

中华人民共和国金融行业标准

JR/T 0253—2022

---

金融服务 生僻字处理指南

Financial services—Guidelines for processing rarely used Chinese characters

2022 - 6 - 24 发布

2022 - 6 - 24 实施

---

中国人民银行 发布

## 目 次

前言.....	IV
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.....	3
5 总体原则与策略.....	3
5.1 总体原则.....	3
5.2 信息系统处理汉字的通用架构.....	4
5.3 生僻字处理分级.....	4
5.4 生僻字处理策略.....	4
6 生僻字的输入.....	4
6.1 输入法.....	5
6.2 机读设备输入.....	5
6.3 其他方法输入.....	5
6.4 信息系统输入配备.....	6
7 生僻字的显示.....	6
7.1 字库.....	6
7.2 信息系统字库的配备.....	7
8 生僻字的打印.....	7
8.1 柜台 PC 通用打印机.....	7
8.2 报表高速打印机.....	8
8.3 打印机字库升级方法.....	8
9 生僻字的信息交换.....	9
9.1 机构内部系统间的信息交换.....	9
9.2 机构与外部系统的信息交换.....	10
10 生僻字的存储及内部处理.....	10
10.1 数据库存储.....	10
10.2 文件存储.....	11
10.3 系统内部处理.....	11
11 内部培训和投诉处理.....	11
12 生僻字处理方法的开源.....	12
附录 A（资料性）引用方法和示例.....	13
A.1 生僻字处理成熟度评估.....	13
A.2 UCS 汉字编码概况.....	13
A.3 GBK 52 个双码字.....	14
A.4 人名用生僻字全字符集示例.....	15

A.5 常见编码和伪码格式比较.....	16
A.6 常用字符集“实心点”字符的编码.....	17
附录 B（资料性）生僻字问题改造实例.....	18
B.1 实例 1——中信银行全系统生僻字改造.....	18
B.2 实例 2——中国银联全渠道系统生僻字改造.....	19
B.3 实例 3——中国农业银行核心银行系统生僻字改造.....	19
参考文献.....	21

## 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。

本文件由中国人民银行科技司提出。

本文件由全国金融标准化技术委员会（SAC/TC 180）归口。

本文件起草单位：中国人民银行科技司、北京金融科技产业联盟、招商银行股份有限公司、中信银行股份有限公司、中国工商银行股份有限公司、中国农业银行股份有限公司、中国银行股份有限公司、中国建设银行股份有限公司、建信金融科技有限责任公司、中国人民银行营业管理部、中国人民银行济南分行、中国人民银行重庆营业管理部、中国人民银行杭州中心支行、中国人民银行广州分行、中国人民银行长沙中心支行、中国人民银行乌鲁木齐中心支行、交通银行股份有限公司、中国邮政储蓄银行股份有限公司、中国科学院软件研究所、中国金融电子化集团有限公司、成方金融科技有限公司、北京国家金融标准化研究院有限责任公司、中国银联股份有限公司、北京银联金卡科技有限公司、重庆国家金融科技认证中心有限责任公司、北京北大方正电子有限公司、北京郑码世纪信息技术有限公司。

本文件主要起草人：李伟、杨富玉、聂丽琴、纪熙东、马良有、曲维民、冯蕾、刘子群、刘江涛、胡达川、李寻、李言平、徐晓剑、孙炎森、梁宇、柯尚锋、杨启龙、李学鹏、张立建、王丽静、王学群、郭贞、柏杨、邱程昱、江山、马懿、赵磊、马征、刘妍、韩婷婷、刘启滨、刘赐麟、杨志、孙坚、叶茜、张伟宁、胡沐创、谭旺、刘曼齐、戴雪龙、许健、张嘉威、谢谨、潘以桢、谢晋、张兰英、胡军锋、张兰英、朱礼华、刘汇丹、刘书元、孙建智、李家琪、陈达炜、谢彦丽、白璐、邱鹏、缪海波、王琪、于鸽、李博文、李远、史艳语、毕小文、秦逞、吴娟、张建国、张国荣、陈恳、郑珑、蓝飞。

## 引 言

随着经济社会数字化程度越来越高，以及实名制要求越来越严格，姓名中含有生僻字的客户在办理金融业务时，因输入、显示、打印、存储、交换等一个或多个环节中无法正常处理生僻字，可能造成的障碍包括以下内容。

- a) 身份证鉴别仪读取客户证件信息失败，无法完成联网核查。
- b) 服务人员知道客户姓名，但使用通用输入法找不到相应汉字。
- c) 跨行转账户名一字多码，户名不能准确匹配，无法完成自动入账。
- d) 信息交换时户名被当作非法字符或被转换成“？”，无法正确识别。
- e) 与银行往来的第三方支付、社保、证券、保险等业务无法正常实名处理。

本文件旨在针对上述情况，提供金融业处理生僻字的通用方法指南，提高金融业信息系统对生僻字的处理能力，提升金融业服务水平。

本文件内容可能涉及信息系统关联的注册公司、产品名称或商标，仅作一般描述使用，无意侵权，更不表示推荐或不推荐相关产品。



# 金融服务 生僻字处理指南

## 1 范围

本文件提供了金融业信息系统生僻字处理指南，包括生僻字处理总体原则与策略，生僻字的输入、显示、打印、信息交换、存储和内部处理方法，以及生僻字内部培训和投诉处理、生僻字处理方法开源的机制。

本文件适用于为客户提供金融服务的机构及参与金融服务信息交换的机构。

## 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 18030 信息技术 中文编码字符集

GB/T 2312—1980 信息交换用汉字编码字符集 基本集

GB/T 13000 信息技术 通用多八位编码字符集（UCS）

ISO/IEC 10646 信息技术 通用编码字符集（UCS）（Information technology — Universal Coded Character Set (UCS)）

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: Latin alphabet No.1）

## 3 术语和定义

下列术语和定义适用于本文件。

### 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，通常也指按照这种规则确定的文字的有序集合。

**示例：**1. GB 18030 是我国制订的以汉字为主并包含多种我国少数民族文字（例如藏、蒙古、傣、彝、朝鲜、维吾尔文等）的大型中文编码字符集标准，业界对该标准的全字符集的技术实现一般称作 GB18030 编码，该编码包含单字节字符、双字节字符、四字节字符，其中双字节字符编码的技术实现又称作 GBK 编码。

