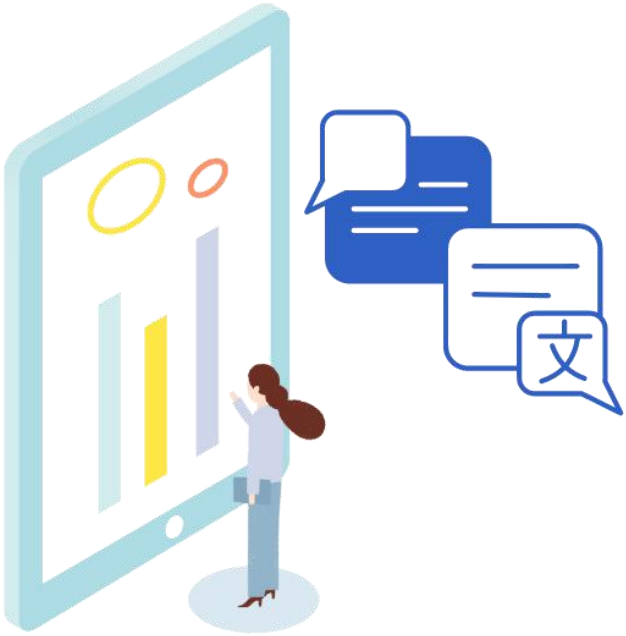
信息技术　生僻字处理指南

（第一版）



全国信标委字符集与编码分委会

二〇二三年八月

编写单位（排名不分先后）

中国电子技术标准化研究院
中信银行股份有限公司
文化艺术出版社有限公司
北京冠群信息技术股份有限公司
万达信息股份有限公司
蚂蚁科技集团股份有限公司
北京搜狗科技发展有限公司
北京汉仪创新科技股份有限公司
北京国家金融科技认证中心
广发银行股份有限公司
电子科技大学
潍坊北大青鸟华光照排有限公司
福昕鲲鹏（北京）信息科技有限公司
江苏中威科技软件系统有限公司
麒麟软件有限公司
北京点聚信息技术有限公司
北京百度网讯科技有限公司
拉萨净土信息安全服务中心有限公司
中国科学院软件研究所
招商银行股份有限公司
中国出版集团中华字库工程办公室
思旦达（北京）技术有限公司
永中软件股份有限公司
深圳市腾讯计算机系统有限公司
北京北大方正电子有限公司
北京金融科技产业联盟
北京银行股份有限公司
重庆国家金融科技认证中心
DICOM 标准国家地方联合工程实验室
中科方德软件有限公司
北京金山办公软件股份有限公司
北京数科网维技术有限责任公司
赞同科技有限公司
友虹（北京）科技有限公司
佳能医疗系统(中国)有限公司
北京卓玛之裕科技有限公司

编写组成员（排名不分先后）

范科峰	董　建	杨　磊	黄姗姗	王　欣	陈　壮
陈亚军	陈晓研	崔晓琳	何正安	吴　健	刘汇丹
江激扬	徐晓剑	刘文涛	孙炎森	刘伟骅	马良有
纪熙东	陈永聪	陶　扬	薛馨枫	刘江涛	谢史颉
陈　克	穆玉静	梁　勇	王长胜	吴庆敏	陆碧波
刘　杰	任　翔	鲁　剑	冯　雷	尹方卉	张建国
陈　恳	张国荣	陈金娣	冷怀晶	郑乃玉	胡达川
李　寻	李　振	李博文	祖宇飞	王子健	周宗明
孙　梦	赵华伟	毕小文	秦　逞	吴　娟	邹见效
蒲立新	徐　昆	范计朋	何明杰	高忠军	殷建民
张　宏	康　凯	王新政	胡　昆	冯建茹	冯倩倩
严　伟	施东海	蔡亚军	梁维添	陈　林	战　茅
杨丽霞	刘　丹	王少康	秦　聪	陆　猛	刘赛赛
陈庆功	李　伟	范敏虎	常敏强	赵兰平	黄　岩
王威凯	屈伟晨	杨秋虹	卓玛普赤	李　巍	刘巧焕
孟　岩	蒋增增	代　威			

版权声明

本文件版权属于全国信息技术标准化技术委员会字符集与编码分委会（SAC/TC28/SC2，简称“全国信标委SC2”）。

使用说明：未经全国信标委SC2书面授权，不得以任何方式复制、抄袭、影印、翻译本文件的任何部分。凡转载或引用本文的观点、数据，请注明“来源：全国信标委字符集与编码分委会”。

目录 CONTENTS

前言.....1
第一章 认识信息系统“生僻字”问题.....2
1.1 信息系统“生僻字”的概念.....2
1.2 造成的影响.....2
1.2.1 生僻字受困群众面临的困难.....2
1.2.2 服务机构的业务困境.....6
第二章 问题形成的原因.....7
2.1 历史原因.....7
2.1.1 中文信息处理技术的发展.....7
2.1.2 相关标准的发展过程.....8
2.1.3 汉字编码的工作机制.....9
2.2 现状成因.....10
2.2.1 产品不支持生僻字的情况.....10
2.2.2 信息系统不支持生僻字处理的情况.....13
2.2.3 管理层面的原因.....13
2.2.4 人员层面的原因.....14
2.3 原因总结.....15
第三章 相关工作基础.....16
3.1 各行业已开展的工作.....16
3.2 相关标准/文件.....18
第四章 编码字符集使用指南.....21
4.1 编码字符集标准的使用.....21
4.2 码位的使用.....21
第五章 信息处理产品支持生僻字指南.....23
5.1 通用要求.....23
5.1.1 编码字符集要求.....23
5.1.2 汉字字型要求.....23
5.1.3 汉字输入要求.....23
5.2 系统软件.....23
5.3 支撑软件.....25
5.4 应用软件.....26
第六章 信息处理系统建设改造指南.....27
6.1 通用原则.....27
6.2 各环节改造要求.....27
6.2.1 输入.....27
6.2.2 显示.....29
6.2.3 打印.....29
6.2.4 信息交换.....30
6.2.5 内部处理.....32
6.2.6 存储.....33
第七章 组织和个人生僻字处理指南.....35

