

CY

中华人民共和国新闻出版行业标准

CY/T XXX.3—XXXX

中文古籍数字出版规范
第3部分：长期存储

Standard for digital publishing of Chinese ancient books—Part 3: Long-term storage

(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 缩略语 1

5 数据类型 1

6 存储原则 2

 6.1 完整性 2

 6.2 有效性 2

 6.3 规范性 2

 6.4 一致性 3

7 存储要求 3

 7.1 接收 3

 7.2 检验 3

 7.3 存储 3

 7.4 复核 3

8 存储管理原则 3

9 存储环境要求 3

10 数据备份策略 4

参考文献 5

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3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3部分：长期存储

1 范围

本文件规定了中文古籍长期存储的数据类型、存储原则、存储要求、存储环境及存储备份策略等。本文件适用于中文古籍数字化资源文件的长期存储。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/Z 23283—2009 基于文件的电子信息的长期保存

GB/T 23286.1—2009 文献管理 长期保存的电子文档文件格式 第1部分：PDF1.4(PDF/A-1)的使用

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.2—XXXX 中文古籍数字出版规范 第2部分：元数据

3 术语和定义

下列术语和定义适用于本文件。

3.1

中文古籍 ancient Chinese books

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。

[来源：WH/T 24—2006，3.1，有修改]

3.2

长期存储级数字资源 long-term conservation digital resource

作为档案存储及出版用的数字资源，不用于发布服务，可作为格式转换和复制的母本。

3.3

分辨率 Dots Per Inch; DPI

图像每英寸长度内的像素点数。

3.4

无损压缩 loss less compression

无失真压缩

利用数据的冗余进行压缩，可完全恢复原始数据而不引起任何失真。

4 缩略语

下列缩略语适用于本文件。

PDF：可携带文档格式（Portable Document Format）

TIFF：标签图像文件格式（Tag Image File Format）

XML：可扩展标记语言（Extensible Markup Language）

5 数据类型

长期存储的中文古籍数字出版数据类型及存储格式主要包括。

- a) 古籍元数据：包括用于中文古籍数字出版的描述元数据和管理元数据，宜采用 XML 格式存储。
- b) 古籍内容结构化信息：包括内容组成结构和各级结构的内容及属性，宜采用 XML 格式存储。
- c) 古籍版式结构化信息：包括古籍版式要素组成结构和各级结构坐标位置等信息，宜采用 XML 格式存储。
- d) 固定版式数据：包括固定版式信息以及专用字库和色彩空间信息等数据，宜采用 PDF 格式存储。
- e) 图片对象数据：包括古籍的插图、生僻字图和表格等对象数据，宜采用 TIFF 格式存储。
- f) 说明注释类文件：包括数字化加工过程中原始文献内容缺失、无法辨识等情况的记录数据，宜采用 TXT 格式存储。

6 存储原则

6.1 完整性

长期存储数据应遵循完整齐备原则，主要包括。

- a) 内容完整，长期存储数据应涵盖古籍的所有部件内容及相应的元数据。
 - 1) 页面版式信息完整，包括封面、书牌、封底、书名页、正文页、有板框的空白页、书耳、天头、天头字、地脚、象鼻、鱼尾、版心、印章、板框、界行等。
 - 2) 古籍内容完整，包括牌记、序跋、正文、题记、批校文字等。
 - 3) 古籍附件完整，包括夹签、活页、附件等。
 - 4) 元数据应包含古籍描述元数据和管理元数据。
- b) 文件齐备，即与中文古籍数字出版有关的数据文件应齐备。

6.2 有效性

长期存储数据文件应遵循有效可用原则，主要参考标准GB/Z 23283—2009的要求，具体包括。

- a) 文件格式版本有效：文件格式版本更新有序，向前兼容。
- b) 文件存储机制有效：文件存储具有有效的容错及校验机制，有明确的数据刷新存储机制及存储设备故障淘汰机制。
- c) 文件具有良好的可迁移性：与设备、操作系统和硬件平台无关，使用与平台无关的压缩算法、容错机制等。

