国家数字图书馆工程标准规范成果

计算机中文信息处理规范和应用指南

蒋贤春　翟喜奎　主编

國家圖書館出版社

计算机中文信息处理规范和应用指南

蒋贤春　翟喜奎　主编

國家圖書館出版社

图书在版编目(CIP)数据

计算机中文信息处理规范和应用指南/蒋贤春,翟喜奎主编. —北京:国家图书馆出版社,2012.11

(国家数字图书馆工程标准规范成果)

ISBN 978-7-5013-4871-8

Ⅰ.①计… Ⅱ.①蒋… ②翟… Ⅲ.①汉字信息处理系统 Ⅳ.①TP391.12

中国版本图书馆CIP数据核字(2012)第231159号

责任编辑:高爽

书名 计算机中文信息处理规范和应用指南

著者 蒋贤春 翟喜奎 主编

出版 国家图书馆出版社(100034 北京市西城区文津街7号)

(原北京图书馆出版社)

发行 010-66114536 66126153 66151313 66175620

66121706(传真),66126156(门市部)

E-mail btsfxb@nlc.gov.cn(邮购)

Website www.nlcpress.com→投稿中心

经销 新华书店

印刷 北京科信印刷有限公司

开本 787×1092(毫米) 1/16

印张 7.25

版次 2012年11月第1版 2012年11月第1次印刷

字数 100(千字)

书号 ISBN 978-7-5013-4871-8

定价 58.00元

丛书编委会

本书编委会

主　编：蒋贤春　翟喜奎

编　委：郑　珑　朱人杰　蓝　飞　谢术清　郭胜霞　张秀欣
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

本标准规范是国家数字图书馆工程标准规范项目研制成果之一。

本标准规范由国家图书馆提出，委托北京中易中标电子信息技术有限公司研制。

本标准规范由北京中易中标电子信息技术有限公司起草，主要起草人为：蒋贤春、郑珑、朱人杰、蓝飞、谢术清、张秀欣、郭胜霞。

目　录

前　言 …… (1)

第一部分　计算机中文信息处理规范 …… (1)

1　范围 …… (3)
2　引用标准 …… (3)
3　术语和定义 …… (3)
4　汉字编码 …… (4)
5　汉字排序 …… (24)
6　存储格式 …… (29)
7　传输格式 …… (35)
8　全文显示 …… (42)
附录 A　(资料性附录)汉语拼音和韦氏拼音对照表 …… (44)
附录 B　(资料性附录)汉语拼音和注音对照表 …… (48)
参考文献 …… (52)

第二部分　计算机中文信息处理规范应用指南 …… (53)

1　汉字编码 …… (55)
2　文献排序 …… (71)
3　文件存储 …… (76)
4　文件传输 …… (88)
5　全文显示 …… (96)

后　记 …… (101)

第一部分　计算机中文信息处理规范

1 范围

本规范规定了计算机中文信息处理领域中文件格式、存储格式、传输格式、全文显示、汉字规范化信息、文献排序的规范。

本规范适用于国家图书馆进行中文信息处理、信息交换、汉字输入、文献排序以及在计算机系统上建立文件、检索、显示、打印输出时所需的排序要求等。在使用这一规范时，可根据本规范和国家图书馆的具体需要补充制定相应的细则。

2 引用标准

GB 2312—80 信息交换用汉字编码字符集 基本集

GB 18030—2000 信息交换用汉字编码字符集 基本集的扩充

GB 18030—2005 信息技术 中文编码字符集

GB/T 13016—1991 标准体系表编制原则和要求

GB/T 1.1—2000 标准化工作导则

GB/T 16680—1996 软件文档管理指南

GB/T 13418—92 文字条目通用排序规则

GB/T 12200.1 汉语信息处理词汇

GB/T 12200.2 汉语信息处理词汇

ISO 7098—1991 中文的罗马化

GF 0012—2009 GB 13000.1 字符集汉字部首归部规范

GF 0013—2009 现代常用独体字规范

GF 0014—2009 现代常用字部件及部件名称规范

GF 3002—1999 GB 13000.1 字符集汉字笔顺规范

GF 0011—2009 汉字部首表

3 术语和定义

3.1 文本(Text)

通过文字、符号的形式表现、传递信息的方式。读者在文本数据中通过对文字、符号的阅读来获取信息。

3.2 格式(Formats)

用来存储信息的各种方法。文件格式是用来对数据以及相关信息(包括结构、布局、压缩算法等)进行编码的软件算法。

3.3 文本文件(Text File)

用字符内码存储的文件。它是计算机中最常见也是最原始的文件格式,组成简单,存储体积极小。文本文件中可以含有超链接。通过超链接,文本文件中可以包含各种多媒体文件,如图像、音频、视频等。

3.4 集外字(Gaiji Outside the Font Set)

指特定的字符集以外的汉字。本规范的集外字指超出 GB 18030—2005 字符集的汉字。

3.5 系统外字(Gaiji)

