

国家数字图书馆工程标准规范成果

古籍用字（包括生僻字、避讳字） 属性字典规范和应用指南

张力伟 翟喜奎 主编



國家圖書館出版社

古籍用字(包括生僻字、避讳字) 属性字典规范和应用指南

张力伟 翟喜奎 主编

圖 國家圖書館出版社

图书在版编目(CIP)数据

古籍用字(包括生僻字、避讳字)属性字典规范和应用指南/张力伟, 翟喜奎主编. —北京:国家图书馆出版社, 2010. 10

(国家数字图书馆工程标准规范成果)

ISBN 978-7-5013-4397-3

I. ①古… II. ①张… ②翟… III. ①古籍—异体字—规范—中国
IV. ①G256.1-65 ②H124.3-65

中国版本图书馆 CIP 数据核字(2010)第 134267 号

责任编辑:金丽萍 王涛 高爽

书名 古籍用字(包括生僻字、避讳字)属性字典规范和应用指南

著者 张力伟 翟喜奎 主编

出版 国家图书馆出版社(原北京图书馆出版社)

(100034 北京市西城区文津街7号)

发行 010-66139745 66151313 66175620 66126153

66174391(传真) 66126156(门市部)

E-mail cbs@nlc.gov.cn(投稿) btsfxb@nlc.gov.cn(邮购)

Website www.nlcpress.com→投稿中心

经销 新华书店

印刷 河北三河弘翰印务有限公司

开本 880×1230(毫米) 1/32

印张 3

版次 2010年10月第1版 2010年10月第1次印刷

字数 60千字

书号 ISBN 978-7-5013-4397-3

定价 35.00元

丛书编委会

主 编：国家图书馆

编委会：

主 任：周和平

执行副主任：詹福瑞

副主任：陈 力 魏大威

成 员(按姓氏拼音排名)：卜书庆 贺 燕 蒋宇弘

梁蕙玮 龙 伟 吕淑萍 申晓娟 苏品红

汪东波 王文玲 王 洋 杨东波 翟喜奎

赵 悦 周 晨

本书编委会

主 编：张力伟 翟喜奎

编 委：洪 涛 王晓波 赵 伏 申作宏 富 平
毛雅君 胡昱晓 李 杉 赵 悦

总 序

数字图书馆涵盖多个分布式、超大规模、可互操作的异构多媒体资源库群,面向社会公众提供全方位的知识服务。它既是知识网络,又是知识中心,同时也是一套完整的知识定位系统,并将成为未来社会公共信息的中心和枢纽。数字图书馆建设的最终目标是实现对人类知识的普遍存取,使任何群体、任何个人都能与人类知识宝库近在咫尺,随时随地从中受益,从而最终消除人们在信息获取方面的不平等。“国家图书馆二期工程暨国家数字图书馆工程”是国家“十五”重点文化建设项目,由国家图书馆主持建设,其中国家数字图书馆工程的建设内容主要包括硬件基础平台、数字图书馆应用系统和数字图书馆标准规范体系。

标准规范作为数字图书馆建设的基础,是开发利用与共建共享资源的基本保障,是保证数字图书馆的资源和服务在整个数字信息环境中可利用、可互操作和可持续发展的基础。因此,在数字图书馆建设中,应坚持标准规范建设先行的原则。国家数字图书馆标准规范体系建设围绕数字资源生命周期为主线进行构建,涉及数字图书馆建设过程中所需要的主要标准,涵盖数字内容创建、数字对象描述、数字资源组织管理、数字资源服务、数字资源长期保存五个环节,共计三十余项标准。

在国家数字图书馆标准规范建设中,国家图书馆本着合作、开放、共建的原则,引入有相关标准研制及实施经验的文献信息机构、科研机构以及企业单位承担标准规范的研制工作,这就使得国家数字图书

馆标准规范的研制能够充分依托国家图书馆及各研制单位数字图书馆建设的实践与研究,使国家数字图书馆的标准规范成果具有广泛的开放性与适用性。本次出版的系列成果均经过国家图书馆验收、网上公开质询以及业界专家验收等多个验收环节,确保了标准规范成果的科学性及实用性。

目前,国内数字图书馆标准规范尚处于研究与探索性应用阶段,国家图书馆担负的职责与任务决定了我们在数字图书馆标准规范建设方面具有的责任。此次将国家数字图书馆工程标准规范研制成果付梓出版,将为其他图书馆、数字图书馆建设及相关行业数字资源建设与服务提供建设规范依据,对于推广国家数字图书馆建设成果,提高我国数字图书馆建设标准化水平,促进数字资源与服务的共建共享具有重要意义。

