ICS 35.020
CCS L 70

T/CQAE

团　　体　　标　　准

T/CQAE XXXX.2

信息技术　生僻字处理
第 2 部分：业务系统要求

Information technology－Processing rarely used Chinese characters－Part 2:business system requirements

（报批稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中国电子质量管理协会　　发 布

目　　次

前言 III
引言 V
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 通则 1
6 各环节建设及改造要求 2
6.1 输入 2
6.2 显示 2
6.3 打印 2
6.4 信息交换 3
6.5 内部处理 4
6.6 存储 4
参考文献 6

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。

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。

本文件由中国电子技术标准化研究院提出。

本文件由中国电子质量管理协会归口。

本文件起草单位：中国电子技术标准化研究院、中国科学院软件研究所、中信银行股份有限公司、北京金融科技产业联盟、北京银行股份有限公司、中国民航信息网络股份有限公司、九江银行股份有限公司、文化艺术出版社有限公司、深圳市腾讯计算机系统有限公司、蚂蚁科技集团股份有限公司、微软（中国）、电子科技大学、上饶师范学院等。

本文件主要起草人：王欣、黄姗姗、陶扬、崔晓琳、刘汇丹、吴健、马良有、刘伟铧、王震、胡达川、李寻、王子健、陈陈、周祥伟、彭志强、白非非、陈永聪、蒋增增、陆碧波、林琳、范计朋、黄起豹、周宗明、孙梦等。

引　　言

T/CQAE XXXX《信息技术　生僻字处理》旨在针对信息系统生僻字问题，分别从软件产品、业务系统改造和服务机构的层面，提出处理生僻字的产品设计要求，解决信息系统生僻字问题。本文件是T/CQAE XXXX《信息技术　生僻字处理》的第2部分。T/CQAE XXXX拟由三个部分构成。

——第 1 部分：软件产品要求。目的在于对新交付的软件产品提出生僻字处理相关的技术要求，指导软件企业的研发工作，并对有关机构在采购相关产品时提供参考标准。

——第 2 部分：业务系统要求。目的在于围绕业务系统在生僻字处理方面的实际需求，提出业务系统建设及改造要求，为业务系统建设单位及需求单位提供指导。

——第 3 部分：服务机构要求。目的在于向服务机构提出生僻字处理的相关要求，包括技术管理机制、服务管理机制等。

信息技术　生僻字处理　第2部分：业务系统要求

1　范围

本文件规定了业务系统处理生僻字的建设及改造要求。
本文件适用于涉及到中文信息化处理及交换的各类业务系统。

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/T 13000　信息技术　通用编码字符集（UCS）
GB 18030　信息技术　中文编码字符集
T/CQAE XXXX.1　信息技术　生僻字处理　第1部分：软件产品要求

3　术语和定义

T/CQAE XXXX.1界定的以及下列术语和定义适用于本文件。

3.1

业务系统 business system
政务服务和公共服务行业为满足用户需求和目标，包含一个或多个信息系统之间关联交互的系统。

4　缩略语

下列缩略语适用于本文件。
APP：移动应用程序（Mobile Application）
BOM：字节顺序标记（Byte Order Mark）
CCSID：编码字符集标识（Coded Character Set Identifier）
CJK：中日韩统一表意文字（China, Japan and Korea unified ideographs）
EBCDIC：扩展二进制编码十进制交换码（Extended Binary Coded Decimal Interchange Code）
FTP：文件传输协议（File Transfer Protocol）
JSON：JavaScript对象简谱（JavaScript Object Notation）
MBCS：多字节字符集（Multi-Bytes Character Set）
NFC：近场通信技术（Near Field Communication）
OCR：光学字符识别（Optical Character Recognition）
OFD：版式文档（Open Fixed-layout Document）
PC：个人电脑（Personal Computer）
PDF：便携式文档格式（Portable Document Format）
PUA：用户自定义区（Private Use Area）
SQL：结构化查询语言（Structured Query Language）
UCS：通用编码字符集（Universal Coded character Set）
XML：可扩展置标语言（Extensible Markup Language）

5　通则

有关机构在业务系统建设或存量系统改造过程中，为支持生僻字的处理，宜遵守以下原则。
a)　遵循标准。应支持GB 18030，宜兼容GB/T 13000的相应编码，字符集以最新版本为准；

b) 易于扩展。使用可扩展和安全可控的技术框架和方案；
c) 经济适用。以满足用户实际需要为基础，配置适用的字库、输入法、接口设备、输出设备等；
d) 兼容处理。宜兼容处理涉及民生的各类信息系统中的生僻字问题；
e) 包容普惠。宜考虑农村与偏远地区居民、老年人、残障人士、少数民族等群体的需求；
f) 接口统一。对超出支持范围的字符转义表示宜采用统一方案。