2. EBCDIC 是大型主机的 8 比特单字节或双字节编码字符集。

**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。

[来源：ISO/IEC 8859-1:1998，4.5，有修改]

### 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。

- 示例：1. “1388”表示GB 18030中的强制部分用EBCDIC编码实现的字符集。  
2. “1392”表示GB 18030中的强制部分的字符集。

### 3.3

#### 字库 font library

建立在计算机存储媒体上的字形数据集合。

- 注：1. 字库在存储方式上一般分硬字库、软字库。硬字库指预烧录在只读存储器等介质中且不能再更改的字库，软字库指以文件形式存储在光盘或者硬盘上的字库。  
2. 字库一般以ttf、otf、ttc等字体格式文件的形式存在。ttf指True Type Font字体文件，otf指Open Type Font字体文件，ttc指True Type Collection字体文件。

### 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+F000～U+FFFFD、U+10000～U+10FFFFD。

- 注：1. 一般指人口信息字库中使用的PUA编码，在人口信息字库中，户籍管理部门对未收录进GB/T 13000但实际使用的生僻汉字利用PUA编码予以补充，人口信息字库通过转换对照表提供编码转换解决方案。  
2. 人口信息字库中部分PUA编码字符陆续被通用编码字符集收录而拥有正式编码，会导致一个字符同时存在正式编码和PUA编码，造成一字多码的情况。

[来源：GB 18030，3.3，有修改]

### 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GB 13000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“癸”“鸪”“镬”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/IEC 10646，最新版本为ISO/IEC 10646:2020。我国现行GB/T 13000—2010采标自ISO/IEC 10646:2003。  
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。

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）。
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（见附录A.2）。
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位以四字节表示。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”（Big Endian, 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。
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+FFFF称为基本多文种平面（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)
- CTID: 网络电子身份证 (Cyber Trusted ID)
- 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)
- GDI: 图形设备接口 (Graphics Device Interface)
- HTML5: 超文本标记语言第5版 (HyperText Markup Language 5)
- JDK: Java语言开发工具 (Java Development Kit)
- MFC: 微软基础类库 (Microsoft Foundation Classes)
- OCR: 光学字符识别 (Optical Character Recognition)
- PC: 个人电脑 (Personal Computer)
- PUA: 用户自定义区 (Private Use Area)
- SDK: 软件开发工具 (Software Development Kit)
- UCS: 通用编码字符集 (Universal Coded character Set)
- XML: 可扩展标记语言 (Extensible Markup Language)

## 5 总体原则与策略

### 5.1 总体原则

提供金融服务的机构在处理生僻字时，宜遵守以下原则。

- 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。

- 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。
- 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。
- 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，同时向客户详细说明一字多码情况。

注：部分居民身份证件姓名数据包含的字符分布在UCS的CJK扩充A~G范围内，部分超出现行GB 18030—2005强制要求的字符集范围。

## 5.2 信息系统处理汉字的通用架构

信息系统处理汉字的通用架构见图 1，包括客户与柜台、客户与前置中台、柜台与前置中台、前置中台与后台系统、后台系统与外联系统、外联系统与其他机构等交互环节。在客户与柜台、客户与前置中台交互环节，输入、显示、打印处理涉及生僻字。在柜台与前置中台交互环节，流程、交换处理涉及生僻字。在前置中台与后台系统、后台系统与外联系统交互环节，开放、主机系统处理涉及生僻字。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。

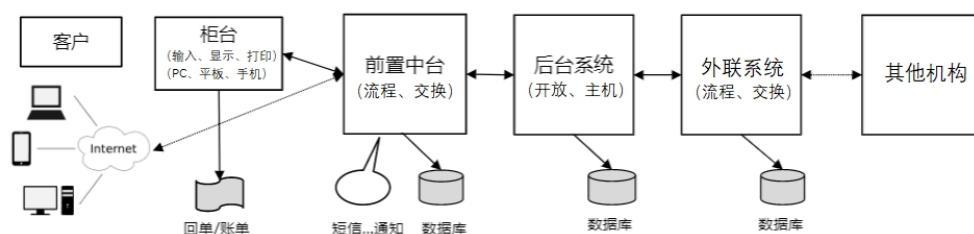


图 1 信息系统处理汉字的通用架构图

信息系统通常需要在 GBK、GB18030、EBCDIC、UTF-8、UTF-16 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、EBCDIC 不支持的汉字需实现无损透传处理。

Unicode 字符编码详见附录 A. 2。

## 5.3 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。

## 5.4 生僻字处理策略

生僻字的显示和打印宜通过操作系统支持完整级汉字，生僻字的输入宜支持实用级汉字。

## 6 生僻字的输入

## 6.1 输入法

### 6.1.1 输入法字符范围

提供金融服务的机构宜选择实用级或完整级的汉字输入法，至少支持基础级的汉字输入法。

### 6.1.2 输入法编码

适用于生僻字输入的输入法宜采用以下方法。

- a) 使用拼音、笔画、字形等方法对汉字字符进行编码。
- b) 综合使用部件拆分、拆字拼音、笔画等多种方法对汉字字符进行编码。
- c) 按照汉字的字频、所处编码区块及其他属性对候选汉字进行排序。
- d) 对候选的 PUA 编码字、繁体字、异体字等给出标识，进一步提示其对应的正式编码字、简体字、规范字。

注：生僻字读音通常难以识别，完整级汉字如使用拼音输入，因同音候选字最多可达3,000多个，造成查找和选字困难。

### 6.1.3 输入法实现形式

提供金融服务的机构选用的输入法，可使用如下形式实现。

- a) 常规输入法软件。通过外接键盘或软键盘输入字符的输入法软件，可配置在操作系统的输入法候选列表中并可切换选择。
- b) 第三方软件。使用带有字符输入功能的第三方软件，用户通过软件界面操作以笔画、部件拆字等方式查询到候选字，使用拷贝或其他方式粘贴到信息系统的录入框中。
- c) 内嵌输入法。在信息系统中自行实现的输入法。
- d) 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输入码，云输入客户端根据输入码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。

### 6.1.4 少数民族姓名间隔符的输入

少数民族姓名间隔符须按照《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号文）要求的格式输入，统一用“·”（UCS编码U+00B7，GB18030编码A1A4）。考虑到常用字符集中“实心点”字符有多个，宜在用户输入的前端检测少数民族姓名间隔符为非U+00B7的“实心点”时，自动转换成U+00B7。