6.3 规范性

长期存储数据文件应遵循规范性原则，主要包括。

- a) 数据规格要求如下。
 - 1) 文本类数据规格：长期存储文本类数据应采用 PDF、XML、TXT 文件格式，PDF 文件应符合 GB/T 23286.1—2009 第 1 部分的规定，不存在编码混乱、图像失真、精度不足、关联错误等问题；XML 文件应遵循 XML1.0 以上版本规范。
 - 2) 图像类数据规格：长期存储图像类数据应采用 TIFF5/TIFF6 等以上版本格式，文件规格要求见表 1。

表1 中文古籍数字出版长期存储图像类数据规格要求

序号	图像分辨率 (DPI)	色彩位深	文件格式压缩算法
1	不低于 300	8 位、24 位或更高	TIFF5/TIFF6 不压缩或无损压缩

- b) 元数据：元数据应符合 CY/T XXX. 2—XXXX 的规定。
- c) 命名要求：长期存储的中文古籍数字出版数据文件的名称应由以下 3 部分组成。
 - 1) 名称前缀应由数据类型、数据格式的两级命名段组成，两段之间应采用英文半角“.”分隔。
 - 2) 前后缀间的连接符使用“-”。
 - 3) 名称后缀应为数据编号。

具体命名规则见表 2。

表2 具体命名规则

序号	类别	命名规则
1	数据类型	两位数字组成，从 01 记起，可参照如下数字顺序命名： a) 古籍元数据：01； b) 古籍内容结构化信息：02； c) 古籍版式结构化信息：03； d) 固定版式数据：04； e) 图片对象数据：05； f) 说明注释类文件：06。
2	数据格式	三位数字组成，从 001 记起，可参照如下数字顺序命名： a) PDF：001； b) XML：002； c) TIFF：003。
3	数据编号	六位数字组成，从 000001 记起，在同一数据类型、数据格式、数据资源类型下按顺序编号命名

6.4 一致性

长期存储数据文件应遵循一致性原则，主要包括。

- a) 版本一致：数据文件的各个版本与不同版本、版次的古籍内容保持一致。
- b) 内容一致：数据文件中的文字、符号、图片、位置、版式等与古籍内容保持一致。
- c) 存储方法一致：存储的内容文件应有统一的存储结构、命名规则和检验方法。

7 存储要求

7.1 接收

接收需长期存储的数据，应进行分类、核对、检查、登记等过程，同时形成规范完整的记录。

7.2 检验

中文古籍数字出版长期存储数据的检验应遵循GB/T 38548.5—2020的规定。

7.3 存储

存储时应在同一存储空间按照统一的命名规则进行存储，应定期对文件进行备份，防止文件损坏、丢失。

7.4 复核

对存储介质、长期存储数据的有效性进行定期的检查，在发现问题时及时采取备份数据进行恢复或更换存储介质。

8 存储管理原则

数据存储管理原则主要包括：

- a) 长期存储数据及其介质应指定专人负责保管；
- b) 长期存储数据应按照备份策略定期、完整、真实、准确地转储到永久性存储介质上，并按照命名规范为存储介质编码并进行明显标识；
- c) 长期存储介质应分布式存储，按照各系统规定的保存期限存放。

9 存储环境要求

长期存储的环境基本要求包括以下内容：

- a) 宜采用硬盘存储介质进行长期存储；

- b) 存储介质应在指定场所保管，应具备防火、防热、防潮、防尘、防磁、防盗设施；
- c) 存储介质应远离腐蚀气体、易燃易爆物；
- d) 存储环境的温度和湿度应严格控制；
- e) 存储环境内的静电电位不应过大，避免损坏存储介质。

10 数据备份策略

宜采用完全备份与增量备份相结合的策略，短周期进行增量备份，长周期进行完全备份。

参 考 文 献

- [1] WH/T 24—2006 图书馆古籍特藏书库基本要求
-