中居左展示的文字。 〈居左〉你好</居左〉
9	居中	center-align	居中是指将元素放置在页面的中心位置，使左右两侧的空间相等。	用于标注文中居中展示的文字。 〈居中〉你好</居中〉
10	居右	right-align	居右是指在设计中将元素放在画面的右侧，使得页面看起来更加平衡。	用于标注文中居右展示的文字。 〈居右〉你好</居右〉
11	粗体	bold	粗体是一种文字排版方式，通过加粗文字的字形来强调或突出某些内容。在文档、网页或印刷品中，粗体常用于标题、关键词或需要特别强调的文本。	用于标注文中加粗的文字。若整段为加粗，则不需要再有分段标签。 〈粗体〉时令已快到<斜体>惊蛰</斜体>，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。〈/粗体〉
12	斜体	italic	斜体是一种文字排版方式，通过将文字倾斜来呈现不同的视觉效果。斜体常用于表示书名、人名、引文或需要稍微强调但不至于过于突兀的文本。	用于标注文中斜体的文字。若整段为斜体的话，则不需要再有分段标签。 〈加粗〉时令已快到<斜体>惊蛰</斜体>，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。〈/加粗〉
13	文字边框	border	带文字边框的段落或内容	用于标注带文字边框的内容。 〈文字边框〉 语句（1）（2）中含有变量…… 〈/文字边框〉
14	段落底色	bg-color	带底色的段落或内容	用于标注带底色的内容。 〈分段〉 〈段落底色〉语句（1）（2）中含有变量…… 〈/段落底色〉 在无法判断它们的真假 …… 〈/分段〉
15	小字	small-style	字体小于周围正常文字的内容	用于标注字体小于周围正常文字的内容。 〈分段〉 〈小字〉在</小字〉 这寒假之前…… 〈/分段〉

16	大字	big-style	字体大于周围正常文字的内容	用于标注字体大于周围正常文字的内容。 <问候语> <大字>亲爱</大字> 的唐冉: </问候语>
17	强调	hi	用于标注仅有排版样式差异、无明确语义的文本。	用于标注与正文字体不一致的内容,通常为楷体 <强调>我走着,只觉得全身空虚,轻飘飘的, 有时若不倚着东西,就怕会向前扑下去。遇到这样 的情形时,我就倚着电杆,暂时不动,等好了一点才走。 这种感觉起先只是在近午时才有,后来就时时有了,甚至 于倚着电杆,亦觉得身在半空似的,四下的土地都在移动 颠簸</强调>

B.2.6 辅助标签使用规则

辅助标签中设计语种标签,同图书同名标签。

B.3 期刊 XML 标签使用规则

B.3.1 期刊标签分类

期刊标签按用途分为元数据标签、结构标签、呈现标签、样式标签和辅助标签。

B.3.2 元数据标签使用规则

用于标引期刊版权元数据信息。根标签为<版权元数据>或<copyright-meta>。

表 B.9 期刊元数据标签使用规则

序号	中文名称	英文标签	标签释义	标签用法(示例)
1	期刊唯一标识符	identifier	在特定上下文环境中,给予期刊资源的一个明确的标识。一般采用字符串或数字代码。建议采用符合正式标识体系的字符串进行标识。	<期刊唯一标识符>××××××</期刊唯一标识符>
2	正刊名	periodical-title	期刊的全称标识。	<正刊名>文化教育 </正刊名>
3	主办者	sponsor	在经过审批和授权后,负责该期刊的运营、管理和出版工作的机构或组织。	<主办者>××传媒集团</主办者>
4	创刊日期	founding-date	期刊第一次正式出版发行的年月日	<创刊日期>1950-08-15</创刊日期>
5	出版频率	publication-frequency	期刊出版的固定频率,如月刊、季刊、半月刊。	<出版频率>月刊</出版频率>
6	刊期	issue	一种期刊每年出版的频次。	<刊期>12</刊期>
7	卷	volume-number	期刊按年度或内容划分的卷次编号。	<卷>第 45 卷</卷>
8	期	issue-number	期刊在某一卷内的出版顺序号。	<期>第 3 期</期>
9	总期号	total-issue-number	从创刊号开始累计的跨年度总序号。	<总期号>156</总期号>
10	CN	cn-number	中国报刊管理部门为统计管理编制的代号体系,以“CN”为国别标识,由地区号、序号及分类号组成,1988 年正式实施。	<CN>CN11-1786/N</CN>
11	ISSN	issn	国际标准连续出版物号。	<ISSN>0577-6686</ISSN>
12	出版日期	publication-date	本期期刊正式出版发行的年月日。	用于描述期刊的出版日期。 <出版日期>2024 年 5 月 16 日</出版日期>
13	定价	price	由出版者印制在出版物上的价格。	用于描述期刊的售价。 <定价>CNY 10.00</定价>
14	出版者	publisher	从事出版活动的专业机构。	用于描述期刊的出版机构或组织。 <出版者>健康忠告杂志出版社</出版者>
15	出版者电子邮箱	publisher-email	出版者的电子邮箱。	用于描述期刊出版者电子邮箱。 <出版者电子邮箱>contact@jkzg.com</出版者电子邮箱>