注：有些文字处理软件中U+00B7复制到其他应用有可能变成U+2022。

### 6.1.5 括号的输入

企业名称等信息中括号的输入宜兼容处理全角和半角。

## 6.2 机读设备输入

提供金融服务的机构选用的身份证件识读设备，宜使用GB18030编码、UTF-8编码、UTF-16编码的解码驱动程序进行识读，不宜使用GBK编码的解码驱动程序。外国人永久居留证件识读设备、CTID的二维码读取器等，宜达到同等要求。

## 6.3 其他方法输入

提供金融服务的机构使用OCR、语音识别、手写识别等方法识别汉字时，宜以人工核对、修正后的结果为准。

注：由于识别引擎使用的字符集通常未收录全部汉字，同音字或形近字较多，准确率难以达到100%。

## 6.4 信息系统输入配备

### 6.4.1 PC 输入配备

PC的输入配备宜实现以下要求。

- a) 按照 6.1 配备输入法。
- b) 对于有 PUA 编码标识但所用编码与人口信息字库 PUA 编码不同的输入法，宜关闭录入 PUA 编码汉字的功能。
- c) 不宜使用含有 PUA 编码汉字录入功能但没有 PUA 编码标识的输入法。
- d) 若配备身份证件识读设备，宜参考 6.2。
- e) 若使用 OCR、语音识别、手写识别技术，宜参考 6.3。
- f) 信息系统宜对姓名、地址等栏位输入的字符做检测。字符集限制宜以最新 UCS 汉字编码范围为准，不宜使用 U+4E00~U+9FA5 的范围来控制只输入 GBK 汉字。若输入的字符包含“？”，宜记录日志，并提示服务人员和客户可能存在客户姓名不匹配的情况，相关业务宜转为人工处理；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，若在公安系统身份实名核查失败，宜向客户详细说明一字多码情况。
- g) 面向客户的网页（Web）应用中，宜在需要录入姓名、地址的页面内嵌云输入法。
- h) 若信息系统无法显示所输入生僻字的字形，不宜用拼音代替，宜使用 UCS 编码代替，以保证汉字编码的唯一性。

### 6.4.2 APP 输入配备

APP宜提供内嵌输入法、云输入法以支持生僻字的输入。

- a) 对于使用 HTML5 页面实现的 APP，宜提供内嵌云输入法。
  - b) 对于其他 APP，宜提供内嵌本地输入法或云输入法。
- APP 宜对姓名、地址等栏位输入字符做检测，并参考 6.4.1 中 f) 给出相应提示。

### 6.4.3 ATM、智能终端等自助设备输入配备

ATM、智能终端等自助设备的输入宜实现以下要求。

- a) 若通过实体键盘或软键盘录入，宜按照 6.1 选择输入法。
- b) 若配备身份证件识读设备，宜按照 6.2 选择相关设备。
- c) 若配备 OCR、语音识别、手写识别输入设备，宜参考 6.3。
- d) 信息系统宜对姓名、地址等栏位输入字符做检测，并参考 6.4.1 中 f) 给出相应提示。

## 7 生僻字的显示

### 7.1 字库

#### 7.1.1 字符范围

字符范围宜覆盖基础级和实用级汉字，鼓励提供金融服务的机构覆盖完整级汉字。

#### 7.1.2 字形

PUA 编码字与正式编码字的字形宜作出明显区分，或在字形外以其他标记区分。

生僻字的字形品质宜提高到精品出版印刷用字库的水平，即字形美观、结构端正、重心平稳、大小一致、黑白均匀。

### 7.1.3 字形描述技术

针式打印机配置的硬字库使用点阵字形技术。

其他情况使用曲线轮廓字形技术，一般使用ttf格式或otf格式。

### 7.1.4 兼容性

宜兼容提供金融服务的机构使用的所有操作系统。

单个ttf文件或otf文件因受65,536个字形限制，宜用实用级字库以便能跨平台兼容处理。

### 7.1.5 实现形式

字库可以有以下3种实现形式。

- a) 本地字库。安装在操作系统上，所有应用均可使用的字库。
- b) 应用嵌入字库。随APP安装，并仅限于APP内使用的字库。
- c) 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。

## 7.2 信息系统字库的配备

### 7.2.1 PC字库的配备

PC本地字库配备宜符合7.1要求。信息系统和显示生僻字的控件宜支持UCS编码。支持生僻字显示的方法如下。

- a) 通过辅助软件对操作系统字形显示逻辑进行配置，从操作系统层面解决生僻字显示问题，以便信息系统在用缺省字体（如宋体）的情况下能够显示生僻字。
- b) 通过使用控件绑定生僻字字库的方式，实现生僻字显示。

在遇到字库缺字时宜以可识读的标准UCS编码方式显示其字形，占一个汉字宽度位置。以“张□韦华（U+2E9F5）三”为例，缺字时以UCS编码显示字形见图2。




图2 缺字时以UCS编码显示字形

### 7.2.2 APP字库的配备

APP宜通过应用嵌入字库或云字库的方式实现对生僻字的显示。

APP内部显示人口信息生僻字的控件宜绑定生僻字字库。

### 7.2.3 ATM、智能终端等自助设备字库的配备

ATM、智能终端等自助设备宜预装本地字库。

自助设备中基于Web技术实现的应用也可利用云字库技术实现对生僻字的显示。

## 8 生僻字的打印

### 8.1 柜台PC通用打印机

柜台PC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	优点	缺点
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。	a) 免硬件升级。 b) 打印速度快。	信息系统改造复杂。
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。	日常办公类的非针式打印机，如激光打印机、喷墨打印机等。	a) 字库依赖操作系统，与打印机硬字库无关。 b) 信息系统改造方案简单，依赖打印机驱动即可。	对于传统串口、并口打印机速度较慢。
纯文本方案	升级存折打印机字库，字库支持完整级汉字。	带硬字库的针式打印机，如存折打印机、宽行打印机等。	打印速度快。	需升级硬字库，后续升级困难。

