ICS 35.240
L 70

中华人民共和国国家标准

GB/T 31219.2—2014

图书馆馆藏资源数字化加工规范 第2部分:文本资源

Specification of library collections digitization—Part 2:Text resources

2014-09-30 发布　　2015-01-01 实施

中华人民共和国国家质量监督检验检疫总局
中国国家标准化管理委员会　发布

目　　次

前言 …… Ⅰ
1　范围 …… 1
2　规范性引用文件 …… 1
3　术语和定义 …… 1
4　加工级别及内容编码 …… 2
5　加工准备 …… 3
6　资源采集与处理 …… 3
7　元数据加工 …… 4
8　命名规则 …… 6
9　质量管理 …… 6
参考文献 …… 7

前　言

GB/T 31219《图书馆馆藏资源数字化加工规范》分为五个部分：

——第1部分：总则；

——第2部分：文本资源；

——第3部分：图像资源；

——第4部分：音频资源；

——第5部分：视频资源。

本部分为GB/T 31219的第2部分。

本部分按照GB/T 1.1—2009给出的规则起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本部分由中华人民共和国文化部提出。

本部分由全国图书馆标准化技术委员会(SAC/TC 389)归口。

本部分起草单位：国家图书馆、首都图书馆、北京大学图书馆、中国科学院文献情报中心、上海图书馆上海科学技术情报研究所、浙江大学图书馆、汉王科技股份有限公司、北京方正阿帕比技术有限公司。

本部分起草人：李晓明、龙伟、赵四友、朱云、陈建新、王炜、张春红、刘秀文、张建勇、周静怡、徐强、黄晨、李明敬、魏丕。

图书馆馆藏资源数字化加工规范
第2部分:文本资源

1 范围

GB/T 31219的本部分规定了图书馆文本资源数字化加工遵循的技术标准。

本部分适用于以文字为主要表达形式,可存在少量图表的文本文献(不包括古籍善本、手稿等特殊文献)的数字化加工。

注:数字化加工对象可以是一般印刷型文献,也可以是印刷型文献经过数字转换后的图像文件。

本部分适用于图书馆文本资源数字化加工,其他文献信息机构的文本资源数字化加工也可参照使用。

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。凡是注日期的引用文件,仅注日期的版本适用于本文件。凡是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 2312 信息交换用汉字编码字符集 基本集

GB/T 4894—2009 信息与文献 术语

GB 13000 信息技术 通用多八位编码字符集(UCS)

GB 18030 信息技术 中文编码字符集

GB/T 25100—2010 信息与文献 都柏林核心元数据元素集

ISO/IEC 10646 信息技术 通用多八位编码字符集(UCS)[Information technology—Universal Multiple-Octet Coded Character Set (UCS)]

3 术语和定义

下列术语和定义适用于本文件。

3.1

文献 document

在文献工作过程中作为一个单位处理的记录信息或实物对象。

[GB/T 4894—2009,定义4.1.2.2]

3.2

文本 text

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

[GB/T 4894—2009,定义4.1.1.2.4]

3.3

图像 image

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

3.4

光学字符识别　optical character recognition

又称 OCR 识别，自动识别通过扫描仪、数码相机、摄像机等得到的图像中的字符，便于存储、编辑和检索。

3.5

点/英寸　dots per inch

dpi

扫描仪(打印机)在水平方向上和垂直方向上的每英寸都能扫描(打印)的点数。

[GB/Z 19736—2005，定义 3.4]

4　加工级别及内容编码

4.1　加工级别

文本资源数字化加工级别分为长期保存级和发布服务级：

——长期保存级。用于文本资源的长期保存，在必要时用于编辑及格式转换。长期保存级的文件格式主要有：

- XML 格式，适用于标识文件的版面信息，描述文件的内容或结构。
- TXT 格式，是最常见的一种文本格式，其文件体积小，存储方便，不易被病毒感染。
- PDF 格式，适用于各种档次的印刷，文本文档的保护、打印、网络显示及长期保存等。