序号	中文名称	英文标签	标签释义	标签用法（示例）
16	出版者电话	publisher-phone	出版者的联系电话号码。	用于描述期刊出版者的联系电话号码。 <出版者电话>020-88888888</出版者电话>
17	出版者地址	publisher-address	出版者的详细地址，可能包括街道、城市、省份或国家等信息。	用于描述期刊出版者的地址。 <出版者地址>北京中关村大街 31 号</出版者地址>
18	语言种类	language	期刊的主语言种类。	<语言种类>中文</语言种类>

B.3.3 结构标签使用规则

期刊结构标签用于标注期刊的逻辑结构组成部分，主要体现期刊中保留下来的物理结构和组成期刊完整内容的分块，可分为封面、封底、书脊、序等。

表 B.10 期刊结构标签使用规则

序号	中文名称	英文标签	标签释义	标签用法示例
1	期刊	periodical	定期出版的刊物。期刊结构化文件的根节点。	<期刊> ... <编辑组>×××编辑组</编辑组> ...</期刊>
2	封面	cover	书刊的外层，指期刊的封一。	用于标注期刊封面的内容，直接以图片形式保存。嵌套书页图进行标注。 <封面><书页图> <图片链接>/img/封面.jpg</图片链接> </书页图></封面>
3	封底	back-cover	书刊外表的背面部分，指期刊的封四。	用于标注期刊封底页面上的内容，直接以图片形式保存。嵌套书页图进行标注。 <封底><书页图> <图片链接>/img/书页图.jpg</图片链接> </书页图></封底>
4	书脊	spine	书背两侧的凸起部分。	用于标注书脊的内容，直接以图片形式保存。嵌套书页图进行标注。 <书脊><书页图> <图片链接>/img/书脊.jpg</图片链接> </书页图></书脊>
5	编委会	editorial-board	定义编委会。	<期刊> ... <编委会>《××学报》编委会</编委会> ...</期刊>
6	编辑组	editing-group	定义编辑组。	<期刊> ... <编辑组>《××学报》编辑组</编辑组> ...</期刊>
7	宣传语	slogan	宣传语内容。	<期刊> ... <宣传语>总有一天，你读过的书会铺成你脚下的路。</宣传语> ...</期刊>
8	序	preface	仅用于纪念刊、特刊或文学类期刊。置于正文前的独立文章。	用于标注期刊的序言部分的内容，序言中的内容根据实际情况嵌套呈现标签和样式标签。 <序> <分段>这是一段文本内容</分段> <插图> <图片链接>/img/插图 1.jpg</图片链接> </插图> ... </序>
9	目次	contents	按一定次序编排以供查考的名目。	用于标注期刊的目录部分的内容，包含章节标题及对应的页码。实际标注时通过加书签工序生成的书签 XML 导入。 <目次> <分段>第一章 绪论……001</分段> <分段>第二章……002</分段> ... </目次>

序号	中文名称	英文标签	标签释义	标签用法示例
10	索引	index	汇集期刊中包含的字词、语句、名词、事件、编号等主题,以适当方式编排,指引读者查找的检索工具。	用于标注期刊中的索引部分内容,主要是关键信息对应的页面页码。 <索引> <表格图><图片链接>/img/表格图.jpg</图片链接></表格图> </索引>
11	参考文献	bibliography	在正文之后或各部分之后一一列出的参考、引用资料的名单。	用于标注期刊中引用的文献列表。 <参考文献> <分段>【序号】作者.书名【M】.出版地:出版社,出版年份:起止页码.</分段> </参考文献>
12	附录	appendix	附在正文后面的有关文章、图片、资料。	用于标注正文后的补充性材料和附加信息,附录中的内容根据实际情况嵌套呈现标签和样式标签。 <附录> <一级标题>附录一</一级标题> <分段>这是一段文本内容</分段> <插图><图片链接>/img/插图1.jpg</图片链接></插图> <一级标题>附录二</一级标题> <表格图><图片链接>/img/表格图1.jpg</图片链接></表格图> ... </附录>
13	插页	inset	独立于期刊正常页码排序之外插入的单页或折页,常用于附赠海报、大幅拉页图片或特殊版式内容的展示。	<正文内容> ... <插页>...</插页> ...</正文内容>
14	广告页	advertising	期刊中专门用于发布商业广告、产品宣传等营销信息的页面。	<正文内容> ... <广告页>...</广告页> ...</正文内容>
15	活动页	activity	刊载与期刊相关的营销活动、读者互动、线下沙龙或征集比赛等信息的专属页面。	<正文内容> ... <活动页>...</活动页> ...</正文内容>
16	正文内容	body	期刊或文章的主要部分,包含作者想要传达的核心信息、观点、故事或研究数据等。	用于标注期刊的正文部分的内容。 <正文内容> <一级标题>...</一级标题> <分段>...</分段> <二级标题>...</二级标题> <分段>...</分段> </正文内容>
17	文章	article	期刊或出版物中独立的、有主题的文字作品,通常围绕一个特定的话题展开论述、报道或表达观点。	<正文内容><文章><标题>...</标题><副标题>...</副标题> ...</文章></正文内容>

