Q/GZYH

赣州银行股份有限公司企业标准

Q/GZYH 003-2024

赣州银行生僻字处理指南

2024-10-30 发布　　2024-10-20 实施

赣州银行股份有限公司　　发布

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》给出的规则起草。

本文件由赣州银行股份有限公司科技部提出并归口。

本文件起草单位：赣州银行股份有限公司科技部

本文件主要起草人：

本文件于 2024 年 10 月 25 日首次发布。

引言

生僻字问题是指信息系统无法处理含生僻字的姓名、地址数据等，导致公民无法在各类服务系统中正常办理业务的情况，已成为数字化转型中“数字鸿沟”的表现形式之一。为了贯彻落实“十四五规划纲要”要求，基于《金融服务　生僻字处理指南》（JR/T 0253—2022）金融行业标准要求，保障姓名中带有生僻字的公民享有金融服务的基本权益，切实提高人民群众对金融服务的获得感和满意度，赣州银行股份有限公司制定《赣州银行生僻字处理指南》，对信息系统在输入、显示、打印、存储、交换等一个或多个环节提出统一的处理规范要求。本文件可作为信息系统的建设和改造依据、涉及中文信息处理的外部产品引入采购参考。

本文件仅限于应用在赣州银行股份有限公司。

生僻字处理指南

1 范围

本文件规定了赣州银行生僻字处理规范总体要求，以及提出了生僻字在信息系统的显示、输入、打印、存储、信息交换等方面的指南。

本文件适用于赣州银行股份有限公司的信息系统建设，以及与中文编码字符集相关产品采购。

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。

ISO/IEC 8859-1 信息技术—8 位单字节编码图形字符集—第 1 部分: 1 号拉丁字母(Information technology—8-bit single-byte coded graphic character sets—Part 1:Lain alphabet No. 1) ISO/IEC 10646 信息技术 通用编码字符集 (UCS)(Information technology-Universal Coded Character Set(UCS))

GB 18030-2022 信息技术 中文编码字符集

GB/T 13000 信息技术 通用多八位编码字符集 (UCS)

JR/T 0253-2022 金融服务 生僻字处理指南

3 术语和定义

下列术语和定义适用于用于本文件。

3.1

编码字符集 coded character set

一组无歧义的规则，用以建立一个字符集和该字符集中的字符及其编码表示之间的对应关系，也指 按照这种规则确定的文字的有序集合。

注：GB 18030（含空格）指《信息技术 中文编码字符集》标准；GB18030（无空格）指具体字符编码。

3.2

编码字符集标识 coded character set identifier

标识大型主机当前字符使用的编码字符集 (3. 1)编号。

3.3

字库 font library

建立在计算机存储媒体上的字形数据集合。

3.4

人口信息字库 font library of population information

户籍管理部门针对人口信息 (人名、 地名等)数据数字化而定制的字库 (3.3)，采用 GB/T 13000 编码。

3.5

用户自定义区 private use area;PUA

未在通用编码字符集（见 3.7 定义）中指定，由私有规则决定字符用途的一系列码点，使用三个编 码区块：U+E000~ U+F8FF、U+F0000~U+FFFFD、U+100000~U+10FFFD。

3.6

生僻字 rarely used Chinese characters

GB/T 13000 编码区间(U+4E00~U+9FA5，20,902 字)之外的汉字。

注：1993年发布的GB13000收录了U+4E00~U+9FA5共20,902个汉字，1995年发布的《汉字内码扩展规范》(以下简称 GBK)含21,003个汉字(增加了101个汉字及偏旁部首，包括“”，“”，“”等52个汉字)， 现已被GB 18030 代替；由于GBK字符集内的20, 902个汉字已能被国内外绝大部分技术产品和国内的应用系统所支持，而其他的 汉字往往会遇到问题， 故一般认为在20,902个汉字之外的汉字为生僻字。

3.7

通用编码字符集 universal coded character set

国际通用的多八位编码字符集。

注 1：通用编码字符集(UCS)标准由国际标准化组织(ISO)与国际电工委员会(IEC)制订，编号为 ISO/IEC10646,最新版本为 ISO/IEC 10646:2020。我国现行 GB/T 13000-2010 采标自 ISO/IEC10 646:2003。