旧有的串口、并口打印机宜逐步更新为USB接口或本地网络接口的打印机。每个服务网点宜至少配备一台支持实用级或完整级汉字集合的打印机。

## 8.2 报表高速打印机

使用文本方式打印的报表高速打印机，宜由厂家升级字库以支持实用级或完整级汉字。

## 8.3 打印机字库升级方法

### 8.3.1 硬字库

硬字库宜以可改写存储介质形式存储，方便及时更新。

硬字库存储空间测算：通用打印一般使用 24×24 点阵字库，单个汉字字形数据占 72 字节（B），以 50,000 个汉字为例，需要 50,000×72B=3,600,000B≈3.5MB。

注：此方式有一定的硬件成本，扩展升级工作量较大，但打印速度较软字库更快。

### 8.3.2 软硬字库混合

需要结合打印机硬字库汉字范围，使用混合指令来支持，对不同厂商打印机图形指令集支持，在开发上有一定的复杂度和工作量。生成汉字图形方法如下。

- a) 判断待打印文本中是否含有生僻字，若有则转图形模式打印，否则用字符模式打印。
- b) 设置图片要素，为确保图片打印效果尽量接近字符打印效果，生成图片需要做特殊处理，用不同的字体或拉伸图片解决：
  - 默认宋体字体，支持的汉字范围更广；
  - 倍高拉伸，先生成正常字体，然后拉伸高度至原来的 2 倍，再合并图片；
  - 倍宽拉伸，先生成正常字体，然后拉伸宽度至原来的 2 倍，再合并图片。
- c) 生成图片，可使用系统函数或方法调用生成图片。

### 8.3.3 软字库

如果打印机能支持图形打印，只要在操作系统下安装字库，即可选择其中任意字符进行打印。

示例：基于 GDI 等绘图引擎，将仿真指令打印方式转换为基于绘图引擎的图形打印，在绘图时使用支持生僻字的字库绘制出图形，再转换为点阵数据并使用系统驱动打印出凭证内容。该方式生成的打印点阵图较大，发送给打印机数据较大，对打印速度有一定影响，但有利于扩展字库，只要操作系统进行了字库扩展，就能够支持新的生僻字。

## 9 生僻字的信息交换

### 9.1 机构内部系统间的信息交换

#### 9.1.1 概述

宜支持 GB 18030 或 GB/T 13000（一般用 UTF-8 编码）的汉字无损透传处理。

在交换之前，请求方宜采用生僻字的正式标准汉字编码进行归一化处理。如果原内部系统间接口为 GBK 或 EBCDIC 等小字符集的编码，且改造成本过大，可以保留，此时生僻字宜改用伪码表示和交换。


转接系统在转接时，如果输入、输出双方编码不同，需要做编码转换时，宜注意下述情况。

- a) 不发生丢弃某些字符或转成替代符“?”的有损转换。
- b) 不发生报文丢弃的情况。

#### 9.1.2 报文

提供金融服务的机构内部系统间交换的报文常见格式和注意事项见表 2。

表2 提供金融服务的机构内部系统间交换报文常见格式和注意事项

报文格式	注意事项
字段定长 无分隔符	生僻字改用伪码表示和交换时，1 个四字节的 GB18030 编码汉字用伪码可能需要占用更多存储空间，如果原先字段定义长度不够，则需考虑超长问题。
变长字段 有分隔符	分隔符不宜与伪码的特殊符号相同，且不宜与汉字编码的一部分相同。 如果字段分隔符是“ ”，则与某些汉字编码的一部分冲突，如“毀”字的 GBK 编码第 2 字节是 16 进制的 0x7C，与常用竖线分隔符“ ”的编码一样。
XML	需注意头部的 encoding 编码设置须与内容采用的编码一致，以免 XML 解析器解码错误，如用 UTF-8 编码或 GB18030 编码，头部宜改为： <pre>&lt;?xml version="1.0" encoding="UTF-8"?&gt;</pre> 或 <pre>&lt;?xml version="1.0" encoding="GB18030"?&gt;</pre>
JSON（一种 数据格式）	JSON 使用 UTF-8 编码，不支持 GB18030 编码。可使用转义方式表示特殊字符，包括回车（\u000D）、换行（\u000A）等不可见字符，BMP 之外的辅助平面字符采用代理对转义字符串表示，如 U+20164（  亩心）字表示为“\uD840\uDD64”（参见 RFC 8259: The JavaScript Object Notation (JSON) Data Interchange Format），所以，BMP 外生僻字若不用伪码表示时，宜使用前述转义字符串方式表示，不宜用 UTF-8 的四字节表示，以免某些 JSON 解析器解析错误。

#### 9.1.3 文件

文件内容的格式及注意事项见 9.1.1。

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。BOM标记与编码方式的对应关系如下。

- a) 在 UTF-16LE 中 BOM 为 0xFF FE，UTF-16BE 中 BOM 为 0xFE FF。
- b) 在 UTF-32LE 中 BOM 为 0xFF FE 00 00，UTF-32BE 中 BOM 为 0x00 00 FE FF。
- c) 在 UTF-8 带 BOM 的版本中，BOM 为 0xEF BB BF。

注：在某些操作系统自带记事本编辑保存时，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。

#### 9.1.4 信息交换转码

信息系统间使用不同编码字符集进行数据交换时，需要进行无损透明转码，可借助于中间件(如FTP)或自行编写程序进行转码。

以FTP方式交换文件不需要转码时，可设定为二进制流方式；如需转码时，宜设定相应的编码集，以保证无损透传。

### 9.2 机构与外部系统的信息交换

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，使用伪码处理生僻字，应避免生僻字丢弃或转为无意义的替代字符（如问号、空格等）。

当外部系统无法正常显示或打印生僻字时，宜在打印凭证的对应位置手写相应汉字。

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（如“=?GBK?B?”或“=?GB2312?B?”），宜使用UTF-8（即“=?utf-8?B?”开头）。

注：业界对ISO/IEC 8859-1的技术实现一般称作ISO8859-1编码。

## 10 生僻字的存储及内部处理

### 10.1 数据库存储

#### 10.1.1 概述

数据库存储宜使用 UTF-8 编码，其次使用 GB18030 编码。

如使用 GBK 编码，在不改变数据库字符集设置的情况下，对超出 GBK 范围的生僻字，宜在系统层面用伪码编码后，再写入数据库。从数据库读出数据时，宜将伪码还原成汉字。