7.1 组织.....35
7.1.1 技术管理机制.....35
7.1.2 服务管理机制.....35
7.2 个人用户.....36
第八章 测试评估.....37
8.1 信息处理产品的标准符合性测试.....37
8.1.1 编码字符集测试.....37
8.1.2 字型测试.....39
8.1.3 输入法测试.....40
8.2 信息处理系统生僻字处理能力评估.....40
8.2.1 定义.....40
8.2.2 标准符合性.....40
8.2.3 易扩展性.....41
8.2.4 经济适用性.....41
8.2.5 兼容性.....41
8.2.6 易用性.....41
8.2.7 统一性.....42
8.3 服务机构业务生僻字处理能力评估.....42
8.3.1 定义.....42
8.3.2 技术管理机制.....42
8.3.3 服务管理机制.....42
第九章 实用工具及资源.....44
9.1 对本章所列资源的说明.....44
9.2 软件产品.....44
9.2.1 超大字符集字库.....44
9.2.2 输入法.....47
9.3 相关工具.....50
9.3.1 字符转码工具.....50
9.3.2 字符转义工具.....50
9.3.3 公共组件.....50
9.4 论文文献.....51
第十章 实施案例.....53
10.1 公共服务机构的系统内改造案例.....53
10.2 大规模、跨系统场景的生僻字问题解决.....58
10.3 用字单位解决内部用字需求的办法.....62
10.4 软件产品支持生僻字的解决方案.....64
附件：提交补充编码汉字的技术要求.....68
一、待编码汉字的范围.....68
二、提交申请的单位.....68
三、接受待编码汉字申请的单位.....68
四、申请资料的技术要求.....68

前 言

国务院办公厅《关于全面加强新时代语言文字工作的意见》提出，到2025年，语言文字规范化、标准化、信息化水平要进一步提高。

国家通用语言文字作为信息交互的载体，在数字经济发展建设中发挥着越来越重要的作用。特别是在政务服务和公共服务行业，随着“互联网+”业务模式日趋成熟，人民群众随时可以通过网络实名认证等方式在线办理各类业务，极大地提升了办事效率和服务体验。然而，在语言文字信息技术高速演进的同时，信息系统生僻字问题的严峻性也在逐渐凸显。由于一些信息系统对于生僻汉字的支持不充分，致使人名、地名中的生僻汉字难以正常输入、显示、打印、传输，相关群众在办事时屡屡受阻，形成了难以逾越的“数字鸿沟”。作为影响产业发展与民生福祉的一大痼疾，信息系统生僻字问题的解决时不我待，势在必行。

为形成适用于各行业领域信息系统的综合、协调、务实、可行的生僻字解决方案，全面实现人名、地名生僻汉字在政务服务和公共服务行业内的互联互通，全国信标委字符集与编码分委会组织相关产学研用单位，共同编写形成《信息技术　生僻字处理指南（第一版）》（以下简称“指南”）。

指南主要提出了以下内容：一是针对信息系统生僻字问题的现状与成因进行系统性总结与分析；二是对信息系统生僻字处理提供全方位指导；三是梳理与信息系统生僻字处理相关的标准文件、工具资源以及各行业实施案例。同时，对于无标准编码的生僻汉字给出了申请编码的技术要求与渠道。

指南力求为各相关方提供有效指导和参考，加快推动信息系统生僻字问题的解决，进一步促进国家通用语言文字信息处理水平的提升，助力国家语言文字事业和信息技术产业高质量发展。

编写组

2023年8月

第一章 认识信息系统“生僻字”问题

1.1 信息系统“生僻字”的概念

从甲骨文到楷书，汉字经历了数千年的演变和传承。悠久的历史也导致了汉字的复杂性，不仅数量巨大，还难读、难写、难认。即便是最专业的汉字研究者与文字工作者，对汉字的认知也存在盲区。

随着计算机技术的发展，汉字插上了信息技术的翅膀，从最初能录入计算机，到现在信息系统的广泛使用。汉字信息处理除了给人们生活带来了便利，也引发了新的问题。在信息时代，汉字需要在信息系统上输入、存储、处理、显示、打印，在不同系统之间流转共享，这一系列过程要由信息技术来保障汉字被正确处理，其中最主要的环节就是汉字编解码，也就是需要保证文字处理周期各个环节遵循同样的汉字编码标准。

在日常生活中，人们一般把不常见的或自己不熟悉的汉字称为生僻字。对信息系统来说，“生僻字”是指需要通过一系列特殊处理才能被信息系统正确地输入、输出、处理、传输的汉字。一个字如果在输入、存储、打印、显示、交换等任意一个环节中处理不当，就会造成业务不能正常办理、信息核验失败等问题，给相关人群的生活与工作带来诸多不便。

1.2 造成的影响

1.2.1 生僻字受困群众面临的困难

在信息时代，信息化系统已经深入到人们生活的方方面面。个人或机构的信息需要在各类信息化系统中录入、存储、流转，比如公安人口信息管理系统、居民身份证制证系统、社保系统、金融系统、工商系统、税务系统以及各种出行及政务服务系统。