简称为"外字",指用户需要处理,但在计算机当前的操作系统中并不存在的汉字。

3.6 内码(Standard Code)

内码是指系统中使用的二进制字符编码,是沟通输入、输出与系统平台之间的交换码。通过内码可以达到通用和高效率传输文本的目的。内码由国际编码演化而来。

3.7 外码(Input Code)

汉字的输入码称为"外码"。输入码即指我们输入汉字时使用的编码。

4 汉字编码

4.1 汉字内码编码

4.1.1 常用的汉字内码编码标准

表 1 常用的汉字内码编码标准

编码标准	使用的国家和地区	说 明
ISO/IEC 10646	全球	国际标准。ISO/IEC 10646—2003《信息技术通用多八位编码字符集》。简称 UCS。

续表

编码标准	使用的国家和地区	说　明
Unicode	全球	国际标准。分为 Unicode－16(2 个字节编码)和 Unicode－32(用 4 个字节为字符编码)。
GBK	中国内地	中国内地。《汉字内码扩展规范》(简称:GBK)。国家技术监督局标准化司、电子工业部科技与质量监督司 1995 年 12 月 15 日颁布和实施。
GB 2312—80	中国内地	中国内地。《信息交换用汉字编码字符集 基本集》。1980 年由国家技术监督局审批颁布和实施。
GB 18030—2005	中国内地	中国内地《信息技术 中文编码字符集》。由国家质量监督检验总局和中国国家标准化管理委员会于 2005 年 11 月 8 日颁布,2006 年 5 月 1 日实施。
CNS	中国台湾、香港	中国台湾。《全字库中文标准交换码》。1986 年台湾地区审定颁布。
HKSCS	中国香港	中国香港。《香港增补字符集—2004》。在 2005 年 5 月,香港特区政府推出。
JIS	日本	日本。指 Shift-JIS,日本电脑系统常用的编码表。
GB 12345—90	中国内地	中国内地。《信息交换用汉字编码字符集 第一辅助集》。国家技术监督局 1990 年颁布。繁体字的编码标准。
GB 13000	中国内地	中国内地。GB 13000.1—93《信息技术 通用多八位编码字符集(UCS)第一部分:体系结构与基本多文种平面》。国家技术监督局 1993 年 12 月 23 日颁布。
GB 13131	中国内地	中国内地。指 GB 13131—1991《信息交换用汉字编码字符集　第三辅助集》。国家技术监督局 1991 年颁布。
GB 13132	中国内地	中国内地。指 GB 13132—1991《信息交换用汉字编码字符集　第五辅助集》。国家技术监督局 1991 年颁布。

注:建议汉字内码采用 Unicode 编码标准。

4.1.2　汉字内码编码表

表 2　部分汉字内码编码表

UCS	字	Unicode	GBK	GB 2312	GB 18030	BIG5	JIS	CNS
04E00	一	4E00	D2BB	D2BB	D2BB	A440	88EA	1－4421
04E01	丁	4E01	B6A1	B6A1	B6A1	A442	929A	1－4423
04E02	丂	4E02	8140	(无)	8140	(无)	(无)	4－2126
04E03	七	4E03	C6DF	C6DF	C6DF	A443	8EB5	1－4424

续表

UCS	字	Unicode	GBK	GB 2312	GB 18030	BIG5	JIS	CNS
04E04	丄	4E04	8141	(无)	8141	(无)	(无)	3-2126
04E05	丅	4E05	8142	(无)	8142	(无)	(无)	3-2125
04E06	丆	4E06	8143	(无)	8143	(无)	(无)	(无)
04E07	万	4E07	CDF2	CDF2	CDF2	C945	969C	2-2126
04E08	丈	4E08	D5C9	D5C9	D5C9	A456	8FE4	1-4437
04E09	三	4E09	C8FD	C8FD	C8FD	A454	8E4F	1-4435
04E0A	上	4E0A	C9CF	C9CF	C9CF	A457	8FE3	1-4438

本规范给出了 GB 18030—2005 所包括的全部汉字的内码编码表,详见本书第二部分。

4.2 汉字外码编码

汉字外码包括音码、形码、音形码、部首笔画、部件笔画编码等。汉字外码主要用于汉字输入、排序与检索。

4.2.1 外码编码原则

汉字外码编码的基本原则:规范、实用。

表 3 汉字外码编码原则

外码类型	编码原则
音码	有单字、词和短语的编码。“音”符合现代汉语拼音。韦氏拼音、注音除外。
音形码	有单字、词和短语的编码。“音”符合现代汉语拼音;“形”拆分规范符合《GB 13000.1 字符集汉字部首归部规范》《现代常用独体字规范》《现代常用字部件及部件名称规范》。
形码	部件拆分规范符合国家语言文字规范,易学易记、输入快速、具有通用性和可扩展性。
部首笔画	部首拆分符合《现代常用字部件及部件名称规范》《汉字部首表》和《GB 13000.1 字符集汉字部首归部规范》。汉字笔顺需符合本规范汉字笔画编码规则(参见第一部分 4.2.9.1 节)。
部件笔画	部件拆分规范符合《现代常用字部件及部件名称规范》《汉字部首表》和《GB 13000.1 字符集汉字部首归部规范》,汉字笔顺需符合本规范汉字笔画编码规则(参见第一部分 4.2.9.1 节)。