伪码宜使用易于还原、占用空间小的 UCS 编码，伪码格式参考附录 A.5。

伪码仅限在数据库内部使用，外部访问宜还原为接口标准编码，以保证透传、通用。

少数民族姓名的间隔符宜按 6.1.4 的规范形式存储，不规范的旧数据宜定期迁移。

注：下文所列 MySQL、DB2、Oracle 均指数据库产品名称。

#### 10.1.2 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。

注：utf8mb4 编码是 utf8mb3 编码的超集，兼容 utf8mb3 并且能够存储四字节 UTF-8 的字符。

#### 10.1.3 DB2 数据库

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伪码处理。

开放平台 DB2 数据库宜升级支持 UTF-8 或 GB18030 编码。



注：目前大型主机系统使用 CCSID 1388 编码，汉字使用双字节表示，支持至 CJK 扩充 A 字符集。

#### 10.1.4 Oracle 数据库

Oracle 数据库宜将字符集值设置成 AL32UTF8。

注：目前 Oracle 数据库字符集默认值为 ZHS16GBK，仅支持 GBK 字符集。

#### 10.1.5 其他数据库

其他数据库宜使用 UTF-8、GB18030 或 ISO 8859-1 等支持全字符集的编码。

### 10.2 文件存储

文件存储宜采用 UTF-8 或 GB18030 编码。

### 10.3 系统内部处理

#### 10.3.1 编程语言对生僻字的处理

常用编程语言宜对生僻字特殊情况做以下处理。

- a) 在使用 JAVA（一种面向对象的计算机编程语言）编程时，宜使用 JDK1.5 及以上版本。当字符串包含 BMP 外字符时，不宜使用“str.length()”方法计算字符个数，宜使用“str.getBytes().length”方法，依据输入的 GBK、GB18030、UTF-32 等编码参数返回字节数进行判断。
- b) 使用 VC++（一种计算机编程语言开发平台）MFC 编写的应用程序，宜在“\_Unicode”模式下编译，不宜在“\_MBCS”模式下编译。

#### 10.3.2 联网核查居民身份证信息

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宜做兼容处理。

对于“一字多码”的生僻字，当有正式编码存在时，宜向客户详细说明一字多码情况。

对于联网核查未通过情况，必须按照《关于切实做好联网核查公民身份信息有关工作的通知》（银发〔2007〕345 号文）要求采取其他方式进一步核实身份。

注：1. “一字多码”的生僻字存在二代居民身份证芯片与人口信息数据库的生僻字编码不一致的问题。

2. 部分“一字三码”的生僻字如表 3 所示，可能需要核查 3 次才能成功。

表3 “一字三码”生僻字示例表

生僻字，拆字	PUA 码 1	PUA 码 2	UCS 正式编码
𠂇, 𠂇从	U+E3FE	U+E579	U+2B4E9 (扩充 C)
𠂇, 𠂇才为	U+E05D	U+F429	U+39D1 (扩充 A)
懋, 懋心	U+E56B	U+EAF0	U+2285F (扩充 B)

#### 10.3.3 业务处理中的姓名比对

对于“一字多码”的生僻字，系统宜支持一字多码互相认同的智能比较。

对于系统不支持处理生僻字的情况下，宜转人工处理，需要时可联系客户核实处理。

## 11 内部培训和投诉处理

提供金融服务的机构及参与金融服务信息交换的机构宜建立处理生僻字的培训、业务操作、客服投诉等管理机制，配备专门的生僻字处理人员，设立沟通交流小组开展经验分享，提升服务人员、技术支持人员对生僻字的认知和处理水平。可采取以下措施。

- a) 在机构内部建立生僻字专业知识库，定期收集、分析生僻字客户的投诉以及处理过程，开展内部分享和交流。
- b) 开展培训，使服务人员了解汉字信息处理标准、编码知识及生僻字输入、查询工具使用方法等知识。

## 12 生僻字处理方法的开源

提供金融服务的机构及参与金融服务信息交互的机构可参考金融业生僻字开源项目或其他方式完成系统改造。

提供金融服务的机构及参与金融服务信息交互的机构宜积极参与金融业生僻字开源项目，分享生僻字处理相关的代码和方案。提供金融服务的机构及参与金融服务信息交互的机构开展开源项目活动时，应遵循开源社区的规章制度。

鼓励字库厂商、输入法厂商及其他机构共同参与开源社区建设。

注：金融业生僻字开源项目官方网站是金融业生僻字信息平台，访问地址为“[rucc.org.cn](http://rucc.org.cn)”。

附 录 A  
(资料性)  
引用方法和示例

### 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		
人员认知和服务	培训机制 柜员、客户服务 指导客户正确用字 疑难处理及知识共享	培训课件、服务手册包含相关内容	
客户教育	网站指引 工具提供 定向宣传	能通过系统或人工方式获取相关信息及工具	
行业知识贡献	共享知识	能反馈生僻字案例、参与同业交流	
	标准定制	参与行业、团体、企业标准制修订	
	工具提供	能提供或应用生僻字处理开源工具	

### A.2 UCS 汉字编码概况

在本文件编写时，已使用 UCS 汉字编码范围，可用于字库等编码限制参考，见表 A.2。

表A.2 已用UCS汉字编码表

字符集 (含其他区域)	编码范围	字数(个)	版本	备注	对应国内标准
CJK UR0	U+4E00~U+9FA5	20,902	1.0	1993年发布。	GBK (21,003字)
CJK Ext A	U+3400~U+4DB5	6,582	3.0	1999年发布。	2000年版本 GB 18030
CJK Ext B	U+20000~U+2A6D6	42,711	3.1	2001年3月1日发布。	2005年版本 GB 18030

表A.2 已用UCS汉字编码表（续）

字符集 (含其他区域)	编码范围	字数(个)	版本	备注	对应国内标准
CJK Ext C	U+2A700~U+2B734	4, 149	5.2	2009年10月1日发布。	-
CJK Ext D	U+2B740~U+2B81D	222	6.0	2010年10月1日发布。	-
CJK Ext E	U+2B820~U+2CEAF	5, 762	8.0	2015年6月1日发布。	-
CJK Ext F	U+2CEB0~U+2EBE0	7, 473	10.0	2017年6月20日发布。	-
CJK Ext G	U+30000~U+3134A	4, 939	13.0	2020年3月3日发布。	-
URO+	U+9FA6~U+9FFF	90	14.0	基本集尾, 补字区。	-
CJK Ext A+	U+4DB6~U+4DBF	10	13.0	扩充A尾, 补字区。	-
CJK Ext B+	U+2A6D7~U+2A6DF	9	14.0	扩充B尾, 补字区。	-
PUA	U+E000~U+F8FF	约 4, 700	-	人口信息字库。	-
符号	U+00B7	1	-	标准姓名分隔符。	-
	U+FF00~U+FFEF	240	-	全角字母及符号。	-
	U+4DC0~U+4DFF	64	-	易经64卦符号, 宜禁用。	-
兼容区	U+F900~U+FAFF	512	-	CJK兼容区, 宜禁用。	-
	U+2F800~U+2FA1D	542	-	CJK兼容区, 宜禁用。	-