6 各环节建设及改造要求

6.1 输入

6.1.1 输入法/输入设备要求

业务系统配备的输入法/输入设备应符合T/CQAE XXXX.1《信息技术 生僻字处理 第1部分：软件产品要求》中的相关规定。其中：

a) PC 端应用应不限定输入法，允许使用拼音、笔画、字形等多种输入方法，宜优先配备支持生僻字的常规输入法。常规输入法产品无法满足应用需求时，应配置第三方软件形式的输入法、云输入法或 APP 内嵌输入法或提供其他指引，以保证至少有一种方法可将生僻字录入到系统中；
b) 移动应用、智慧柜台等触摸屏应用宜在输入焦点进入输入域后自动弹出系统默认输入法界面，在有多种输入法时允许用户切换。

6.1.2 不同场景下的输入要求

如下要求适用于不同场景和输入情形：

a) 针对客户需使用实体身份证进行核验的场景，宜采用机具读入姓名等身份信息。若因风控限制手工输入姓名，也应考虑机具驱动程序缺陷、客户证件芯片编码错误等异常场景，提供其它补充手段；
b) 通过 OCR、语音识别、手写识别等方式输入的，应提供人工核对、修正功能；
c) 对于移动应用输入身份证信息的场景，建议增加利用客户手机端 NFC 功能读取身份证芯片信息到机构后端解密后自动导入的功能；
d) 对于支持输入法输入信息的字段，应支持复制粘贴的录入方式；
e) 少数民族姓名中的间隔符应按照《关于在政府管理和社会公共服务信息系统中统一姓名采集应用规范的通知》（民委发〔2016〕33 号）要求的格式输入，统一用“•”（GB 18030 编码 A1A4,GB/T 13000 编码 00B7）。常用字符集中实心“点”字符有多个，宜在用户输入的前端进行自动检测及转换。

6.2 显示

6.2.1 一般要求

业务系统在汉字信息的显示方面的一般要求包括：

a) 可显示 GB 18030 规定的全部汉字；
b) 宜参考 ISO/IEC 10646 最新版本覆盖新增汉字；

6.2.2 特殊情况的处理

生僻字信息在显示时如遇到以下特殊情况，可采用下列处理方式：

a) 单个字型文件限制字形数量时，宜通过操作系统的字体回退机制或者应用软件自行实现的字体回退机制实现生僻字的显示；
b) 在必须显示 PUA 编码汉字的情况下：
   1) 宜对 PUA 编码汉字字形与正式编码字形作出明显区分；
   2) 对于人名地名等的 PUA 编码生僻字，宜采用 GB 18030 实现级别 3 的字库予以显示，供用户确认。

6.3 打印

6.3.1 不同类型的打印机生僻字处理方法

通用打印机包括针式打印机、激光打印机和喷墨打印机等，不同类型打印机在处理生僻字时，可使用以下三种方法，见表1。

表 1 打印机生僻字处理方法

实现方案	实现方式	适用范围
文本图形混合方案	a)在硬字库支持范围内，用文本打印模式。 b)在硬字库支持范围外,由应用端程序转换成图片后再打印。	带有硬字库的针式打印机，如存折打印机、宽行打印机等。
纯图形方案	依赖操作系统的图形输出，打印机按照图形输出进行打印。	日常办公类的打印机，如激光打印机、喷墨打印机等。
纯文本方案	升级打印机字库，字库支持GB 18030实现级别3。	带硬字库的针式打印机，如存折打印机、宽行打印机等。

6.3.2 打印机字库

打印机是否能正确打印生僻字信息，与打印机内置字库和/或系统字库有关。

a) 生产厂商应及时跟踪国家标准最新版本升级点阵打印机内置字库，实现对生僻字的支持；
b) 部分生僻字笔划较多，应避免采用过小的点阵字体（32×32 点阵以下）导致因减笔划而造成有法律效应的打印件产生纠纷；
c) 在点阵字库不支持的情况下，应通过图形打印的方式确保生僻字被正确打印；
d) 打印文件为 PDF、OFD 等文件时，应使用符合 GB 18030 实现级别 3 的字库以字体嵌入方式生成 PDF、OFD 文件，避免生僻字打印结果与客户信息不一致。

6.4 信息交换

6.4.1 一般要求

业务系统在汉字信息交换方面的一般要求包括：

a) 信息系统间使用不同编码字符集进行数据交换时，应支持 GB 18030 的汉字无损透传处理，同时宜 UTF-8 编码；
b) 原使用 GBK 编码的报文及文件交换应升级为 GB18030 编码，同时宜支持 UTF-8 编码；
c) 转接系统在转接时，需要做编码转换时，不应发生：
   1) 丢弃某些字符或转成替代符“?”的有损转换；
   2) 报文丢弃。

6.4.2 特殊情况的处理

生僻字信息在交换时如遇到以下特殊情况，可采用下列处理方式：

a) 原内部系统间接口为 GBK 或 EBCDIC CCSID 1388 等小字符集的编码，且改造成本过大，可以保留，此时可借助中间件或改用转义格式对生僻字进行表示和交换；