注 2：统一码(Unicode)是由统一码联盟依据 UCS 制定的可以容纳世界上所有文字和符号的编码字符集，Unicode 比 UCS 额外定义了与字符有关的语义符号学内容。

注 3：UCS 将中国、日本、韩国等国使用的汉字通称为中日韩统一表意文字(CJK)。

注 4：CJK 按编码区块分为基本集(URO)、扩充 A~G、兼容区，急用汉字会在各编码区块末尾增补。

注 5：UCS 在技术实现上，使用 UTF-8、UTF-16、UTF-32 三种编码方式对字符进行编码。UTF-8 是一种以一个或多个 8 位为单元的编码方式；UTF-16 是一种以一个或两个 16 位为单元的编码方式;；UTF-32 是一种以一个 32位为单元的编码方式。16 位以 2 字节表示，32 位以 4 字节表示。对于多个字节的排列顺序， 如果第一个字节是整数二进制中的最高位字节，最后一个字节是整数二进制中的最低位字节，则该字节序称为“大端”(BigEndian,BE)；如果第一个字节是整数二进制中的最低位字节，最后一个字节是整数二进制中的最高位字节，则该字节序称为“小端”(Little Endian, LE)。UTF-16 分 UTF-16BE 和 UTF-16LE 两种方式， UTF-32 分 UTF-32BE 和 UTF-32LE 两种方式。

注 6：本文件以 U+XXXX 或 U+XXXXX 表示 UCS 的一个码点或字符，如 U+0000~U+FFF 称为丛本多文种平面(BMP)，U+20000~U+2FFFF 称为辅助表意文字平面。

4 缩略语

下列缩略语适用于本文件。

APP：移动应用程序（Mobile Application）
ASCII：美国信息交换标准代码（American Standard Code for Information Interchange）
ATM：自动柜员机 (Automatic Teller Machine)
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)
JDK：Java语言开发工具（Java Development Kit）
PC：个人电脑(Personal Computer)
PUA：用户自定义区(Private Use Area)
UCS：通用编码字符集(Universal Coded character Set)

5 生僻字处理模式概述

5.1 信息系统处理汉字的通用模式

信息系统处理汉字的通用模式，包括客户与终端、终端与总线、总线与后台系统、后台系统与外联系统、外联系统与其他机构等交互环节。在客户与终端交互环节，输入、显示、打印处理涉及生僻字。在柜台与总线交互环节，流程、交换处理涉及生僻字。在总线与后台系统、后台系统与外联系统交互环节，交换处理涉及生僻字。在外联系统与其他机构交互环节，流程、交换处理涉及生僻字。

信息系统通常需要在GBK、GB18030、UTF-8等编码间转换处理汉字，因不同类型编码所支持的字符集不同，GBK不支持的汉字需实现无损透传处理。

5.2 生僻字处理分级

生僻字处理分为以下三个级别。

注：本文件的编码字符集指导准则有两份，一是国家强制性标准《信息技术 中文编码字符集》（GB 18030-2022）中的实现级别三级分类，二是金融行业标准《金融服务 生僻字处理指南》（JR/T 0253—2022）中的生僻字处理级别三级分类。为保证本文件前后描述一致，本条采取金融行业标准的生僻字处理级别三级分类，国家强制性标准的实现级别三级分类放至规范性附录。