B.3.4 呈现标签使用规则

呈现标签用于标注期刊包含的实质性文字和图片元素,结合业务应用实际和标注难易度,分为图片、段落、公式、注释等。

表 B.11 期刊呈现标签使用规则

序号	中文名称	英文标签	标签释义	标签用法示例
1	译注	translator-note	作者对作品以及译者对译文中语汇、内容、引文出处等所作的说明。	<正文内容> ... <译注>...</译注> ...</正文内容>
2	自注	author-note	作者对其著述所加的注。	<正文内容> ... <自注>...</自注> ...</正文内容>

序号	中文名称	英文标签	标签释义	标签用法示例
3	编者注	editor-note	编者对原稿、原注或译文、译注中的内容所作的说明和对原注或译注失误、不足的再注释。	<正文内容> ... <编者注>...</编者注> ...</正文内容>
4	落款	signature	常见于前言、时间、序等地方，章节内容编写者的署名等。	用于标注文章结束的署名信息及日期等。 <落款>编者</落款> <落款>2022 年 1 月 31 日</落款>
5	补字图	glyph	对计算机无法输出的字符用图片形式展示。	用于标注计算机无法输出的字符 <分段> 惊 <补字图><图片链接>/img/蛰.jpg</图片链接></补字图>，是二十四节气中的第三个节气。 斗指丁，太阳到达黄经 345°，于公历 3 月 5—6 日交节。 </分段>
6	表格图	table-fig	表格内容以图片形式展示。	用于标注文中出现的表格，以表格图方式进行标注。 <附录> <分段>附录二</分段> <表格图><图片链接>/img/表格图 1.jpg</图片链接></表格图> ... </附录>
7	公式图	graphic-formula	在图形中嵌入数学公式或表达式的图表。	对于无法用文字进行标注的公式，用公式图进行标注。 <附录> <分段>附录二</分段> <公式图><图片链接>/img/公式图.jpg</图片链接></公式图> ... </附录>
8	插图	figure	插在书刊文字中间用于说明内容的图画。	用于标注插入在文中的图片。 <作者简介> <分段>时令已快到惊蛰，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。</分段> <插图><图片链接>/img/插图.jpg</图片链接></插图> ... </作者简介>
9	落款图	signature-picture	常见于前言、序等地方，章节内容编写者的无法用简单文字表达的落款。	用于标注文中内容的落款图片。标签中页码属性表示该图出现的页码。 <分段> <落款图 页码="15"><图片链接>imageCut/img00015001.jpg</图片链接></落款图> ... </分段>
10	图题	picture-caption	图片的标题。	用于标注图片题目，通常出现在图片下方。 <附录><分段>附录二</分段><插图 页码="15"> <图片链接>/img/插图.jpg</图片链接><图题>麦克斯韦方程组</图题><插图> ...</附录>
11	图注	picture-footnote	图片的注释。	用于标注图片补充性注释、图例等，通常出现在图题下方。 <附录><分段>附录二</分段><插图 页码="15"> <图片链接>/img/插图.jpg</图片链接><图题>麦克斯韦方程组</图题><图注>使用高斯单位制</图注></插图> ...</附录>
12	表题	caption	表格题目。	表格题目，通常出现在表格正上方。 <有线表> <表题>苗语句典表</表题> ... </有线表>

序号	中文名称	英文标签	标签释义	标签用法示例
13	表注	table-footnote	表格注释。	用于标注表格注释，通常位于表格下方。 <有线表> <表题>苗语句典表</表题> <表正文>… </表正文> </有线表> <表注>特指 2025 年数据</表注>
14	表头	table-head	表格表头，用于标注表格顶部第一行的单元格区域，相当于数据表的字段名称集合。	<有线表> <表题>语言句典表</表题> <表头> <单元格头 行范围="2">条目</单元格头> <单元格头 列范围="3" 行范围="1">普通话</单元格头> <单元格头 行范围="2">苗语</单元格头> <单元格头 行范围="2">分词情况</单元格头> <单元格头 行范围="1">北京话</单元格头> <单元格头 行范围="1">广东话</单元格头> <单元格头 行范围="1">上海话</单元格头> </表头> </有线表>
15	表正文	table-body	表格数据，用于标注表格内容数据。	<有线表> … <表正文> <行> <单元格>条目 1</单元格> <单元格>第一行第一列</单元格> <单元格>第一行第二列</单元格> <单元格>第一行第三列</单元格> </行> <行> <单元格>条目 2</单元格> <单元格>第二行第一列</单元格> <单元格>第二行第二列</单元格> <单元格>第二行第三列</单元格> </行> </表正文>… </有线表>
16	表尾	table-foot	表格总结、汇总等，通常出现在表内结尾。	用于标注底部区域，是表格内容结束后的部分，如：汇总、结论、合计等。 <有线表> … <表尾> <行> <单元格>合计</单元格> <单元格>50.00</单元格> <单元格>70.00</单元格> <单元格>100.00</单元格> </行> </表尾>… </有线表>
17	行	table-row	以行为单位，为表格内的单元格分组。	用于标注表格中单元格隔行。 <有线表> … <表正文> <行> <单元格>条目 1</单元格> <单元格>第一行第一列</单元格> <单元格>第一行第二列</单元格> <单元格>第一行第三列</单元格> </行> …</表正文> </有线表>