国家图书馆馆长 周和平
2010年8月

前 言

本规范是国家数字图书馆工程标准规范项目研制成果之一。

D001 汉字规范处理项目包含五个部分内容: D001—01 汉字属性字典规范, D001—02 中文文献全文版式还原与全文输入 XML 规范, D001—03 古籍用字规范, D001—04 计算机中文信息处理规范, D001—05 生僻字、避讳字处理规范。本规范涵盖 D001—03 古籍用字规范, D001—05 生僻字、避讳字处理规范。

古籍用字规范(计算机用字标准)是确保数字图书馆工程的重要手段之一。针对数字图书馆古籍文献类型复杂、使用汉字字符数量大等特点,古籍用字规范要求处理的汉字范围大于 GB18030—2005 (UNICODE 5.1、ISO10646—2003)与古籍相关的全部汉字,超出部分为新造字符。

D001—05 生僻字、避讳字处理规范也是确保数字图书馆工程的重要手段之一。针对数字图书馆数字文献类型复杂、使用汉字字符数量大等特点,生僻字、避讳字处理规范要求处理的汉字范围大于 GB18030—2005 (UNICODE 5.1、ISO10646—2003)与生僻字、避讳字相关的全部汉字,超出部分为新造字符。

由于古籍用字与生僻字关系密切,而避讳字亦是古籍用字的一个类型,属性数据处理有许多共同的地方,因而本规范将他们合并在一起,建立一个统一的古籍用字(包括生僻字、避讳字)属性字典。规范则对古籍用字(包括生僻字、避讳字)属性字典的格式及各种属性作出规定。

《古籍用字(包括生僻字、避讳字)属性字典规范》依据 D001—03《古籍用字规范(计算机用字标准)技术规范书》《古籍用字规范(计算机用字标准)需求规格说明书》及 D001—05《生僻字、避讳字处理规范技术规范书》《生僻字、避讳字处理规范需求规格说明书》要求,根据《古籍用字(包括生僻字、避讳字)属性字典》研究编制过程中采用的各项处理原则编制。这些处理原则尽可能地依据了国际与国内的相关标准与规范。对于已有标准与规范所未能涵盖的内容,则说明其处理方略,同时在《古籍用字(包括生僻字、避讳字)属性字典应用指南》中加以详细说明。

《古籍用字(包括生僻字、避讳字)属性字典规范》具体体现于《古籍用字(包括生僻字、避讳字)汉字属性字典》中。该字典包括了:

- (1)古籍用字(包括生僻字、避讳字)属性的完整数据;
- (2)古籍用字(包括生僻字、避讳字)属性字典的文本(PDF 文档);
- (3)古籍用字(包括生僻字、避讳字)属性字典查看和维护工具。

本规范由国家图书馆提出,委托中华书局进行研制。

本规范由中华书局起草,主要起草人为:张力伟、洪涛、王晓波、赵伏、齐蕾。

引 言

随着信息化的发展,各种类型的数字化图书及其他出版物在图书馆及各类读者群中已全面普及,数字图书馆的建设正在有序进行。数字化的中国古籍,将构成中文数字图书馆的核心资源。数字化古籍在规范化、标准化的基础上,可以发挥海量存储、多元检索等巨大优势。古籍用字的规范处理是数字图书馆对各种古代文献进行数字加工、制作、存储、保存、利用的依据。

国际标准化组织负责的汉字编码工作已进行了近二十年,此前由中、日、韩等国及台湾地区等分别提出过各自的汉字编码方案,全部被编码的汉字超过七万多。这些编码汉字支撑了数字图书的制作,但也造成了编码不一、多平台共存、相互间难以转换的问题,给数字图书馆管理与应用造成很大困难。而古籍用字因其状况复杂,是汉字编码的一个主要难点。古籍原本中存在大量的古今字、通假字、隶古定字、异体字、笔形微异字、俗字、讹字等生僻字,不同时代的版本往往反映出不同时代的用字偏好。生僻字不经过上下文、同时代文献、同一典籍不同版本间的对比,往往难以确认其形音义以及与通行文字的字际关系。在以保存古籍原貌为原则的专业古籍数字化工作中,文字形态的差异也是保存原貌的一个重要方面,字形差异要尽量予以保留。事实上,生僻字处理是古籍数字化中难度最大的一项基础工作。

古籍用字属性字典为进入国际标准的全部古籍用字逐一标出它们的国际标准代码和不同技术环境中的变体,以及在各国各地区编码体系中的代码,为各类编码的转换提供了基础。对不见于已有字符集

中的古籍用字,也为其指定了相应的编码。古籍用字属性字典在现有各语言文字规范的基础上,对收入的汉字确定了字形、字音、字际关系等各类属性,为中文信息处理和古籍数字化制作、管理、应用提供了基础性的规范。