5.2.1 基础级

基础级包括：

——CJK 基本集和扩充 A，其中包含 52 个 GBK 双码字。

——《通用规范汉字表》全部汉字（含补字区、CJK 扩充 B~E 共 199 字）。

——人口信息字库 PUA 编码部分对应的正式编码汉字（含 CJK 扩充 B~G）。

5.2.2 实用级

实用级（包含基础级，增加实际在用汉字）包括：

——CJK 扩充 B~G 中已知的人名、地名在用汉字。

——人口信息字库 PUA 编码部分，虽有正式编码但仍在用 PUA 编码的汉字。

——人口信息字库 PUA 编码部分，没有正式编码只能使用 PUA 编码的汉字。

5.2.3 完整级

完整级包括：

UCS 收录的全部 CJK 汉字，包含实用级。

6 总体要求

对于本行管控的信息系统，以及涉及编码字符集的外部引进产品，应遵守以下要求。

a）遵循标准。以 GB 18030、GB/T 13000 为基础，便于客户和服务人员识读、辨别生僻字，便于信息系统持续优化。

1）信息系统、数据库，编码字符集采用 UTF-8mb4 编码。

2）信息系统在本行内部的通讯报文，编码字符集采用 UTF-8 编码。

3）信息系统在本行内部、对外报送的交换文件，编码字符集采用 UTF-8 编码。

b）易于扩展。使用可扩展和安全可控的技术框架和方案，便于提升系统服务效率和客户体验。

1）信息系统的新建立项或框架升级，在软硬件选型规划阶段应统筹考虑字符集要求，提前做好标准适配。

2）对于改造困难的、存量信息系统，在过渡期内可采取码值转义兼容方案适配，过渡期期间应规划重新立项或下线退出。转义方式只局限在系统内部，不能对外使用转义方式。

c）经济适用。以满足客户实际需要为基础，配置实用的字库、输入法、接口设备等。

1）调研字库市场现状和需求，根据本行采购办法要求，引入满足各方要求的字库产品。

2）字库产品包括输入法、字库、云输入法、字库及转码等工具软件，由外部专业厂商提供，字库产品应根据本行未来发展方向需要，支持 Windows、Linux 主流系统（麒麟、统信等）、安卓、IOS、鸿蒙等版本操作系统，支持 PC 端、移动端的使用，不限制接入终端及接入场景。

d）兼容处理。当在用的 PUA 字符正式编码发布后及时使用正式编码。在核验环节，兼容处理一字多码的互相认同。

7 生僻字的输入

7.1 输入整体规范

7.1.1 字库产品输入规范

对于各类设备的生僻字输入，通过输入法产品完成，输入法产品属于字库产品，应满足下述要求。

a）输入法字符集范围，应达到《信息技术 中文编码字符集》（GB18030-2022）第三级别要求，以及达到《金融服务 生僻字处理指南》（JR/T 0253—2022）实用级及以上要求，包含公安部人口信息字库中的全部汉字。

b）输入法编码，应支持多种方式输入，包括但不限于拼音输入（支持全拼、简拼）、笔画输入、拆字输入（部首拆字），支持录入符号。支持智能联想、智能纠错、智能组词等智能功能。

c）输入法实现形式，应支持 2 种形式，包括本地输入法、云输入法。

1）本地输入法。输入法和字库提供独立离线安装包，具备较好的兼容性，输入习惯与主流输入习惯保持一致。支持软键盘，能够提供金融场景常用词库。可配置在操作系统的输入法候选列表中并可切换选择。

2）云输入法。信息系统集成云输入客户端，用户在云输入客户端录入，云输入客户端根据输入从云输入服务器端查询到候选字，由用户选择录入信息系统的录入框中。

7.1.2 信息系统输入处理规范

对于涉及客户姓名、地址输入的信息系统，在输入处理时应满足下述要求。

a）信息系统字符检测，应对姓名、地址等栏位输入的字符做检测。字符集限制应以最新 UCS 汉字编码范围为准，不应使用 U+4E00~U+9FA5 的范围来控制只输入 GBK 汉字。

b）信息系统字符兼容，对于身份证件识读设备读取的汉字为 PUA 编码但已有 UCS 编码，应满足 6 的 d)要求，支持一字多码兼容处理；对于生僻字落库存储，应优先存储正式码，若无正式码则存储 PUA 码。

7.2 PC 终端输入配备

PC 终端的范围为本行管理的设备资产，客户使用的 PC 终端不在此范围内，此类设备的生僻字输入应满足 7.1 的要求，输入法实现形式建议采用安装本地输入法软件。

7.3 自助设备输入配备

自助设备的范围为本行管理的设备资产，应遵循下述要求。

a）对于高拍仪、柜内清、柜外清等设备，应支持 GB18030 编码、UTF-8 编码、UTF-16 编码。

b）对于 ATM、智能柜台等自助设备，生僻字输入应满足 7.1 的要求，应支持安装本地输入法，或通过系统改造升级支持云输入法。

c）对于 PAD 等便携式设备，生僻字输入应满足 7.1 的要求，应支持安装本地输入法。

7.4 其他场景（APP、WEB 、PC等）输入配备

其他场景的范围通常是客户在自有设备上，以 APP、WEB、PC 等作为载体，通过渠道类系统，输入生僻字。此场景下的载体包括 H5、WEB、APP（iOS 环境/Android 环境...）等，渠道类系统包括手机银行、网上银行等系统。