——发布服务级。用于网络浏览、下载及打印。发布服务级的文件格式主要有：

- HTML 格式，一般用于文本资源的网络发布。
- PDF 格式，也适用于文本文件的交换、显示。
- DOC 格式，是一种专属格式，一般用于文本编辑。

4.2　内容编码

文本内容编码应遵循通用的国家标准或国际标准，见表 1。

表 1　文本内容编码标准

标准编号	标准名称	简要说明
GB 2312	信息交换用汉字编码字符集　基本集	规定了汉字信息交换用的基本图形字符及其二进制编码表示。它是一个简化字汉字的编码，共收录 6 763 个汉字，其中一级汉字 3 755 个，二级汉字 3 008 个。
GB 18030	信息技术　中文编码字符集	规定了信息技术用的中文图形字符及其二进制编码的十六进制表示，它是以汉字为主并包含中国多种少数民族文字的超大型中文编码字符集标准，共收录 70 244 个汉字。
GB 13000	信息技术　通用多八位编码字符集(UCS)	规定了 UCS 的总体结构。其编码空间巨大，可以容纳多种文字同时编码，共收录汉字 20 902 个。
ISO/IEC 10646	信息技术　通用多八位编码字符集(Information technology—Universal Multiple-Octet Coded Character Set)	ISO/IEC 10646 标准由国际标准化组织颁布，简称 UCS，用来实现全球所有文种的统一编码。其基本级收录 20 902 个汉字，扩充 A 6 582 个汉字，扩充 B 47 211 个汉字，已有汉字编码超过 7 万个。UCS 与 Unicode 在字符编码上保持一致。

表 1（续）

标准编号	标准名称	简要说明
ASCII	美国信息交换标准码（American Standard Code for Information Interchange）	美国国家标准学会（American National Standard Institute，ANSI）制定的标准的单字节字符编码方案，主要用于显示现代英语和其他西欧语言。ASCII 码使用指定的 7 位或 8 位二进制数组合来表示 128 种或 256 种可能的字符。标准 ASCII 码也叫基础 ASCII 码，使用 7 位二进制数来表示所有的大写和小写字母、数字 0～9、标点符号，以及在美式英语中使用的特殊控制字符。

5 加工准备

在文本资源数字化加工之前应做好以下准备工作：

——加工环境。根据文本资源的类型及数字化加工任务量合理配置相应的软硬件设施，这些设施在功能性、可用性、安全性方面宜满足加工要求。

——数据查重。针对加工对象检查已有的对象数据和元数据，应尽量利用已有的数据，尽量避免重复加工。

——文献保护。根据文献的状况采取适当的保护措施，文本资源数字化加工过程中应尽量减少对文献的损害。

——加工对象。文本资源数字化加工应优先选择文本文献的数字化图像作为加工对象，没有数字化图像的文本文献，可先通过扫描或拍照等数字化手段加工成数字对象，或者直接通过键盘录入文本文献内容。