## A.3 GBK 52个双码字

GBK规范在PUA区域编码了52个汉字（见表A.3，最后一列为PUA编码列对应的GB18030编码），后被GB 18030正式收录为扩充A区域。

表A.3 GBK 52个双码字

汉字	GBK	Unicode	EBCDIC	PUA	GB18030
儻	FE55	3473	CE5B	E81A	8336C936
佻	FE56	3447	CE5C	E81B	8336C937
喝	FE5A	359E	CE60	E81F	8336CA30
囍	FE5B	361A	CE61	E820	8336CA31
囍	FE5C	360E	CE62	E821	8336CA32
憐	FE5F	396E	CE65	E824	8336CA35
恹	FE60	3918	CE66	E825	8336CA36
捌	FE62	39CF	CE68	E827	8336CA37
扞	FE63	39DF	CE69	E828	8336CA38
搜	FE64	3A73	CE6A	E829	8336CA39
捌	FE65	39D0	CE6B	E82A	8336CB30
柄	FE68	3B4E	CE6E	E82D	8336CB31
殍	FE69	3C6E	CE6F	E82E	8336CB32
汰	FE6A	3CE0	CE70	E82F	8336CB33
睽	FE6F	4056	CE75	E834	8336CB36
穆	FE70	415F	CE76	E835	8336CB37

表A.3 52个GBK双码字（续）

汉字	GBK	Unicode	EBCDIC	PUA	GB18030
紬	FE72	4337	CE78	E837	8336CB39
糶	FE77	43B1	CE7D	E83C	8336CC33
糶	FE78	43AC	CE7E	E83D	8336CC34
膊	FE7A	43DD	CE81	E83F	8336CC36
勞	FE7B	44D6	CE82	E840	8336CC37
禛	FE7C	4661	CE83	E841	8336CC38
禛	FE7D	464C	CE84	E842	8336CC39
沂	FE80	4723	CE86	E844	8336CD30
讌	FE81	4729	CE87	E845	8336CD31
賄	FE82	477C	CE88	E846	8336CD32
賄	FE83	478D	CE89	E847	8336CD33
鎬	FE85	4947	CE8B	E849	8336CD35
钶	FE86	497A	CE8C	E84A	8336CD36
钷	FE87	497D	CE8D	E84B	8336CD37
鎬	FE88	4982	CE8E	E84C	8336CD38
鎇	FE89	4983	CE8F	E84D	8336CD39
鎈	FE8A	4985	CE90	E84E	8336CE30
鎉	FE8B	4986	CE91	E84F	8336CE31
闋	FE8C	499F	CE92	E850	8336CE32
闋	FE8D	499B	CE93	E851	8336CE33
阨	FE8E	49B7	CE94	E852	8336CE34
阨	FE8F	49B6	CE95	E853	8336CE35
廡	FE92	4CA3	CE98	E856	8336CE36
鮓	FE93	4C9F	CE99	E857	8336CE37
鮓	FE94	4CA0	CE9A	E858	8336CE38
鮓	FE95	4CA1	CE9B	E859	8336CE39
廡	FE96	4C77	CE9C	E85A	8336CF30
臄	FE97	4CA2	CE9D	E85B	8336CF31
鸱	FE98	4D13	CE9E	E85C	8336CF32
鸱	FE99	4D14	CE9F	E85D	8336CF33
鸱	FE9A	4D15	CEA0	E85E	8336CF34
鸱	FE9B	4D16	CEA1	E85F	8336CF35
鸱	FE9C	4D17	CEA2	E860	8336CF36
鸱	FE9D	4D18	CEA3	E861	8336CF37
鸱	FE9E	4D19	CEA4	E862	8336CF38
鸱	FE9F	4DAE	CEA5	E863	8336CF39

## A.4 人名用生僻字全字符集示例

部分在用人名生僻字全字符集见表A.4，可作为一般测试案例，以具备较完整的覆盖度。

繁体字以编码在 U+4E00~U+9FA5，可以映射为 GBK 的汉字为例。

表A.4 部分在用人名生僻字全字符集

字样	字集	UCS	UTF-8	EBCDIC	GB18030	备注	繁体	编码
癸	PUA	U+E863	EEA1A3	0xCEA5	8336CF39	52 个双码字	龔	U+9F91
	扩充 A	U+4DAE	E4B6AE	0xCEA5	FE9F	-		
契	PUA	U+E032	EE80B2	0x7673	AAD3	《通用规范汉字表》汉字 (扩充 A 6530 字)	-	-
	扩充 A	U+36C3	E39B83	0xD2C5	8230B731			
憇	PUA	U+E0FC	EE83BC	0x7782	ACE1	-	-	-
	扩充 B	U+20164	F0A085A4	-	9532A632	-		
鈞	PUA	U+E3FE	EE8FBE	0x7B94	FCF3	-	縱	U+9F91
	PUA	U+E579	EE95B9	0x7D97	A294	-		
	扩充 C	U+2B4E9	F0AB93A9	-	9838E139	-		
苧	PUA	U+E4AC	EE92AC	0x7C86	FEE5	-	苧	U+8504
	扩充 E	U+2C72C	F0AC9CAC	-	9932BD34	-		
顛	PUA	U+E43E	EE90BE	0x7BD4	FDD5	-	顛	U+9814
	扩充 E	U+2CC56	F0ACB196	-	9933C336	-		
璿	PUA	U+E362	EE8DA2	0x7AB4	FBB5	非《通用规范汉字表》汉字	璿	U+3F06
	扩充 E	U+2C386	F0AC8E86	-	9931DE30			
犇	PUA	U+E43B	EE90BB	0x7BD1	FDD2	-	犇	U+97E1
	扩充 F	U+2E9F5	F0AEA7B5	-	9939C539	-		
枲	PUA	U+EB71	EEADB1	-	8337A030	E9xx~无 GBK	-	-
	扩充 B	U+234C3	F0A39383	-	9632DD33	-		
曉	PUA	U+EC81	EEB281	-	8337BB32	-	-	-
	扩充 B	U+24A4A	F0A4A98A	-	96378E34	-		