其他场景下的生僻字输入应满足 7.1 的要求，输入法实现形式应采用云输入法。

与渠道类系统关联、涉及客户信息存储的其他系统，应满足 7.1.2 的 b)要求。

8 生僻字的显示

8.1显示整体规范

8.1.1字库产品显示规范

对于各类设备的生僻字显示通过字库产品支持，应满足下述要求：

a） 字库字符集范围，应达到《信息技术 中文编码字符集》（GB18030-2022）第三级别要求，以及达到《金融服务 生僻字处理指南》（JR/T 0253—2022）实用级及以上要求。

b） 兼容性，应能兼容本行使用的所有操作系统。

c） 字形，PUA 编码字与正式编码字的字形应作出明显区分。

d） 字库实现形式，应支持 2 种形式，包括本地字库、云字库。

1）本地字库。安装在操作系统上，所有应用均可使用的字库。

2）云字库。存放在服务器端，在客户端需要时才下载到客户端的字库。

8.2 PC 终端字库配备

PC 终端的范围为本行管理的设备资产，客户使用的 PC 终端不在此范围内，此类设备的生僻字显示应 满足 8.1 的要求，字库实现形式采用本地字库安装。

8.3 自助设备字库配备

自助设备的范围为本行管理的设备资产，应遵循下述要求。

a) 对于 ATM、智能柜台、柜外清等自助设备，生僻字显示应满足 8.1 的要求，应支持安装本地字库，或通过系统改造升级支持云字库。

b) 对于 PAD 等便携式设备，生僻字显示应满足 8.1 的要求，应支持安装本地字库。

8.4 其他场景（APP、WEB、PC等）字库配备

其他场景的范围通常是客户在自有设备上，以 APP、WEB、PC 等作为载体，通过渠道类系统，显示生僻字。此场景下的载体包括 WEB、APP（iOS 环境/Android 环境...）、H5 等，渠道类系统包括手机银行、网上银行等系统。

其他场景下的生僻字显示应满足 8.1 的要求，应支持通过系统改造升级支持云字库。

9 生僻字的打印

柜台PC通用打印机类型包括针式打印机、常规打印机和报表高速打印机，各类打印机处理要求如下。

a) 对于针式打印机，优先采用图形或者PDF打印，或者升级内置大字库。

b) 对于常规打印机，通过对连接的PC终端字库配置升级，采用图形或者PDF打印支持生僻字打印，可提供给柜面使用。

c) 对于报表高速打印机，需评估厂家升级字库的改造成本，再参考 a)和 b)的处理方案。

10 生僻字的信息交换

10.1 内部信息系统间的信息交换

内部信息系统应支持 GB18030 或 UCS 字符集(一般用 UTF-8 编码)的汉字无损透传处理，要求如下。

a) ESB 服务总线系统在做转接处理时，如果输入、输出双方编码不同，需要做编码转换时，涉及输入、输出的双方系统不应发生丢弃某些字符或转换替代符的有损转换。

b) 对于规划满足字符集要求的信息系统，需做好各方面的标准适配。

1) 有生僻字输入场景的信息系统，应符合 7.1.2 的要求。

2) 通过 ESB 服务总线接入的信息系统，接口报文应采用 UTF-8 编码。

3) 存在直连调用的信息系统，通讯报文应采用 UTF-8 编码。

c) 对于改造困难的、存量信息系统，可采取码值转义兼容方案适配。有生僻字输入场景的信息系统，应符合 7.1.2 的要求。

10.2 与外部信息系统的信息交换

使用GBK编码的报文及文件交换宜升级为UTF-8或GB18030编码。底层交换可用ISO8859-1编码。在原有接口、系统无法升级的情况下，宜与外部系统协商，使用码值转义处理生僻字，避免生僻字丢弃或转为无意义的替代字符（如问号、空格等）。