对应 GB 18030—2005 所包括的全部汉字,本规范给出:3 种音码编码(拼音、韦氏拼音、注音),2 种形码编码(四角号码、郑码),4 种部首笔画编码(现代部首、康熙部首、笔画、现代部首笔画)和 1 种部件笔画编码(笔画字)。

4.2.2 拼音编码

4.2.2.1 汉字拼音编码规则

表 4　汉字拼音编码规则

序号	编码规则
1	汉字发音使用现代汉语拼音编码,必须有声调。
2	汉字有多个发音时,给出所有发音的现代汉语拼音编码。
3	发音符合现代汉语拼音。
4	汉字发音无法用现代汉语拼音编码时,可以使用罗马音编码。
5	对无法给出拼音和罗马音的汉字,暂不给编码。

4.2.2.2　汉字拼音编码表

表 5　部分汉字拼音编码表

UCS	字	拼音编码	韦氏拼音编码
04E00	一	yī	i
04E01	丁	dīng;zhēng	ting;cheng
04E02	丂	kǎo;qiǎo;yú	k'ao;ch'iao;yü
04E03	七	qī	ch'i
04E04	丄	shàng	shang
04E05	丅	xià	hsia
04E07	万	wàn; mò	wan;mo
04E08	丈	zhàng	chang

本规范给出了 GB 18030—2005 所包括的全部汉字的拼音编码表,详见本书第二部分。

本规范给出了 GB 18030—2005 所包括的全部汉字的韦氏拼音编码表,详见本书第二部分。

4.2.2.3　汉语拼音和韦氏拼音对照表

拼音和韦氏拼音对照表见本部分附录 A。

韦氏拼音没有声调。

4.2.3　注音编码

4.2.3.1　汉字注音编码表

表 6　部分汉字注音编码表

UCS	字	注音编码
04E00	一	ㄧ
04E01	丁	ㄉㄧㄥ;ㄓㄧㄥ

续表

UCS	字	注音编码
04E02	丐	ㄎㄞˇ;ㄑㄧㄠˇ;ㄐㄧㄝˋ
04E03	七	ㄑㄧ ˉ
04E04	丄	ㄕㄤˋ
04E05	丁	ㄊㄧㄥˉ
04E07	万	ㄨㄢˋ;ㄇㄛˋ
04E08	丈	ㄓㄤˋ

本规范给出了 GB 18030—2005 所包括的全部汉字的注音编码表,详见本书第二部分。

4.2.3.2 拼音和注音对照表

拼音和注音对照表见本部分附录 B。

注音有声调,标记在最后,“ˉ”表示 1 声;“ˊ”表示 2 声;“ˇ”表示 3 声;“ˋ”表示 4 声;没有声调表示轻声。

4.2.4 四角号码编码

4.2.4.1 四角号码检字法

第一条:笔画分为十种,用 0 到 9 十个号码来代表。

表 7 四角号码检字法

号码	笔名	笔形	举例	说明	注意
0	头	亠	言 主 广 疒	独立的点和横相结合	1、2、3 都是单笔,0、4、5、6、7、8、9 都是二笔以上的单笔合为一复笔的,凡能成为复笔的,须取复笔,切勿误作单笔;如亠应作 0 不作 3,寸应作 4 不作 2,厂应作 7 不做 2,丷应作 8 不作 3、2,小应作 9 不作 3、3。
1	横	一 乚	天 土 地 江 元 風	包括横、挑(趯)和右钩	
2	垂	丨 丿 亅	山 月 千 則	包括直、撇和左钩	
3	点	丶	宀 礻 冖 之 衣	包括点和捺	
4	叉	十 乂	草 杏 皮 刈 大 对	两笔相交	
5	插	扌	扌 戈 中 史	一笔通过两笔以上	
6	方	口	國 鳴 目 四 甲 由	四边齐整的方形	
7	角	厂 冖	羽 門 厅 阴 雪 衣 舅 罕	横和垂的锋头相接处	
8	八	八 丷 人	分 頁 羊 余 灾 汆 足 午	八字形和它的变形	
9	小	小 ⺌ 个 忄	尖 糸 辨 杲 惟	小字形和它的变形	

第二条:每字只取四角的笔形,顺序为:左上、右上、左下、右下角。

照四角的笔形和顺序,每字得四码。例:颜 =0128;截 =4325;烙 =9786。

第三条:字的上部或下部,只有一笔或一复笔时,无论在何位,都作左角,它的右角作 0。每笔用过后,如再充他角,也作 0。