序号	中文名称	英文标签	标签释义	标签用法示例
18	单元格头	table-cell-header	表格内单元格头。	<p>用于标注表格中最小单元格，跨行或跨列的单元格需设置属性（跨行数/列数）。</p> <p>〈有线表〉 …〈表头〉 〈单元格头 行范围="2"〉条目〈/单元格头〉 〈单元格头 列范围="3" 行范围="1"〉普通话〈/单元格头〉 〈单元格头 行范围="2"〉苗语〈/单元格头〉 〈单元格头 行范围="2"〉分词情况〈/单元格头〉 〈单元格头 行范围="1"〉北京话〈/单元格头〉 〈单元格头 行范围="1"〉广东话〈/单元格头〉 〈单元格头 行范围="1"〉上海话〈/单元格头〉 〈/表头〉… 〈/有线表〉</p>
19	单元格	table-cell	表格内单元格。	<p>用于标注表格中最小单元格，跨行或跨列的单元格需设置属性（跨行数/列数）。</p> <p>〈有线表〉 …〈表正文〉 〈行〉 〈单元格〉条目 1〈/单元格〉 〈单元格〉第一行第一列〈/单元格〉 〈单元格〉第一行第二列〈/单元格〉 〈单元格〉第一行第三列〈/单元格〉 〈/行〉 〈行〉 〈单元格〉条目 2〈/单元格〉 〈单元格〉第二行第一列〈/单元格〉 〈单元格〉第二行第二列〈/单元格〉 〈单元格〉第二行第三列〈/单元格〉 〈/行〉 〈/表正文〉… 〈/有线表〉</p>
20	直接引语	quotation	正文直接引用他人文献的具体文本片段，通常以不同字体或引号呈现。	<p>用于标注引用其他文献的内容，一般字体不同于正文文字。</p> <p>〈直接引语〉我走动时，只觉得全身空虚，轻飘飘的，有时若不倚着东西，就怕会向前扑下去。遇到这样的情形时，我就倚着电杆，暂时不动，等好了一点才走。这种感觉起先只是在近午时才有，后来就时时有，甚至于倚着电杆，亦觉得身在半空似的，四下的土地都在移动颠簸〈/直接引语〉</p>
21	书信	letter	用于交流的具备特殊格式的文本内容。	<p>用于标注书信内容。</p> <p>〈书信〉 〈书信标题〉手连手心连心〈/书信标题〉 〈开头称谓〉亲爱的唐冉〈/开头称谓〉 〈问候语〉你好！〈/问候语〉 〈书信内容〉 在这寒假之前，我们学校举行了一次“手拉手”的活动…… …… 〈/书信内容〉 〈结束语〉身体健康，学习进步！〈/结束语〉 〈书信落款〉你的朋友陈研〈/书信落款〉 〈书信落款〉2018年2月26日〈/书信落款〉 〈/书信〉</p>
22	开头称谓	salutation	书信开头的称谓、提称语。	<p>用于标注书信开头称谓。</p> <p>〈书信〉 〈书信标题〉手连手心连心〈/书信标题〉 〈开头称谓〉亲爱的唐冉〈/开头称谓〉 〈问候语〉你好！〈/问候语〉 〈书信内容〉 在这寒假之前，我们学校举行了一次“手拉手”的活动…… …… 〈/书信内容〉 〈结束语〉身体健康，学习进步！〈/结束语〉 〈书信落款〉你的朋友陈研〈/书信落款〉 〈书信落款〉2018年2月26日〈/书信落款〉</p>

序号	中文名称	英文标签	标签释义	标签用法示例
				</书信>
23	问候语	greeting	书信的开头语、启事敬辞。	用于标注书信中内容开始前的问候语。 <书信> <书信标题>手连手心连心</书信标题> <开头称谓>亲爱的唐冉</开头称谓> <问候语>你好！</问候语> <书信内容> 在这寒假之前，我们学校举行了一次“手拉手”的活动…… …… </书信内容> <结束语>身体健康，学习进步！</结束语> <书信落款>你的朋友陈研</书信落款> <书信落款>2018年2月26日</书信落款> </书信>
24	书信内容	letter-body	书信的主体内容。	用于标注书信内容。 <书信> <书信标题>手连手心连心</书信标题> <开头称谓>亲爱的唐冉</开头称谓> <问候语>你好！</问候语> <书信内容> 在这寒假之前，我们学校举行了一次“手拉手”的活动…… …… </书信内容> <结束语>身体健康，学习进步！</结束语> <书信落款>你的朋友陈研</书信落款> <书信落款>2018年2月26日</书信落款> </书信>
25	书信落款	letter-signature	书信的落款。	用于标注书信落款（写信人+时间）。 <书信> <书信标题>手连手心连心</书信标题> <开头称谓>亲爱的唐冉</开头称谓> <问候语>你好！</问候语> <书信内容> 在这寒假之前，我们学校举行了一次“手拉手”的活动…… …… </书信内容> <结束语>身体健康，学习进步！</结束语> <书信落款>你的朋友陈研</书信落款> <书信落款>2018年2月26日</书信落款> </书信>
26	诗词	poem	文学创作的具备特殊格式的文本内容	用于标注期刊中出现的诗词，单句出现在正文中的不需要标诗词。 <诗词> <诗词标题>短歌行</诗词标题> <诗词作者>曹操</诗词作者> <诗词内容> 对酒当歌，人生几何！ 譬如朝露，去日苦多。 </诗词内容> </诗词>
27	诗词标题	poem-title	诗词的标题。	用于标注诗词标题。 <诗词> <诗词标题>短歌行</诗词标题> <诗词作者>曹操</诗词作者> <诗词内容> 对酒当歌，人生几何！ 譬如朝露，去日苦多。