#### A.5 常见编码和伪码格式比较

以“癸”“𠄎𠄎问”字为例的常见编码和伪码格式比较见表A.5。

表A.5 常见编码和伪码格式比较表

分类	简述	“癸”字编码	字节长度	“𠄎𠄎问”字编码	字节长度
UTF-8	网页和数据库常见。	0xE4 B6 AE	3	0xF0 AC 9C AC	4
UTF-16	数据库常见。	0x4D AE	2	0xD8 71 DF 2C	4
GB18030	用于兼容 GBK 数据。	0xFE 9F	2	0x99 32 BD 34	4
C	\u 前缀, UTF-16 UTF32。	\u4DAE	6	\u0002C72C	10
JAVA	\u 前缀, UTF-16 代理。	\u4DAE	6	\uD871\uDF2C	12
CSS3	\前缀, 4 位或 6 位 Hex, 必要时加空格, 可能引起误判。	\4DAE	5	\02C72C	7
XML_DEC	Html, 十进制。	&#19886;	8	&#182060;	9

表A.5 常见编码和伪码格式比较表（续）

分类	简述	“癸”字编码	字节长度	“卅卅”字编码	字节长度
XML_HEX	Html, 十六进制。	&#x4DAE;	8	&#x2C72C;	9
Z1	[U+xxxx], X 字符集。	[U+4DAE]	8	[U+2C72C]	9
Z2	`H 前缀, HEX 定长 5 位。	`H04DAE	7	`H2C72C	7

注：1. 以 ASCII 码为主的数据，使用 UTF-8 最省空间。以 BMP 汉字为主的数据，使用 UTF-16 最省空间。如伪码只用于扩充 B 及以上的少数生僻字，考虑边界可辨识性、标准码可读性，宜使用 Z1。

2. 为方便阅读，表 A.5 中“0x”前缀表示 16 进制内容。

3. X 字符集指在信息互联、交换时可以使用的字符集，由英文大写字母、小写字母、数字及特殊字符构成，其中特殊字符包括“ . , - \_ ( ) / = + ? ! & \* ; @ # : % [ ] (换行符) (回车符) (制表符) (空格符)”，最初在环球银行金融电信协会（SWIFT）使用，后来 XML 也使用此字符集。

4. 分类中的 C 为一种计算机编程语言，CSS3 指层叠样式表第三版，XML\_DEC、XML\_HEX 分别表示 XML 的十进制、十六进制表达形式，Z1、Z2 分别表示自定义的两种形式。

## A.6 常用字符集“实心点”字符的编码

常用字符集“实心点”字符的编码见表A.6。

表A.6 常用字符集“实心点”字符编码

序号	式样	GB18030	Unicode	UTF-8	EBCDIC	说明
1	•	A1A4	00B7	C2B7	4345	GB/T 2312—1980 第一区第 4 字。用于少数民族姓名间隔符的标准字符。
2	.	A3AE	FF0E	EFBC8E	424B	GB/T 2312—1980 第 3 区，对应 ASCII 半角字符小数点的全角符号。
3	·	A842	02D9	CB99	CD43	GBK 补充符号。
4	.	A971	FE52	EFB992	CDB2	GBK 补充符号。
5	.	2E	002E	2E	4B	ASCII 半角字符（英文小数点及句号）。
6	•	8136A631	2022	E280A2	-	1993 年 Unicode 1.1 增加，但 2005 年版本 GB 18030 发布时无该字形的符号。
7	.	8136A633	2024	E280A4	-	1993 年 Unicode 1.1 增加，但 2005 年版本 GB 18030 发布时无有字形的符号。
8	.	8136A634	2027	E280A7	-	常用符号。
9	·	8136D337	2219	E28899	-	1993 年 Unicode 1.1 增加，但 2005 年版本 GB 18030 发布时无有字形的符号。
10	·	8136E136	22C5	E28B85	-	1993 年 Unicode 1.1 增加的数学运算符号。
11	·	8130CA31	0387	CE87	-	希腊语和科普特语符号。
12	·	8133EE33	1427	E190A7	-	加拿大土语音节符号。
13	•	8134B731	16EB	E19BAB	-	古代北欧装饰用符号。

## 附录 B

### (资料性)

#### 生僻字问题改造实例

#### B.1 实例 1——中信银行全系统生僻字改造

##### B.1.1 背景

金融行业信息系统众多、技术栈复杂，对所有系统进行生僻字改造，面临实施成本高、实施周期长、升级风险大等问题。2020 年之前，中信银行核心系统的大型主机仅支持 EBCDIC 编码，该编码字符集所收录汉字规模相当于 2000 年版本 GB 18030，不支持 CJK 扩充 B 及以上区域的汉字。若将数据库字段改为 UTF-8 编码以支持全字集汉字，则上层应用程序几乎都要修改，改造难度高，工作量巨大。此外，不少外围系统内部、系统间接口、数据库等仍在使用的 GBK 编码，GBK 编码升级为 GB18030 编码或 UTF-8 编码的成本也很高。

中信银行柜面系统使用字库是操作系统自带的字库，PUA 生僻字和 CJK 扩充 C 及以上区域的汉字不能显示；存折打印机也只支持 2000 年版本 GB 18030 的 27,533 个汉字；网上银行、手机银行的字库依赖客户端。