涉及生僻字处理的出版、科技、考古、历史等专业领域工作者们常会碰到所需生僻字无法被正确处理等问题。

在政务服务、公共服务等业务场景，生僻字不能正确处理则会导致身份

联网核查失败、户名校验失败、数据库信息乱码等问题，使群众办事遇到阻碍。姓名和地址中含有生僻字的群众，长期忍受着诸多不便。

下面这些实例，是生僻字人群日常困境的真实写照。

➢ 实例 1：要求实名认证的服务无法正常办理

只要需要实名认证的地方，大多无法正常办理。例如手机号码无法实名制登记；使用购票软件和其他需要实名制登记的软件打不出名字或被拦截、报错，均无法实名；辽事通、盛世通、场所码、银行卡等，涉及医疗、保险、交通、通讯和金融领域，都遇到无法办理的情况，寸步难行。如图 1-1 所示。

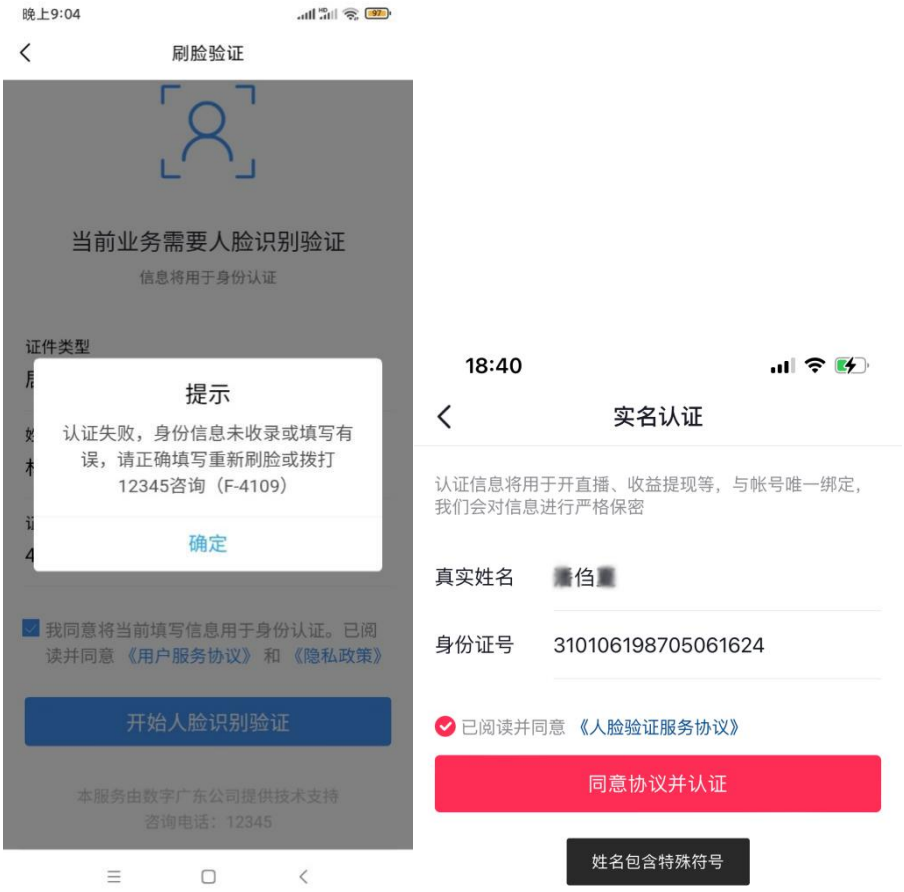


图 1–1　难以通过实名认证（来源：网络）

➢ 实例 2：证件不能正常打印姓名

姓名中有生僻字，准考证上打印出问号，机动车登记证书打印不出姓名，机票/火车票上姓名生僻字只能手写或用拼音替代。如图 1–2 所示。



图 1–2　证照、机票/车票等不能正常打印出姓名生僻字（来源：网络）

➢ 实例 3：无法申请健康码

因名字里有生僻字，八旬老人难办“健康码”。如图 1-3。



图 1–3　新闻截图“八旬老人难办健康码”（来源：湖北电视台）

➢ 实例 4：无法领取退休金

因姓氏是生僻字无法领取退休金，如图 1-4 所示。

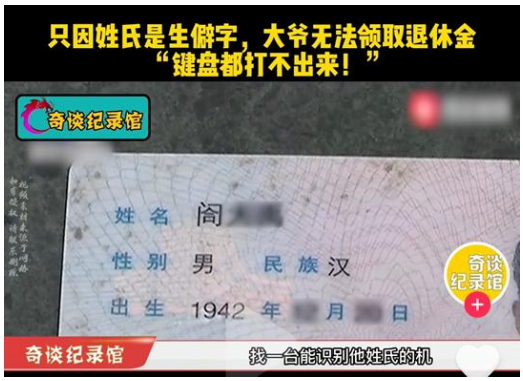


图 1–4　新闻截图“因姓氏是生僻字无法领取退休金”（来源：网络）

➢ 实例 5：无法申请信用卡，税务局无法自动扣税

网友反映：因无法确认征信信息，无法申请信用卡、银行卡；税务局无法线上自动扣税，每次都需要去现场登记办理。网友姓名生僻字举例如图 1-5。



图 1–5　网友姓名中的生僻字（来源：网络）

➢ 实例 6：地名有生僻字无法线上办理出生证

因地名含有生僻字，无法在网上办理孩子的出生医学证明。如图 1-6。



图 1-6　网友住址中的生僻字（来源：网络）

1.2.2 服务机构业务困境

在办理业务过程中碰到生僻字难以处理的情况，群众固然不满意，服务机构也需要投入大量人力和时间来解释缘由、协商处理方式。处理结果也难以让群众满意，引发大量投诉。尽管有些机构设置了私有字库或生僻字处理流程，但是在跨机构、跨系统时仍然会遇到无法校验、显示和核查等问题。