序号	中文名称	英文标签	标签释义	标签用法示例
				</诗词内容> </诗词>
28	诗词内容	poem-body	诗词的主体内容。	用于标注诗词内容。 <诗词> <诗词标题>短歌行</诗词标题> <诗词作者>曹操</诗词作者> <诗词内容> 对酒当歌，人生几何！ 譬如朝露，去日苦多。 </诗词内容> </诗词>
29	诗词作者	poet	诗词的作者。	用于标注诗词作者。 <诗词> <诗词标题>短歌行</诗词标题> <诗词作者>曹操</诗词作者> <诗词内容> 对酒当歌，人生几何！ 譬如朝露，去日苦多。 </诗词内容> </诗词>
30	标题	title	对期刊文章或内容主题的概括性表述，用于吸引读者和提示内容主旨。	<标题>出版业人工智能应用研究</标题>
31	副标题	subtitle	对主标题加以解释说明的部分	<副标题>——一种基于智能体工作流的出版流程优化方法探索</副标题>
32	一级标题	main-title	文章或期刊中最高层次的标题，用于概括整个章节或主要部分的内容。在层级标题体系中，它位于最上方，下面可以有二级、三级等更低层次的标题。	用于标注正文中的一级标题。 <正文内容> <一级标题>…</一级标题> <分段>…</分段> <二级标题>…</二级标题> <分段>…</分段> </正文内容>
33	二级标题	secondary-title	在一级标题之下的标题，用于进一步细分章节或段落的内容。它直接隶属于一级标题，并为其下的三级标题提供上下文。	用于标注正文中的二级标题。 <正文内容> <一级标题>…</一级标题> <分段>…</分段> <二级标题>…</二级标题> <分段>…</分段> </正文内容>
34	三级标题	tertiary-title	内容组织中的一个较细分的层级，通常用于进一步划分二级标题下的内容。	用于标注正文中的三级标题。 <正文内容> <三级标题>…</三级标题> <分段>…</分段> </正文内容>
35	四级标题	quaternary-title	对三级标题内容的进一步细分。当三级标题下的内容仍然需要更细致的划分时，就会用到四级标题。这种层级的标题有助于读者深入具体的细节中，理解每一部分的具体内容。	用于标注正文中的四级标题。 <正文内容> <四级标题>…</四级标题> <分段>…</分段> </正文内容>
36	五级标题	quinary-title	相对较低的层级，主要用于极其详细的内容划分。	用于标注正文中的五级标题。 <正文内容> <五级标题>…</五级标题> <分段>…</分段> </正文内容>
37	六级标题	senary-title	标题层级中的最低一层，通常用于极端详细的内容划分。	用于标注正文中的六级标题。 <正文内容> <六级标题>…</六级标题> <分段>…</分段> </正文内容>
38	分段	p	构成文章的基本单位，通常由几句话或一组句子组成，用于表达一个相对完整的思想或意义。段落之间通常有	用于标注文中出现的文段内容。 <分段> 时令已快到惊蛰，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。</分

序号	中文名称	英文标签	标签释义	标签用法示例
			明显的分隔标记，如换行或缩进等。	段>
39	作者	creator	创作作品的个人或团体。	<文章> ...<作者>张×</作者>...</文章>
40	第一作者单位	unit	第一作者所属的单位。	<单位>北京大学信息管理系</单位>
41	摘要	abstract	对文章内容的简短概括，能让读者快速了解文章的核心内容和主要观点。	<文章> ...<摘要> 本文探讨了在数字人文背景下，古籍数字化标引过程中的标准选择与自定义标签设计问题，提出了一套基于中文 XML 标签的实践方案，有效降低了标引门槛并提升了数据可操作性。</摘要>...</文章>
42	关键词	keyword	文章中具有关键意义的词汇，用于概括文章的核心主题和主要内容。	<文章> ...<关键词>古籍数字化；XML 标引；元数据标准；数字人文</关键词>...</文章>
43	分类号	classification-code	对文章进行分类的代码，用于标识文章所属的学科或领域。	<文章> ...<分类号>G255. 9</分类号>...</文章>
44	日期	date	文章发表或相关事件发生的时间。	<文章> ...<日期>2023-10-25</日期>...</文章>
45	注释编号	annotation-number	对期刊的某些内容或文字所作说明的编号。	<文章> ...<分段> 时令已快到惊蛰<注释编号>⑥</注释编号>，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。 </分段>...</文章>
46	注释内容	annotation-content	对期刊的某些内容或文字所作的说明。	用于标注对于某些词句的解释说明相关内容。 <注释内容> <注释编号>⑥</注释编号>惊蛰，是二十四节气中的第三个节气。斗指丁，太阳到达黄经 345°，于公历 3 月 5—6 日交节。 </注释内容>
47	链接	notelink	注释的链接。	用于建立正文与注释区的跳转关联。 <分段>时令已快到惊蛰<注释编号><链接 target="#note-06">⑥</链接></注释编号>，雪当然再不会存留。</分段>