11 生僻字的存储及处理

11.1 数据库存储

对于本行管控的信息系统，应满足以下要求。

a) 数据库存储优先采用 UTF-8 编码。

b) 如使用 GBK 编码，在不改变数据库字符集设置的情况下，对于超出GBK范围的生僻字，应在系统层面使用伪码转义后，再写入数据库。从数据库读取数据时，应进行伪码还原成汉字。

c) 可采用 UTF-8 编码的数据库包括但不限于下述范围：

1) 使用 MySQL 数据库时，应采用 5.5.3 以上版本，并将 UTF-8 的编码类型设置为 UTF8MB4。

2) 使用 Oracle 数据库时，字符集可设置为 AL32UTF8 格式。

d) 其他类型数据库应使用 UTF-8、GB18030 或 ISO 8859-1 等支持全字符集的编码。

11.2 文件存储及处理

本行的文件来源可分为内部和外部两个途径。

11.2.1 内部文件存储及处理

内部文件存储编码字符集采用UTF-8编码。本行的卸数规范应满足前述要求。

11.2.2 外部文件存储及处理

外部文件的来源主要有两个方式，一是与外部机构交互的数据文件，二是由数据供应商提供的数据文件。外部文件不满足UTF-8或GB18030编码的，优先做格式转换。在对端无法更换情况下，宜与外部系统协商，使用码值转义处理生僻字，避免生僻字丢弃或转为无意义的替代字符（如问号、空格等）。

11.3 系统内部处理

11.3.1 联网核查居民身份信息

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应做兼容处理。

11.3.2 业务处理中的姓名比对

对于“一字多码”的生僻字，系统应支持一字多码互相认同的智能比较。

附　录　A

（规范性）

补充要求

A.1　生僻字处理成熟度评估

提供金融服务的机构及参与金融服务信息交换的机构的生僻字处理成熟度宜按以下几个维度进行评估，见表A.1。

表A.1　生僻字处理成熟度评估表

领域	子领域	评估办法	
系统字符集支持	显示 输入 打印 存储	字集	评分示例（分）
		GBK	2
		GB 18030—2005	3
		+《通用规范汉字表》	3.5
		+人口信息字库	4
		+UCS 全字集	5
“一字多码”兼容处理	存储结果	能对“一字多码”兼容处理	
	联网核查		
	培训机制		

A.2《信息技术 中文编码字符集》（GB 18030-2022）实现级别

A.2.1 实现级别 1

实现级别1支持本文件的单字节编码部分、双字节编码部分和四字节编码部分的CJK统一汉字扩充A（即0x8139EE39—0x82358738）。

A.2.2 实现级别 2

实现级别2包含实现级别1。此外，实现级别2还支持《通用规范汉字表》中的没有包含在实现级别1 之内的编码汉字。《通用规范汉字表》所收汉字在本文件中的代码位置和字形。

系统软件及支撑软件，应至少满足实现级别2的要求。系统软件及支撑软件包括但不限于操作系统、数据库管理系统、中间件（软件产品分类按照GB/T 36475中的规定）。

A.2.3 实现级别 3

实现级别3包含实现级别2。此外，实现级别3还支持本文件规定的全部汉字及表3中的康熙部首。

用于政务服务和公共服务的产品应满足实现级别3的要求。政务服务和公共服务行业包括但不限于 铁路运输业、道路运输业、水上运输业、航空运输业、多式联运和运输代理业、邮政业、货币金融服务、保险业、土地管理业、卫生、国家机构、社会保障等（行业分类按照GB/T 4754中的规定）。

A.3《金融服务生僻字处理指南（JRT 0253-2022）》生僻字处理分级

本文件将生僻字处理分为以下三个级别。

a）基础级：

——CJK 基本集和扩充 A，其中包含 52 个 GBK 双码字。

——《通用规范汉字表》全部汉字（含补字区、CJK 扩充 B～E 共 199 个字）。

——人口信息字库 PUA 编码部分对应的正式编码汉字（含 CJK 扩充 B～G）。

b）实用级（包含基础级，增加实际在用汉字）：

——CJK 扩充 B～G 中已知的人名、地名在用汉字。

——人口信息字库 PUA 编码部分，虽有正式编码但仍在用 PUA 编码的汉字。

——人口信息字库 PUA 编码部分，没有正式编码只能使用 PUA 编码的汉字。

c）完整级：UCS 收录的全部 CJK 汉字，包含实用级。