古籍用字规范对现有字符集的未收的字尽量予以正形,以减少因笔形轻微变异而产生过多的生僻字形。对不见于现有各字符集的古籍数字化过程中收集到的有上下文用例的生僻字和异体字,有特定含义的人名、地名、书名用生僻字,均收入并造字。古籍用字属性库将是一个开放的字库,可以不断整合在古籍数字化中出现的生僻字。

目次

前 言	(1)
引 言	(1)
古籍用字(包括生僻字、避讳字)属性字典规范	(1)
1 范围	(3)
2 规范性引用文件	(3)
3 术语和定义	(6)
4 古籍用字(包括生僻字、避讳字)属性字典规范格式	(17)
5 古籍用字(包括生僻字、避讳字)属性字典规范	(22)
古籍用字(包括生僻字、避讳字)属性字典规范应用指南	(45)
1 说明	(47)
2 关于引用及参考的标准和规范	(47)
3 古籍用字字型标准化规范	(48)
4 古籍用字总笔画、起笔至末笔笔形值标注规范	(50)
5 古籍用字笔画笔顺排序规范	(51)
6 古籍用字 201 部首归部规范、部首笔画数规范	(51)
7 古籍用字 214 部首归部规范、部首笔画数规范	(53)
8 古籍用字四角号码及附角号码标注规范	(57)
9 古籍用字四角号码排序规范	(57)
10 古籍用字使用频度的统计对象	(58)
11 古籍用字标准发音标注规范;古籍用字的古音标注规范 ..	(61)
12 古籍用字的异体字对应关系定义	(62)

13	古籍用字(包括生僻字、避讳字)属性字典数据描述	(64)
14	数据库	(75)
后	记	(78)

古籍用字(包括生僻字、避讳字)属性字典规范

1 范围

本规范针对中文古籍文献类型复杂、使用汉字字符数量大等特点,对 GB18030—2005 (UNICODE 5.1、ISO10646—2003) 所包括的全部古籍用字的各类属性,在整合已有语言文字标准规范的基础上,参考权威字典辞书,给予严格规定与标注。对于古籍数字化过程中收集到的有上下文用例的生僻字,有特定含义的人名、地名、书名用生僻异体字,均收入并造字。本规范适用于数字图书馆对各种中文古籍文献进行数字加工、制作、存储、保存、利用。也可用于其他数字出版物的制作、辞书编纂、索引编制以及中文信息处理中涉及的汉字规范化工作。

2 规范性引用文件

下列标准和规范中的条款为本规范所遵循。本规范对其中若干条款做出了更进一步的规定。不同标准或规范中的相应条款规定不一致时,本规范选择符合汉字信息处理实际的规定,并在规范中做详细说明。

ISO/IEC 10646—2003 ISO/IEC 10646 International Standard, “Information technology—Universal Multiple-Octet Coded Character Set (UCS)” 2003 ISO/IEC

GB13000—200X 信息技术 通用多八位编码字符集 (UCS)
2006 年修订稿 国家技术监督局

GB18030—2005 信息技术 中文编码字符集 2005 国家技术监督局

GF3003—1999 现代汉语通用字笔顺规范 1997 -04 -07 国家语委、新闻出版署 (语文出版社)

GF3001—1997 信息处理用 GB13000.1 字符集汉字部件规范 1997 -12 -01 国家语委 (语文出版社)

GB12200.1—90 汉语信息处理词汇 01 部分:基本术语 1990 -02 -01 国家技术监督局 (中国标准出版社:《国家语言文字规范和标准选编》)

GB/T12200.2—94 汉语信息处理词汇 02 部分:汉语和汉字 1994 -12 -07 国家技术监督局 (中国标准出版社:《国家语言文字规范和标准选编》)

GB/T13418—92 文字条目通用排序规则 1992 -04 -13 国家技术监督局 (中国标准出版社:《国家语言文字规范和标准选编》)

GB/T15732—1995 汉字键盘输入用通用词语集 1995 -08 -01 国家技术监督局 (中国标准出版社:《国家语言文字规范和标准选编》)

GB/T16159—1996 汉语拼音正词法基本规则 1996 -01 -22 国家语委、国家教委 (语文出版社:《语言文字规范手册》)

GB13000.1—1993 信息技术 通用多八位编码字符集 1993 -12 -24 国家技术监督局 (中国标准出版社)

GB2312—1980 信息交换用汉字编码字符集 基本集 1981 -05 -01 国家标准总局 (中国标准出版社)

GB12345—1990 信息交换用字编码字符集 辅助集 1990 国家标准总局 (中国标准出版社)

GB13131—1991 信息交换用字编码字符集 第三辅助集 1991