B. 3. 5 样式标签使用规则

样式标签用于标注期刊文本内容中的特殊展现样式，包括斜体、粗体、居左、居中、居右等，在实际标注过程中需配合结构标签和呈现标签使用，在结构标签和呈现标签标注的内容中，嵌套样式标签使用。

表 B. 12 期刊样式标签使用规则

序号	中文名称	英文标签	标签释义	标签用法示例
1	上标	superscript	比同一行中其他文字稍高的文字，通常作为一种上角标志的符号。它在期刊中主要有以下几种用途：脚注或引用标记、数学和科学表达、特殊意义的扩展代号。	用于标注文中展示为上标的文字。 <数学公式> a<上标>2</上标>-b<上标>2</上标> =(a+b) (a-b) </数学公式>
2	下标	subscript	出现在正常字体下边的数字、字母或其他标志，常用于期刊中的公式、数学表达式或化学复合物的描述。其主要用途包括：数学和科学表达、化学式。	用于标注文中展示为下标的文字。 <化学公式> 2H<下标>2</下标>O→2H<下标>2</下标>+O2 </化学公式>
3	粗体	bold	一种文字排版方式，通过加粗文字的字形来强调或突出某些内容。在文档、网页或印刷品中，粗体常用于标题、关键词或需要特别强调的文本。	用于标注文中粗体的文字。若整段粗体的话不需要再有分段标签。 <粗体>时令已快到<斜体>惊蛰</斜体>，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。</粗体>

序号	中文名称	英文标签	标签释义	标签用法示例
4	斜体	italic	一种文字排版方式，通过将文字倾斜来呈现不同的视觉效果。斜体常用于表示书名、人名、引文或需要稍微强调但不至于过于突兀的文本。	用于标注文中斜体的文字。若整段为斜体的话不需要再有分段标签。 <粗体>时令已快到<斜体>惊蛰</斜体>，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。</粗体>
5	下划线	underline	在文字下方添加的一条直线，用于强调、标注或链接文本。在文档中，下划线常用于表示超链接、拼写错误或需要特别注意的文本。	用于标注文中带下划线的文字。若整段带下划线的話不需要再有分段标签。 <粗体>时令已快到<下划线>惊蛰</下划线>，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。</粗体>
6	删除线	strike-through	在文字上画一条线，用于表示该文字已被删除或不再适用。在文档编辑中，删除线常用于标记需要删除的文本内容。	用于标注文中带删除线的文字。若整段到删除线的话不需要再有分段标签。 <粗体>时令已快到<删除线>惊蛰</删除线>，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。</粗体>
7	上着重	super-emphasis	对文本中的某些内容进行特别强调或突出。除了使用粗体、斜体或下划线等排版方式外，还可以通过添加特殊标记来实现着重效果。	用于标注文中带着上重号的文字。若整段为上着重的话不需要再有分段标签。 <粗体>时令已快到<上着重>惊蛰</上着重>，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。</粗体>
8	下着重	sub-emphasis	对文本中的某些内容进行特别强调或突出。除了使用粗体、斜体或下划线等排版方式外，还可以通过添加特殊标记来实现着重效果。	用于标注文中带着下重号的文字。若整段为下着重的话不需要再有分段标签。 <粗体>时令已快到<下着重>惊蛰</下着重>，雪当然再不会存留，往往还没等落地，就已经消失得无踪无影了。</粗体>
9	居左	left-align	将元素向页面的左边靠齐，从而在视觉上形成整齐有序的布局。	用于标注文中居左展示的文字。 <居左>你好</居左>
10	居中	center-align	将元素放置在页面的中心位置，使左右两侧的空间相等。	用于标注文中居中展示的文字。 <居中>你好</居中>
11	居右	right-align	在设计中将元素放在画面的右侧，使得页面看起来更加平衡。	用于标注文中居右展示的文字。 <居右>你好</居右>
12	文字边框	border	带文字边框的段落或内容。	用于标注带文字边框的内容。 <文字边框> 语句（1）（2）中含有变量…… </文字边框>
13	段落底色	bg-color	带底色的段落或内容。	用于标注带底色的内容。 <分段> <段落底色>语句（1）（2）中含有变量…… </段落底色> 在无法判断它们的真假 …… </分段>
14	小字	small-style	字体小于周围正常文字的内容。	用于标注字体小于周围正常文字的内容。 <分段> <小字>在</小字> 这寒假之前…… </分段>
15	大字	big-style	字体大于周围正常文字的内容。	用于标注字体大于周围正常文字的内容。 <问候语> <大字>亲爱</大字> 的唐冉： </问候语>

B.3.6 辅助标签使用规则

辅助标签中设计语种标签，同图书同名标签。

附 录 C
(资料性)
知识资源数据建设标准

C.1 知识资源数据建设工作内容

知识资源数据建设工作内容宜包括资源筛选、筛选审核、分词选词、词语审核、词间关系、关系审核、知识标引、标引审核。

C.2 知识资源数据要求

C.2.1 关联体系数据要求

C.2.1.1 词库要求

应遵循分词选词、词语审核流程要求完成规范化词库建设：

- a) 分词选词：应采用人工标注或技术分词方式，对版式文件标准数据或结构化文件中的核心词汇、专业术语等词汇进行选词及确词；
- b) 词语审核：应对分词选词结果进行规范性、准确性、完整性审核，剔除冗余词汇、修正错误词汇。