随着社会数字化程度的提高，要求实名制的场景越来越多。基于人名、身份证号的校验、核查愈发普遍，这就要求汉字的信息处理更加规范、标准。

一些机构通过业务积累或搜集整理的方式建立了生僻字库，能满足一部分业务需求。但这样的做法需要这些机构单独维护生僻字库，投入较高的技术成本与人力成本。

还有一些机构通过拼音、图片、临时造字等方式实现当下的业务办理乃至本系统内的信息流转。但在当前万物互联的时代背景下，跨机构、跨系统的应用越来越多，这样的处理方式治标不治本，问题依然没有解决。

第二章 问题形成的原因

2.1 历史原因

全世界的文字，根据其记录语言的方式和性质，可划分为两类：一类是以英文、法文等为代表的表音文字，用字母来表示语音；另一类就是以汉字为代表的表意文字，用一定体系的象征性符号来表示语义。

信息系统生僻字问题，是表意文字所特有的问题。要理解这一问题产生的原因，要从汉字的信息处理技术的原理及其标准化进程说起。

2.1.1 中文信息处理技术的发展

计算机要处理中文信息，其基本原理可以概况为：一字、一形、一码。

字，是指能被单独认知并反映独立意义的一个汉字；

形，是指字形，是文字的精确显示，让机器能正确输出文字信息；

码，是指编码，让机器认识文字，是文字信息处理的基础。

三者之间一一对应，才能被计算机识别和处理。

汉字作为表意文字，每个字都有相对独特的字形和含义，需要有单独的编码，并单独设计字形。

对汉字进行编码的工作包括搜集、查证、去重、认同、赋码等一系列过程。汉字编码工作起步于上个世纪，我国第一个汉字编码字符集标准 GB 2312-80（现为 GB/T 2312-1980）《信息交换用汉字编码字符集　基本集》收录了 6763 个汉字。往后数年，经过各个国家和地区专家的经验积累，被编码的汉字数量急速增长，在 2005 年达到了 7 万多字，覆盖了绝大多数社会在用字。接下来的工作进度则逐渐趋于缓慢，因为尚未被编码的汉字大多是非常罕见的生僻字，搜集难、查证难，有时需要各个国家和地区专家开展大量考证工作，反复讨论审议，才能形成一致结论，为一个汉字分配合理的编码。

汉字字形的校正工作也需要大量专业知识的积累和运用，要对数万个字

形进行考证、校对，确保字形的正确性、规范性，形成符合我国语言文字规范的汉字字形集，避免汉字在文化传播中发生变异。

汉字的数量又非常庞大——2022 年发布的国家标准 GB 18030-2022《信息技术　中文编码字符集》收录 88115 个汉字（含部首），中宣部“中华字库”工程搜集整理出了 30 万个汉字字形成果。为这样庞大数量的汉字分配编码和开发字库，所需的专业知识和工作量都是巨大的。

2.1.2 相关标准的发展过程

标准化是现代社会一切产业的基本保障，是从手工作坊式的分散产业迈向大工厂生产乃至形成大规模产业链的必要手段。从最开始的 GB/T 2312 起步，我国已累计发布中文信息处理相关国家标准 180 余项，涵盖编码、字型、键盘布局、术语等基础性标准，构建了支持市场以及技术开发和应用的中文信息处理基础标准体系，支撑着我国中文信息处理产业的有序健康发展。

解决生僻字问题首先要依靠编码字符集标准的完善，在标准中收录尽可能多的汉字字符，发展标准字库，推动标准在信息系统各个环节的落地。我国研究与制定的编码字符集标准从单字节到双字节、四字节编码，收录的字汇也在不断扩充，相关标准化历程可以用表 2-1 中的几个重要的国家标准/文件来概括：

表 2-1 汉字编码发展历程中的几个重要标准/文件

年份	国家标准/文件名称	汉字数量（含部首）
1980	GB/T 2312-1980《信息交换用汉字编码字符集　基本集》	6763
1995	《汉字扩展内码规范（GBK）》	21003
2000	GB 18030-2000《信息技术　信息交换用汉字编码字符集　基本集的扩充》	27533
2005	GB 18030-2005《信息技术　中文编码字符集》	70244
2022	GB 18030-2022《信息技术　中文编码字符集》	88115

从表中可以看出编码汉字在数量上的两次跃升：一是 1995 年发布的

GBK 文件，在 GB/T 2312 标准 6 千多个汉字的基础上扩充到了 2 万多个；二是 2005 年的 GB 18030-2005 标准，从其 2000 版的 2 万 7 千多个汉字扩充到了 7 万多个。

在汉字字形的标准化方面，我国相继制订发布了数十项汉字点阵字型国家标准和电子行业标准，规定了汉字在点阵栅格中的字形，适用于汉字在点阵屏幕上的显示输出。其中现行有效的国家标准和行业标准有 40 余项，如 GB/T 5007.1-2010《信息技术　汉字编码字符集（基本集）　24 点阵字型》、GB/T 22320-2019《信息技术　中文编码字符集　汉字 15×16 点阵字型》等。

随着信息技术水平的提升，使用点阵屏幕的设备逐渐减少。目前，我们日常使用的计算机和移动终端等设备上搭载的汉字字库都是曲线字型产品。国家标准 GB/T 11460-2009《信息技术　汉字字型要求和检测方法》对曲线字型产品提出了笔形规范性、字形正确性，以及符合编码字符集标准等要求。

2.1.3 汉字编码的工作机制

从汉字编码的工作机制上来说，自 1990 年代至今，我国的汉字编码工作采用的是“国际标准引领，国家标准跟进”的模式。为了确保我国国家标准与国际标准的兼容性，计划新增的汉字首先提交国际标准 ISO/IEC 10646《信息技术　通用编码字符集（UCS)》，待国际标准批准收录后转化为我国国家标准 GB/T 13000。同时，这些新增汉字也将收录进我国自主研发的强制性国家标准 GB 18030《信息技术　中文编码字符集》。