6 资源采集与处理

6.1 文本资源采集方式

文本资源采集方式主要包括文本录入和光学字符识别。文本录入适合处理字体过小、图文模糊、版面复杂的文本文献；光学字符识别适合处理文字规整、版面清晰的文献。

6.2 文本录入

6.2.1 录入要求

文本录入应遵守以下要求：

——文本应按照内容的逻辑顺序进行录入，如一个表格或者分栏的文本应以单元格或栏目顺序为单位进行录入，而不是逐行录入；

——文本录入时应照实录入，保留原始文献中的错别字及各种文字变体。

6.2.2 校对要求

录入的文本通过校对来保证内容的正确率，以满足质量要求。以下校对方法可以结合运用：

a） 编辑软件自带校对功能，能够提供语法检查及拼写检查之类的错误提示功能。

b） 采用双工录入。一般推荐采用不同的输入法进行录入，再通过对比校对，对差异部分进行人工干预纠正错误。

注：双工录入，即同一份文字资料由两个操作员分别进行录入。

6.3 光学字符识别

6.3.1 图像质量要求

用于光学字符识别的图像应满足下列质量要求：

——适合中文识别软件的图像分辨率应为 150 dpi～300 dpi，拼音文字的图像分辨率应为 300 dpi～400 dpi。

——内容完整，无残破和缺失。图像应与文献原件完全对应，保证页面内容完整，并且无多页或少页。

——颜色深浅适中，字迹清晰。图像不能出现文字缺失或文字重叠，存在透字（因用纸较薄或字迹颜色过重）的页面图像需保证正面文字的清晰。

——酌情对图像去污，噪点、阴影、黑线不影响正文文字。因扫描过程而导致图像上出现的黑线、污点、阴影和指印不能影响正文内容的辨识和阅读。

——无明显倾斜和扭曲。图像倾斜不应超过 3°；在原件页面内空较小或原件较厚导致扫描图像边缘的扭曲、文字变形的情况下，扫描图像也必须确保页面文字能够识别。

6.3.2 字符识别率要求

满足图像质量要求的字符识别正确率应达到 95％以上。

6.3.3 识别校对

识别校对有多种方法，各种校对方法需结合运用，以满足正确率及效率的要求。常用的校对方法有：

——纵向校对。将一个图像或若干个图像中识别成同一个字符的图像列在一起显示，便于发现错误和修改。纵向校对可在校对界面直接修改。校对工作中不必通览全文，就可以方便快捷地进行校对修改工作。

——横向校对。是一种逐行逐字地把识别文本与相应图像作对照的校对方式。横向校对通常是纵向校对工序之后的又一道校对工序，它利用上下文信息对识别文本进行判断。

——对比校对。对同一文本的两遍校对结果进行比对，对差异部分进行干预纠正。

7 元数据加工

7.1 元数据著录

元数据著录应与文本内容、文本文件相关联并成为数字化加工过程中的一部分。通过对文本文件进行分析、选择和记录，产生描述元数据、结构元数据、管理元数据等。其中描述元数据著录可依据 GB/T 25100—2010；结构元数据著录内容包括但不限于文本目录体系、目录和正文的链接信息、文本版式信息以及该文本与其他相关文本的关联信息；管理元数据标记文本资源管理及加工过程中所涉及的管理信息及技术信息，管理元数据著录内容包括但不限于表 2 所列。

表 2　管理元数据著录项目

内容名称	标签	定义	注释
infoResourceIdentifier	信息资源标识符	唯一识别信息资源的标识	一般是特定应用系统内具有唯一识别性的标识符号。可由标识应用系统的前缀(即标识符的类型)与一字符串(即标识符的值)组成。可由系统自动产生或由人工赋予
source	来源	对生成本信息资源的资源或其他实体的参照	可用正式标识体系的字符串表示,如 URI 以及其他标识非数字资源的编码体系(ISBN,ISSN 等)。对于派生自其他数字资源的数字资源来说,其来源信息一般在资源的描述元数据中反映,如 DC 的“关联(relation)”
technicalInfo	技术信息	与信息资源的创建、加工、使用相关的物理参数、技术手段与标准以及硬件环境	可嵌入或链接通行的技术元数据
format	格式	信息资源的物理或数字表现形式	信息资源的内容形式,包括资源内容与其元数据的类型
processingMode	加工方式	数字信息资源的加工方式	如:OCR、文本录入等
characterSet	字符集	信息资源采用的字符集的名称	通常采用的编码标准有 ASCII 、GB 18030 等
Identification Tool	识别工具	识别文本资源的软件	识别工具的名称
agentIdentifier	代理标识符	唯一识别代理的标识	一般是特定应用系统内具有唯一识别性的标识符号,建议由系统自动生成
agentName	代理名称	代理的名称,包括职称、所在单位等	可以是个人、团体或者自动装置。代理可以没有名称或可以不同步其名称
agentType	代理类型	根据代理的定义对其划分的基本大类	建立受控词汇表,规范类型的取值,建议为:个人、团体、软件
eventIdentifier	事件标识符	唯一识别事件的标识	主要用于管理元数据记录内部关联元素之间的链接,如代理与其相关的事件,建议由系统自动生成
eventType	事件类型	根据信息资源生命周期的基本阶段对事件划分的大类	包括:创建、数字化、元数据加工等
action	操作	一个事件中有特定意义的细分的行动	包括复制、修改、删除、合并等,建议由系统自动从日志文件中获取
actionDateTime	操作日期时间	操作发生的日期时间或日期时间的范围	如文档创建时间、文件处理日期等,建议由系统自动从日志文件中获取