##### B.1.2 改造方案

中信银行生僻字改造方案主要分为以下几个阶段。

- a) 制定改造方案。需要进行生僻字改造的信息系统众多，涉及核心系统、柜面系统、支付系统、网上银行、手机银行等多类重要系统。为降低系统改造难度及成本，降低业务风险，中信银行于 2017 年 5 月至 2018 年 9 月起草制定了完整的解决方案和推进计划。
- b) 开发通用 SDK。中信银行内信息系统运行环境涉及多种常见操作系统等，采用的开发语言有 C、JAVA 等，系统间接口使用了 EBCDIC、GBK、GB18030 和 UTF-8 等多种编码。为实现统一转码，加快各系统改造进度，中信银行开发了多个通用 SDK，包括主机系统上的姓名格式化函数、支持“一字多码”的姓名智能比对函数、外围系统 C 语言版本的转码工具库（支持 Linux、Windows 等操作系统）、外围系统 JAVA 语言版本的转码工具库、文件传输平台的转码优化组件等。
- c) 分期分批进行系统改造。涉及改造的信息系统共 89 个，按系统的相关性分 12 批进行改造。2019 年 5 月 31 日，第一批系统改造完成并上线。2020 年 4 月 1 日，最后一批系统改造完成上线后，统一打开生僻字开关。

##### B.1.3 改造成果

考虑到只有少数客户姓名或地址含生僻字，在 EBCDIC 这类小字符集编码中采用了编码扩展转义字符串方式来表示生僻字，通过转码工具，应用系统可实现系统原本不支持的生僻字的转码，解决了生僻字在大字符集编码转小字符集编码时转为问号或丢失等问题。

对于采用 GBK 编码的系统，涉及数据量大、程序改动量小的系统尽可能升级为 GB18030 编码；数据量小、程序改动量小、影响范围小的系统尽可能升级为 UTF-8 编码；数据量大、程序改动量大、升级成本过高的系统保持 GBK 编码不变，采用编码扩展的转义字符串的方式来表示生僻字。

数据交汇层统一处理，减低影响范围。使用新交换平台、通用文件传输平台作为转码、批量文件转换枢纽，减少整体系统改造的工作量。对于难以改造的部分 GBK、EBCDIC 编码接口，生僻字改用编码扩展的转义方案。



采购支持人口信息生僻字的字库、输入法和云字库及云输入法系统，柜面终端安装 Windows 版本字库和输入法，网上银行、手机银行则采用私有云的云字库及云输入法作为解决方案。

银行网点需采购新打印机设备或升级打印机字库来完成改造。对于存折打印机，尽量采购支持人口信息生僻字、CJK 扩充 A~F 区域生僻字的打印机。

“一字多码”生僻字姓名在身份证联网核查时报告“身份证号码存在，但与姓名不匹配”问题，前期需要柜员手工输入另一个编码再次发起联网核查，后期宜实现自动化处理。对于转账入账或其他环节的姓名比较，同样对“一字多码”生僻字做智能兼容处理，提升客户体验。

## B.2 实例 2——中国银联全渠道系统生僻字改造

### B.2.1 背景

2020 年 5 月前，中国银联全渠道系统应用使用的 Linux 操作系统、DB2 数据库采用 GBK 编码，对外接口支持 GBK 编码、GB18030 编码及 UTF-8 编码，当外部系统使用 GB18030 编码或 UTF-8 编码接口传入生僻字时，会出现生僻字在中国银联全渠道系统内部转码时出现编码信息丢失问题，使得中文校验失败。

### B.2.2 改造方案

为减少关联系统的影响性，中国银联全渠道系统应用在改造时采用如下方式。

- a) 对外接口保持不变，将全渠道内部的关联系统改为使用 GB18030 编码或 UTF-8 编码。
- b) 修改应用程序，将外部字符集统一转换成内部 GB18030 编码处理。
- c) 针对用户姓名校验，修改对中文字符的正则校验表达式，增加 UCS 扩充 A、扩充 B、PUA 等区域的字符范围的校验支持。
- d) 针对数据库存储，为减少数据库变更，保持 DB2 数据库 GBK 编码不变，经测试，使用 C 语言操作 DB2 数据库时，写入非 GBK 的字符也不会有问题。但是使用 JAVA 语言操作数据库时，由于数据库驱动问题，无法支持非 GBK 字符，因此对可能涉及特殊字符的字段先做 Base64 编码后再存储，对应从数据库取出时先做 Base64 解码再进行处理。

### B.2.3 经验总结

外部接口的字符集变更需要谨慎，宜采用多种编码兼容的方式逐步过渡，避免外部系统改造进度不一致引发各种问题。

若使用 DB2 数据库，由于数据库存储机制的关系，GBK 编码可以支持该编码范围之外的字符而不会产生数据丢失，应用系统可以利用该特性减少对数据库字符集编码的修改。

针对中文字符存储，若数据库不支持生僻字，则可以采用 Base64 编码处理后再存储，这种方案改造成本较低，前提是数据库字段需要预留充足空间。

## B.3 实例 3——中国农业银行核心银行系统生僻字改造

### B.3.1 背景

2016 年之前，中国农业银行核心银行系统运行在主流大型主机上，采用 DB2 数据库和 EBCDIC 编码，支持 32,443 个字符。营业网点使用 Windows 系统的 XPE（一种计算机操作系统）版本，该操作系统仅支持 1993 年版本 GB13000，包括 20,902 个汉字。

### B.3.2 改造方案

2016 年中国农业银行核心银行系统启动生僻字改造项目，主要改造内容如下。

- a) 通过采购支持生僻字的字库与输入法，解决生僻字在柜面渠道操作系统的输入显示问题。

- b) 修改前端应用程序，解决信息系统不能显示生僻字的问题，将原有页面设置的统一字体宋体，全部修改为支持生僻字的字体。
- c) 在网点电子化转型过程中，采用图形打印方式，一并解决生僻字的打印问题。
- d) 建立 UCS、GBK 与 EBCDIC 之间的映射关系，解决大型主机 DB2 数据库存储问题。设计 C、JAVA 和 C# 版转码函数，解决联机调用转码问题，通过引入新组件进行转码或解码，完成需要支持生僻字的项目。

### B.3.3 经验总结

主机下移时若采用分布式架构，建议同步规划字符转码存储的兼容性。  
亟需解决客户在柜面渠道开户难的问题，同时各渠道的优化改造也需统筹推进。

### 参 考 文 献

- [1] 《中国人民银行 公安部关于切实做好联网核查公民身份信息有关工作的通知》（银发〔2007〕345号）.2007-09-11
- [2] 《国务院关于公布〈通用规范汉字表〉的通知》（国发〔2013〕23号）.2013-06-05
- [3] 《教育部等十二部门关于贯彻实施〈通用规范汉字表〉的通知》（教语信〔2013〕2号）.2013-10-09
- [4] 《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
-