ISO/IEC 10646《信息技术　通用编码字符集（UCS)》是对全球所有文种进行统一编码的国际标准，由 ISO/IEC JTC1/SC2（国际标准化组织和国际电工委员会第一联合技术委员会下属编码字符集分技术委员会）负责修订维护。其中，特别为汉字编码设立了 IRG（表意文字小组），负责审批来自中国、日本、韩国、越南等国家和地区的新增汉字编码申请。我国提交的新增汉字主要来自教育及出版印刷等行业相关单位。在经过全国信标委审核后，新增汉字的编码申请以中国国家提案的形式呈报 IRG，再经过 IRG 多轮严格的技术审查，最终提交 ISO/IEC JTC1/SC2 批准，在国际标准 ISO/IEC 10646 中赋予正式编码。新增汉字编码的审核周期一般为 4—5 年。

注：业界常说的产业规范 Unicode Standard 由美国 Unicode 联盟（Unicode

Consortium）提出，其编码结构、收录字符与国际标准 ISO/IEC 10646（UCS）保持一致，不同之处在于 Unicode Standard 对各文种的编码实现作了更详细的规定。虽然从技术角度看，产业规范 Unicode Standard 与国际标准 ISO/IEC 10646 基本一致，但它并不等同于国际标准。

从汉字编码的工作机制来看，收录生僻字的难点有两方面：一是为避免产生一字多码的问题，标准化组织对新增汉字的审查非常严格，有些字会被认为来源证据不足，或跟已收录的某个字是认同（unify）的，所以不能赋予单独编码；二是有些用字单位不会主动向标准化组织提交待编码的生僻字，标准化组织无从获取这些生僻字的信息。

注：一字多码问题，是指在标准化过程中造成的下列情况：

1）同一标准中同一图形字符出现在不同编码位置；

2）存在编码映射关系的标准之间，因标准修订不同步，导致原映射关系变更至新码位；

3）某些机构早期通过占用 PUA 对一批急用图形字符分配了自定义编码，导致这些字符与现行标准编码不一致。

2.2 现状成因

2.2.1 产品不支持生僻字的情况

软硬件产品在进行中文信息处理和交换时不支持生僻字的情况，可能涉及输入法或字库不支持、软件底层编码不支持、软件前端/后端对代码的过滤，以及网页渲染等原因。下面对这些情况逐一进行分析。

- **输入法**

对于一般用户，输入法是最直接接触的环节，一个对用户友好的输入法包括输入码对应字符或字符序列的范围广度、候选框上屏显示效果、易混淆字形提示和词库联想提示等。

GB 18030 和 GB/T 13000 本身并不直接提供所有字符的读音信息，若想使用汉语拼音做输入码基础来输入所有的字符并不现实。字形特征输入码（形码）也需要开发者做大量的研究工作，且目前任何一种形码都缺乏可以协调规则的技术委员会，如果各个输入法开发方都独力开发，最终输入码的结果可能不一致。有的字符本身不一样，但在小字号时，容易渲染得让用户

分不清。比如“田”的 GB/T 13000 码位为 U+2BB42，GB 18030 码位为 0x99308834，与“田”字（GB/T 13000 码位 U+7530，GB 18030 码位为 0xCCEF）的字形很像，但两者编码不同，是两个不同的字。在小字号渲染时，如果输入法没有提示，用户可能会输入错误。

➢ 字库/字体库/字体文件

基于 OFF（开放式字体格式，参见 ISO/IEC 14496-22）封装的曲线字库封装时都会按照一定的历史字符集、字符子集或字形集（glyph collection）来安排字形（glyph），目前市场上大部分字库产品包含的汉字数量在几千到三万之间不等，不足以满足绝大部分生僻字处理需求。此外由于字库文件有容纳字数的限制，GB 18030-2022 收录的所有字符无法容纳在单个字库文件里，一旦系统或平台的渲染规则没处理好，超出当前选用字库包含字形的字符会回退（fallback）到系统默认字库。如果系统默认字库的字体风格（style）与当前选用字库不同，在排版上会显得突兀；如果系统默认字库不包含该字符的字形，会调用字库中的.notdef（一个替代缺字用的字形）来显示。另外，点阵字库在当代仍有广泛的使用环境。医院药房、地铁、公交等的点阵显示屏仍需要相应的点阵字体去支持显示。

➢ 编码

无论是从 GB/T 2312-1980 的通行编码实现到 GBK 再到 GB 18030，还是对 GB/T 13000（UCS）的编码实现，在发展过程中都发生过编码结构上的变化。目前国内仍有大量以前开发的软件或依据旧框架开发的新软件无法处理这些编码结构上的变化，这导致从根本上限制了软件能处理的字符数。比如，“颐”的 UCS 码位为 U+2CC56，GB 18030 码位为 0x9933C336，如果一个只支持 GBK 的环境接收到一个 GB 18030 编码的文档或信息系统报文，这个字会被处理成“&#65533;3&#65533;6”；如果一个只支持 UCS-2（只包含 UCS 基本多文种平面的一种编码）的环境，接收到一个 UTF-16（覆盖 UCS 中所有平面的一种编码）的文档，这个字会被忽略掉，这些情况都属于没能正确处理编码。

➢ 前端与后端

如果在操作系统与平台已经支持编码的前提下，应用系统选用了范围更窄或范围不对的编码（习惯称为 charset label），也会导致和前述相似的结果。比如，在 Oracle 数据库中选用了 ZHS16GBK，而不是 AL32UTF8；在 MySQL 数据库中选用了 ucs2 或 utf8（utf8mb3），而不是 utf8mb4 或 utf16、utf16le 等；在 html 或 XML 中选用了 GBK 或 gb2312，而不是 gb18030 或 utf-8 等。另一方面，国内很多登录界面习惯通过正则表达式限制输入的字符，常见的写法是[\u4e00-\u9fa5]，这等于把超出 GB/T 13000-1993 或 Unicode 1.0.1 定义的汉字以外的部分及人名中需要的其他符号（如汉字数字“〇”和少数民族姓名分隔符“·”）排除在通过的范围内。