7.2 数据关联

为便于资源的保存和应用，应建立对象数据内部组织结构、对象数据外部组织结构、对象数据与元数据之间、对象数据与原始文献之间的关联。

注：相关资源的关联，可以通过描述元数据中的“Relation”元素描述。

7.3 数据封装

针对由被保存内容及其相关的元数据组成的信息包所进行的整合操作。数据封装内容包括内容信息（文本数字对象和表现信息）、保存描述信息（长期保存需要附加的元数据）、封装信息（信息包的组成部分的关联信息）、描述信息（描述元数据）。比如，文本资源的数据封装内容包括元数据、对象文件、目录（标签）跳转链接、说明文件等各类相关文件。

封装的信息包要求内部数据关联关系明确，数据封装和数据解析规则明确，不依赖于特定的软件或硬件平台。

8 命名规则

文本资源文件命名应遵循以下规则：

——拥有唯一标识符，不能与其他资源标识符重复。

——文件名定义应清晰明确，以便于文件名的标准化与统一管理。

——具备长期可用性。文件命名方式不依赖于某种处理或者系统。文件名包含的信息不应随着时间的推移而改变。

——严格遵守技术限制。符合计算机系统对文件名中特殊字符、空格、日期等字符使用的限制，以及文件名字符长度的限制。

——文件扩展名采用三位字符，字母用小写形式。

——文件名中引用的元数据应另有文件记录，以避免文件跨系统转移时元数据受到损害。

9 质量管理

9.1 质量要求

文本资源数字化加工应达到但不限于以下质量要求：

——文本资源数字化内容忠实于原文献，完整有序；

——元数据著录项目完整，著录信息准确；

——文件格式与编码无误；

——字符的错误率不超过 0.3‰；

——文件夹和文件名命名正确；

——存储文件能正常读取，不携带病毒，介质标识正确。

9.2 过程管理

建立安全管理机制。在文本资源加工各个环节中，确保文本资源安全、有序并及时做好数据备份。

建立文档管理制度。及时撰写、整理和汇总加工过程中的技术与管理文档，在数字化加工完成的同时形成完整、规范的加工记录。

参 考 文 献

[1] GB/Z 19736—2005 电子成像 文件图像压缩方法选择指南

[2] 龙伟,罗云川. 国家图书馆文本数据加工标准和操作指南. 北京:国家图书馆出版社,2012.

[3] 章毓晋. 图像工程(上册):图像处理.第3版. 北京:清华大学出版社,2012.

[4] 郑巧英,王绍平,汪东波. 国家图书馆管理元数据规范和应用指南. 北京:国家图书馆出版社,2010.

[5] 肖珑,申晓娟. 国家图书馆元数据应用总则规范汇编. 北京:国家图书馆出版社,2011.

[6] 熊武一,等.军事大辞海(下).北京:长城出版社.2000.

[7] 彭绪庶,蒋颖. 资源数字化标准问题研究.北京:北京图书馆出版社,2005.

[8] Federal Agencies Digitization Initiative (FADGI)-Still Image Working Group .Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files[EB/OL].[2013-2-25]. http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf

中 华 人 民 共 和 国
国 家 标 准
图书馆馆藏资源数字化加工规范
第 2 部分:文本资源
GB/T 31219.2—2014
*
中 国 标 准 出 版 社 出 版 发 行
北京市朝阳区和平里西街甲 2 号(100029)
北京市西城区三里河北街 16 号(100045)
网址:www.gb168.cn
服务热线:400-168-0010
010-68522006
2014 年 11 月第一版
*
书号:155066·1-50410

GB/T 31219.2-2014