C.2.1.2 词间关系要求

应遵循关系构建、关系审核流程要求，建立词汇间标准化逻辑关系关联：

- a) 关系构建：应基于审核通过的词库词汇，采用人工定义或技术算法方式，构建词汇间标准化语义关系；
- b) 关系审核：对词间关系构建结果进行合规性、合理性、一致性审核，确保关系定义准确、层级清晰、逻辑严谨。

C.2.2 关联关系数据要求

应基于已构建的标准化知识体系数据，对版式文件标准数据及结构化文件中的知识元进行标引，建立标准化关联映射，应遵循以下流程：

- a) 知识标引：应通过人工或技术方法，基于构建完成的规范化词库对文件中的知识元进行精准标引，建立知识元与知识体系词汇间的一一对应或多元关联关系；
- b) 知识审核：应对标引结果进行全面审核，验证标引对象、标引词汇、关联关系的准确性与一致性，确保知识标引结果符合知识标引技术要求，保障语义关联的有效性与可用性。

C.2.2.1 知识标引位置要求

C.2.2.1.1 主题标引位置

主题标引的位置在内容分隔符的前后，通常在较明显和突出的位置，包括但不限于：

- a) 章节标题前后；
- b) 文中小标题前后；
- c) 小序号前；
- d) 重要语句前后。

来源：[GB/T 45257 8.6.1.1]

C.2.2.1.2 扩展标引位置

扩展标引的位置通常为：

- a) 正文：正文为扩展标引的常见标引位置，通常标引在“知识概念”后；
- b) 段末：段末标引该段中的核心知识点，通常标引在段落后；
- c) 文末：文末标引本篇的核心知识点，通常标引在文末。

[来源：GB/T 45257 8.6.1.2]

C.2.2.1.3 关联标引位置

关联标引的位置通常位于所需关联内容要素的附近，如图片、引文、公式后等。无特定关联要素的关联标引也可标引在文后。

[来源：GB/T 45257 8.6.1.3]

C.2.2.2 知识标引密度要求

C.2.2.2.1 主题标引密度

主题标引过程中，内容的一个主题可标引一个或多个知识元，不宜超过 3 个。

C.2.2.2.2 扩展标引密度

扩展标引过程中，内容的一个概念后可标引多个知识元，不宜超过 4 个。内容中的概念在 300 字内重复出现时不重复标引。

[来源：GB/T 45257 8.6.2.2]

C.2.2.2.3 关联标引密度

关联标引过程中，多个关联信息可标引在同一位置，每 300 字内关联标引个数不宜超过 5 处。

[来源：GB/T 45257 8.6.2.3]

C.3 知识资源数据质量要求

C.3.1 关联体系质量要求

C.3.1.1 词库质量要求

在相对独立内容范围内的提词数量规律如下：

- a) 核心词：核心词数量不少于 1 个，宜为 1 个～4 个；
- b) 辅助词：一个核心词涉及的辅助词数量宜为 1 个～4 个；
- c) 相关词：提取数量不做具体约束。平均每个章节提出的词汇的数量宜控制在 5 个～20 个，特殊情况可提取更多。

C.3.1.2 词间关系质量要求

词间关系构建的密集程度宜在 10 个/千字～15 个/千字。

C.3.2 关联关系质量要求

关联关系标引的质量指标如下：

- a) 标引密度：内容资源标引的密集程度的指标；
- b) 标引相关性：内容资源标引的相符度和准确度的指标；

c) 标引位置准确性：内容资源标引的位置准确性的指标。
标引质量基本要求见表 C.1。

表 C.1 标引质量基本要求

标引类型	标引密度 (应不低于最低值)	标引相关性		标引位置准确性
		相符度	准确度	
文字	2 个/千字~10 个/千字	100%	不低于 90%	不低于 90%
图片	1 个/幅~10 个/幅	100%	不低于 90%	不低于 90%
音频	1 个/5 分钟~2 个/5 分钟	100%	不低于 80%	不低于 80%
视频	1 个/5 分钟~2 个/5 分钟	100%	不低于 80%	不低于 80%
模型	1 个/5 兆字节(MB)	100%	不低于 80%	不低于 80%

C.4 高价值语料转换

C.4.1 基础要求

知识资源数据宜转换为高价值语料，可经过资源筛选、清洗等一系列工序，加工分离出特定主题或范围（如垂直领域）的语料，转换为内容及格式符合目标需求的语料，供人工智能进行特定范围内的预训练、微调等任务。针对其他特定训练或应用需求的语料加工，应按照正文第 5 章进行处理。