➢ 网页渲染

一般汉字的视觉断行规则是任意两个汉字之间都可以断开，而拉丁字母等不能在任意两个字母之间断开，除非添加连字符（hyphen）。由于使用非 UCS 正式码的生僻字（即定义在 PUA 和两个 SPUA 区）未定义断行属性，多数平台会直接使用等价于拉丁字母的断行行为。这样的渲染效果可能导致页面显示和打印时发生信息丢失。图 2-1 是一个网页对非标准码汉字进行了错误断行处理的效果：它把第一行空格后面的 5 个汉字认定为一个“单词”，在字符总长超过行长时，将这个“单词”整体放到了下一行。

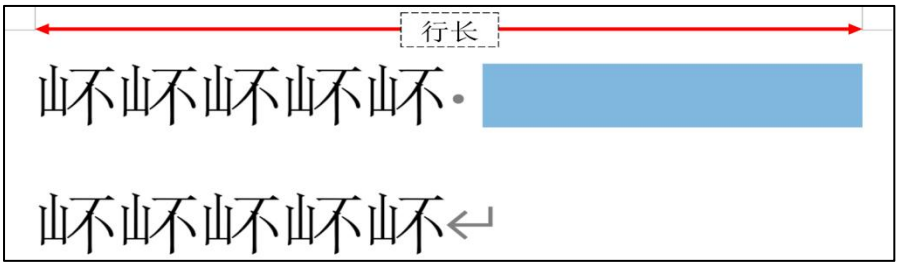


图 2-1　错误的断行效果

正确的断行处理应当如图 2-2 所示。

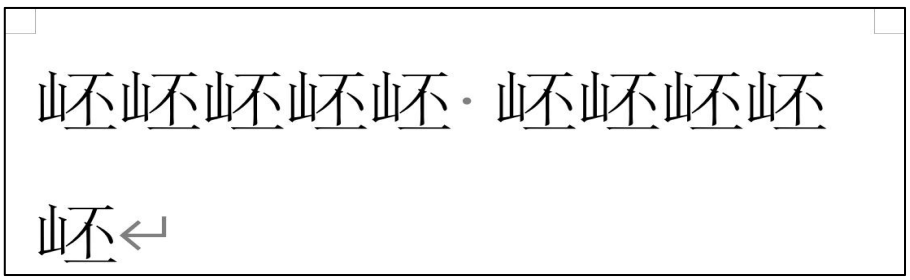


图 2-2　正确的断行效果

2.2.2 信息系统不支持生僻字处理的情况

信息系统不能正确处理生僻字的原因则更加复杂。除了上一节所述软硬件产品本身对生僻字不支持的情况外，还要考虑系统运行的各个环节，包括输入、存储和输出等。

- 输入

信息系统中生僻字相关有效信息的输入既和软件有关，也和信息系统的用户有关。即便软件已开发完善，由于汉字本身的特性，用户从 GB/T 13000 和 GB 18030 标准收录的近十万字中查找到需要的字也并不容易。例如有些字的同音字有几百个。此外，一些输入操作员在遇到生僻字不能输入时会使用替代办法，但不同人员、不同时间使用的替代办法没有固定规则，例如拆字、结构描述、替代字、拼音、替代符号、占用汉字码位修改字形等都有可能。

- 存储

在很多信息系统的编码与转码规则中，当前界面的编码不能支持的字符可能会被一个或多个“□”或半角问号“?”等字符替代。生僻字在被不可逆替代后存储会丢失有效信息。另一方面，输入层面信息的不准确也会给存储带来不便。拆字、结构描述、拼音等方式都可能使字节数超出字段规定的长度，替代字、替代符号、占用汉字码位修改字形等可能会造成信息被误解。

- 输出

当包含生僻字的信息在输出时，会受到软件层面编码、字库、渲染、发音软件等输出能力不足的限制，导致即使在机器意义上信息准确，但显示或打印输出为空白、问号等现象，与一般用户与从业人员的认知严重脱节，影响群众办事进度。当点阵屏和热敏打印机等内置的点阵字库及其编码不支持生僻字时，可能打印出替代用的问号（?），有的还会发生字节丢失或错位解码等情况，使一般用户与从业人员无法从字面识读必要信息。

2.2.3 管理层面的原因

从服务机构的管理层面来看，信息系统生僻字问题普遍存在，其主要原

因还是服务机构各级相关方对生僻字问题的认知不清晰、对技术成因的不理解。比较常见的问题包括：

- **误用 GBK**

GBK 仅收录了 21003 个汉字，绝大部分生僻字都不支持，且其中部分字符编码与现行国际/国家标准中的规定不一致。一些信息系统开发方不理解各个编码字符集标准/文件（包括 GB/T 2312、GBK、GB 18030-2000、GB 18030-2005、GB 18030-2022、ISO/IEC 10646 和 Unicode）之间的差异和要求，误以为 GBK 仍是当下适用的汉字编码标准，在编写行业规范等文档时，规定使用 GBK 编码，使得大多数生僻字都无法被正确处理。

- **错误使用 PUA 区**

PUA（用户自定义区）缺乏有效的管理手段，导致各行业采用各自内部自定义 PUA 编码作为标准，导致无法跨行业互联交换。例如某信息处理产品使用 PUA 区对生僻字进行编码，用户使用该产品可以处理特定生僻字，但由于其编码不是标准码，在和其他系统进行信息交换时，则会引起错误。