注：转义格式是以多个字符的有序组合来间接地表示一个特殊字符的方式。例如，C、Java、Python等语言都使用转义序列“\n”表示换行符。又如，HTML中使用转义序列“
”或“
”表示换行符等。

b) 需要交换的信息包含 PUA 编码汉字，正式编码发布后及时使用正式编码，请求方宜采用生僻字的标准编码对 PUA 编码字符进行归一化处理；
c) 对于“一字多码”的生僻字进行联网核查公民身份姓名信息时：
   1) 对“一字多码”的生僻字做兼容处理；
   2) 应使业务系统支持一字多码互相认同的智能比较；
   3) 对于当前系统未改造尚不支持处理生僻字的情况下，应转人工处理，需要时可联系客户核实处理。

6.4.3 其它要求

针对生僻字的信息交换，还需注意如下技术要点：

a) 避免使用字段定长无分隔符格式报文或文件进行交换；
b) 若采用变长字段有分隔符格式报文或文件进行交换，应考虑分隔符的选取与业务报文内容的字节冲突问题；
c) 对于 XML 报文或文件进行交换，需注意头部的 encoding 编码设置须与内容采用的编码一致，避免 XML 解析器解码错误；
d) 对于 JSON 报文或文件进行交换，需注意其默认使用 UTF-8 编码，而非 GB18030 编码，且辅助平面字符可能采用 UTF-16“代理对”转义字符串表示；
e) 对于 UTF-8、UTF-16、UTF-32 编码的文件，宜检测文件开头是否存在 BOM 标记。若存在，通过 BOM 标记可识别文件的编码方式。某些操作系统自带文本编辑器保存时，会在文件开头自动加上 BOM 标记，应用程序若不支持带 BOM 的文件，文件编辑时往往会报错；
f) 以 FTP 方式交换文件不需要转码时，应设定为二进制（BIN）流方式；如需转码时，宜设定相应的编码集，以保证无损透传；
g) 使用邮件系统交换信息时，Base64 变换前的编码不宜使用 GBK 或 GB2312，宜使用 UTF-8。

6.5 内部处理

关于生僻字信息的内部处理，部分要点按6.4执行。此外应注意：

a) 所用编程语言（包括 SQL）的字符串长度函数/方法得到的结果与字符数、字节数都可能存在差异，原生字符串截取的函数/方法有可能导致半个汉字的异常问题，需要另行开发支持生僻字的函数/方法；
b) 应采用支持生僻字的编程语言进行应用系统开发、编译；
c) 对于如“姓名”字段字符串的实名制比对，不应采用所用编程语言的字符串比较函数/方法（如 C 语言的 strcmp()函数、Java 语言 String 类的 equals() 方法），宜另行开发支持“一字多码”姓名认同的函数/方法。

6.6 存储

6.6.1 一般要求

业务系统在汉字信息存储方面的一般要求：

a) 新建应用系统的数据库的存储和查询应支GB 18030实现级别3所有字符,编码应支持GB18030，同时宜支持 GB/T 13000 的 UTF 编码中的至少一种；
b) 新建应用系统的数据库导入、导出文件编码应支持 GB18030，同时宜支持 GB/T 13000 的 UTF-8 编码；
c) 存量应用系统数据库宜升级为支持 GB18030 或 UTF-8，在难以升级时：
   1) 从数据库读出数据时，应将转义格式还原成汉字；
   2) 转义格式宜基于易于还原、占用空间小的 GB/T 13000 编码；
   3) 转义格式仅限在数据库内部使用，外部访问应还原为汉字，以保证透传、通用。

6.6.2 其他要求

在存储方面还应注意：

a) 信息系统设置中姓名数据项最大长度不少于 25 个汉字；
b) 考虑到转义格式可能会扩展原数据所需宽度，应特别注意字段长度设计。

6.6.3 常见数据库产品的处理要求

常见数据库产品的生僻字处理要求：

a) MySQL 数据库

使用MySQL 数据库时宜采用5.5.3以上版本，并将UTF-8 的编码类型设置为utf8mb4。

b) DB2 数据库
   1) 在大型主机系统中，CJK 扩充 B 及以上扩充区、急用加字区的汉字宜用转义格式处理；
   2) 开放平台 DB2 数据库宜升级支持 UTF-8 或 GB 18030 编码。

c) Oracle 数据库

Oracle数据库宜将字符集值设置成AL32UTF8或AL16UTF16。

d) 其他数据库

其他数据库宜使用GB 18030、UTF-8等支持全字符集的编码。

参 考 文 献

[1] 《关于在政府管理和社会公共服务信息系统中统一姓名采集应用规范的通知》（民委发〔2016〕 33 号文）.2016-04-15