语料转换可基于知识资源数据及其他已标注数据进行，工序包括资源筛选、筛选审核、人工清洗、清洗质检、人工梳理、数据转换、转换验证。

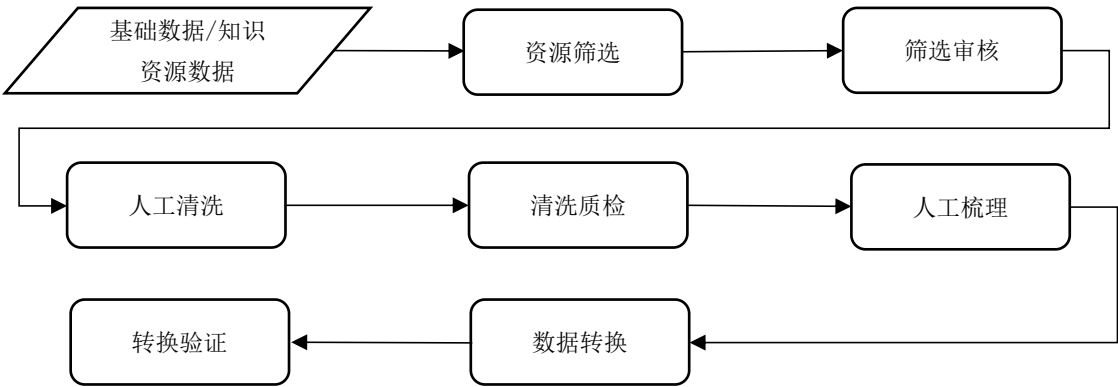


图 C.1 语料转换流程图

C.4.2 数据输入

输入数据应为符合附录 A 规定的 PDF、符合附录 B 规定的 XML 等基础数据，或符合本文件规定的知识资源数据、已完成标注及对齐等语料加工流程的多模态语料等。

C.4.3 语料转换流程

语料转换的工序包括资源筛选、筛选审核、人工清洗、清洗质检、人工梳理、数据转换、转换验证。

C.4.3.1 资源筛选

资深编辑依据出版物内容资源及知识资源等元数据进行批量初选，剔除明显不符合领域要求的资源。

C.4.3.2 筛选审核

由领域专家对初筛资源的目录及样章进行实质性审查，确认知识体系的完整性、权威性及适用性。专家需对确认入库的资源进行“核心知识域”与“一般背景域”的划分。

C.4.3.3 人工清洗

按照数据交付需求，对经过审核的数据进行逻辑性清洗，解决数据规范性问题：

- a) 标签清洗指按交付需求对标签进行必要的整理、增删和映射；
- b) 实体消歧指按交付需求对同名异义或异名同义的知识元进行统一与修正；
- c) 边界整理指按交付需求整理语料边界，确保该领域语料的完整性。

C.4.3.4 清洗质检

对清洗后的数据进行逻辑一致性检查。质检应采用规则校验与人工抽检相结合的方式进行：

- a) 规则校验：使用 Schema 等校验工具检测数据完整性与属性合法性；
- b) 人工抽检：重点检查该领域语料边界内容的逻辑合理性，如知识元、词间关系、知识关联资源是否缺失。

C.4.3.5 人工梳理

在清洗的基础上，由领域专家对语料内容进行核查，通过符合交付要求的方式对领域知识进行必要的补充、修正，如基于样本均衡性要求进行样本补充等。

C.4.3.6 数据转换

将经过人工梳理的结构化数据，基于特定转换逻辑和格式要求转换为适配大模型训练的数据文件。

C.4.3.7 转换验证

对转换完成的语料数据进行质量验证，包括数据完整性验证、内容质量验证及功能可用性测试。验证结果应符合第6章要求。

C.4.4 数据输出

C.4.4.1 文件格式

成品数据宜采用 JSON、JSON Lines(jsonl)、CSV、TSV、LaTeX、MathML、Markdown、XML、Html、Apache Parquet 或纯文本(.txt) 等格式。

C.4.4.2 字段定义

宜采用必要的字段定义，包括唯一标识符、名称、内容、属性、原始资源 URI、元数据等。

CY/T XXX—XXXX

C.4.4.3 编码

应采用 UTF-8 编码。

C.4.5 质量控制

按本文件第 6 章执行。

参考文献

- [1] GB/T 18793—2002 信息技术 可扩展置标语言（XML）1.0
- [2] GB/T 31219.4—2014 图书馆馆藏资源数字化加工规范 第4部分：音频资源
- [3] GB/T 31219.5—2016 图书馆馆藏资源数字化加工规范 第5部分：视频资源
- [4] GB/T 35273—2020 信息安全技术 个人信息安全规范
- [5] GB/T 36344—2018 信息技术 数据质量评价指标
- [6] GB/T 37988—2019 信息安全技术 数据安全能力成熟度模型
- [7] GB/T 42755—2023 人工智能 面向机器学习的数据标注规程
- [8] GB/T 45577—2025 数据安全技术 数据安全风险评估方法
- [9] YD/T 4245—2023 电信网和互联网数据脱敏技术要求和测试方法
- [10] GF 0031—2026 人工智能 语料库 基础术语
- [11] CY/T 101.8—2014 新闻出版内容资源加工规范 第8部分：图书加工
- [12] CY/T 101.9—2014 新闻出版内容资源加工规范 第9部分：报纸加工
- [13] CY/T 101.10—2014 新闻出版内容资源加工规范 第10部分：期刊加工
- [14] CY/T 168.11—2019 新闻出版内容资源加工规范 第11部分：音频加工
- [15] CY/T 168.12—2019 新闻出版内容资源加工规范 第12部分：视频加工
- [16] GY/T 202.1—2004 广播电视音像资料编目规范 第1部分：电视资料
- [17] GY/T 202.2—2016 广播电视音像资料编目规范 第2部分：音频资料