- **系统建设未遵循合理标准**

各类服务机构在系统建设方面，无论是自建还是外购承建，都存在前述的类似问题。而一些行业的复杂系统通常具有数据分布式管理、分系统承建单位不同等特点，在系统内数据交换的各个环节，任一环节不能支持生僻字处理，均会导致问题产生。

2.2.4 人员层面的原因

直接面对生僻字受困群众的服务机构客服人员，普遍存在不了解问题、不会操作输入生僻字以及有概念误区等情况。生僻字受困群众自己，大多对问题的成因一知半解，也没有有效的方法来维护自己的合法权益。

- **一线柜台或后台客服人员**

各个机构缺乏专业知识人员，通常难以建立完整的知识体系并进行有效

的培训。在这方面做的比较好的案例是《广东社保规程》，《规程》对生僻字的范围进行了界定，对服务人员处理生僻字提出了明确的操作指导。

- **生僻字受困群众**

生僻字受困群众往往也没有充足的认知，缺乏有效的方法和手段来维护自身权益。其中有些人会由于长期遭遇各种不便而被迫改姓（如“恵”→“惠”、“茼”→“茼”）、改名（“龚”→“龔”或“衍”、“頔”→“頔”或“迪”、“蘋”→“平”）。

2.3 原因总结

由上述分析可知，信息系统生僻字问题长期存在而难以彻底解决，存在三方面原因：

一是汉字的信息化难度高。汉字数量庞大，对每个汉字进行编码和字形规范化，工作量大、专业度高、标准化工作周期长。一个生僻字从被发现，到提交到标准化组织，再到被赋予编码，进入标准，要经历长期的过程；

二是软硬件产品和业务系统对标准的支持程度良莠不齐。一个业务系统包含各类软硬件产品以及输入、存储、处理、输出和控制等诸多环节，其中任何一个环节不支持生僻字处理，都可能导致业务办理不成功；

三是大众对生僻字问题的认知不足。服务机构的管理人员和客服人员缺乏相关知识，采取的处理办法通常治标不治本，而生僻字人群本身也大多不了解问题根源，没有有效的方法和手段维护自身权益。

第三章 相关工作基础

3.1 各行业已开展的工作

信息系统生僻字问题给相关群众的生活带来极大不便，多年来相继有两会代表委员提出提案，要求解决信息系统生僻字问题；不同行业主管部门也早已注意到这一问题，尝试解决用字问题。例如：

➢ 国家民委及相关部委联合发文

2016 年，国家民委联合多个有关部委发布《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），提出了在信息系统设计研发工作中要实现对国家标准编码汉字 GB 18030 或 GB 13000 的全覆盖等要求。

➢ 公安人口信息专用字库

公安部建立了公安人口信息专用字库，有效解决了绝大多数姓名中含有生僻字的公民正常办理户口登记和居民身份证的问题，但其他部门的信息系统无法正常处理且在跨系统时无法联动。

➢ 社保、医保、社保卡

社保、医保、社保卡方面的三个管理部门以及医院本世纪初统一使用一个造字软件支持生僻字的录入和显示，确保社保信息流闭环内生僻字处理规则的一致性，但这个软件已不再更新。2020 年起，人社部要求人力资源社会保障信息系统全部支持国家标准 GB 18030。在采集个人基础信息时，如果其姓名中包含生僻字，则信息系统中应采用该字的半角大写拼音字母加上半角中括号代替；在备注字段中对拼音标注、字形、读音等进行描述。例如“王乆颋”，在信息系统的姓名字段中采用“王[JIU][DI]”，在备注字段中说明“生僻字[JIU]字形上九下乙，音同九；生僻字[DI]字形左由右页，音同迪”。要求

下属机构对生僻字的甄别严谨细致，不得将非生僻字的汉字擅自按照生僻字的方案处理。

➢ 电信

在电信领域实名登记方面，工业和信息化部为解决身份信息核验中生僻字的问题，一方面指导电信企业加快信息系统升级改造，另一方面采取相似字替换、添加备注等方式为用户提供便利。

➢ 铁路客运

在铁路售票系统中，铁路部门积极研究解决方案，在确认旅客身份证号码无误的前提下，对生僻字进行模糊化对比处理，并进一步完善售票渠道功能，使姓名中包含生僻字的旅客通过互联网购票后，在人工窗口、自动售票机上均可办理取票业务。

➢ 金融

2022 年 6 月，中国人民银行发布了《金融服务　生僻字处理指南》（JR/T 0253-2022）金融业标准，标准提供了行业通用的处理生僻字的方法，为金融业信息系统生僻字处理提供了指引，以改善生僻字受困群众在办理金融业务时遇到的各种障碍，助力提升金融业服务水平。同时，要求遵循新版《信息技术　中文编码字符集》（GB 18030-2022）强制性国家标准。2022 年 9 月 29 日，由国家金融标准化技术委员会秘书处指导北京金融科技产业联盟组织召开了金融业生僻字治理工作推进会，进一步落实相关要求。

为填平生僻字这一历史和技术原因造成的“数字鸿沟”，解决人民群众“急难愁盼”问题，一些地方政府也进行了积极探索。例如：

➢ 上海市

2021 年 9 月上海市经济和信息化委员会、上海市大数据中心联合发出了《关于开展信息系统生僻字问题改造的通知》，针对群众日常生活中常见服务事项涉及的信息系统进行生僻字支持改造，通过集中采购公安系统人口信息专用字库并依托“一网通办”技术体系建立生僻字云服务平台，支撑相关

政府部门以及医院、药店等医疗机构，试点银行、保险等金融机构，水电煤等公共服务企业的信息系统对生僻字进行技术处理。改造后的各部门信息系统对生僻字的处理对标公安的身份证处理方法，在成本可接受的范围内从根本上解决姓名中含有生僻字的群众便捷办理公共服务事项的问题。

3.2 相关标准/文件

面向信息系统生僻字治理工作，可将支撑各类产品研发和信息系统建设的标准/文件划分为四大类，分别是 A 基础技术、B 产品要求、C 应用要求和 D 测试评估。其中：

A 基础技术：包括编码字符集、字型、键盘布局等支撑生僻字处理技术的基础标准；

B 产品要求：围绕生僻字处理需求，分别面向软、硬件产品提出要求；

C 应用要求：面向行业应用的标准、文件等；

D 测试评估：相关测试方法、评估指南等。

该支撑体系框架如图 3-1 所示。

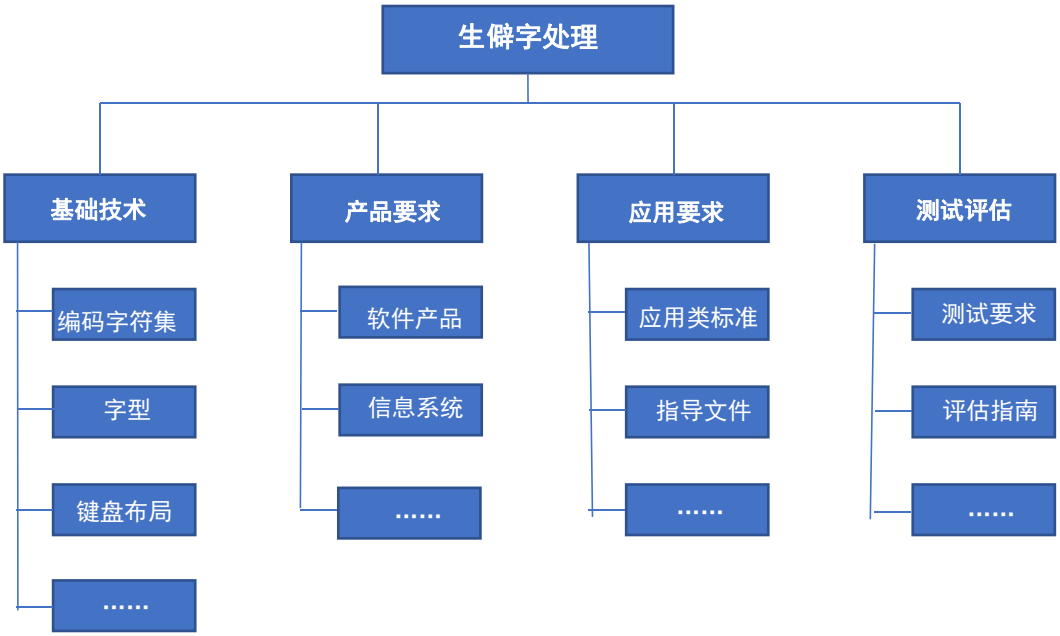


图 3–1　生僻字处理标准/文件体系框架

在该体系下列出已发布的标准/文件清单如表 3–1 所示。

表 3-1　生僻字处理相关标准/文件清单

大类	小类	标准/文件名称	标准号/文件号	文件性质
A 基础技术类	编码字符集	信息技术　中文编码字符集	GB 18030-2022	强制性国家标准
		信息技术　通用多八位编码字符集（UCS）	GB/T 13000	等同采用国际标准 ISO/IEC 10646
		信息技术　通用编码字符集（UCS）	ISO/IEC 10646	国际标准
	字型	汉字点阵字型标准	多项（略）	推荐性国家标准
	键盘布局	汉字及少数民族文字键盘布局标准	多项（略）	推荐性国家标准
B 产品要求	通用要求	信息技术产品国家语言文字使用管理规定	教育部第 54 号令	
		信息技术　软件产品中文信息处理要求及测试方法	拟立项	推荐性国家标准
	软件产品	信息技术　数字键盘汉字输入通用要求	GB/T 18031	推荐性国家标准
		信息技术　通用键盘汉字输入通用要求	GB/T 19246	推荐性国家标准
		联机手写汉字识别系统技术要求与测试规程	GB/T 18790	推荐性国家标准
		中文语音识别系统通用技术规范	GB/T 21023	推荐性国家标准
		印刷体汉字识别系统要求与测试方法	GB/T 17961	推荐性国家标准
		信息技术　汉字字型要求和检测方法	GB/T 11460	推荐性国家标准
		信息技术　生僻字处理要求　第 1 部分：软件产品	拟立项	推荐性国家标准

	信息系统	信息技术　生僻字处理要求　第 2 部分：业务系统	拟立项	推荐性国家标准
	……			
C 应用要求	通用	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	
	金融行业	金融服务　生僻字处理指南	JR/T 0253-2022	
	人力资源和社会保障	人力资源社会保障信息系统姓名生僻字处理方案(暂行)	人社网信函〔2019〕17 号	
	……			
D 测试评估	测试要求	信息技术　软件产品中文信息处理要求及测试方法	拟立项	推荐性国家标准
		信息技术　汉字字型要求和检测方法	GB/T 11460	推荐性国家标准
	评估指南	信息技术　生僻字处理要求　第 3 部分：服务机构	拟立项	推荐性国家标准
	……			

第四章 编码字符集使用指南

4.1 编码字符集标准的使用

我国现行有效的汉字相关编码字符集国家标准主要包括 GB/T 2312、GB/T 13000 和 GB 18030 三项。

GB/T 13000《信息技术　通用多八位编码字符集（UCS）》等同采用国际标准 ISO/IEC 10646。业界常说的 Unicode 规范与国际标准 ISO/IEC 10646 在编码和字汇层面保持一致，当前版本收录全球各文种字符约 14 万。国际标准 ISO/IEC 10646(GB/T 13000)定义的交换码包括 UTF-8、UTF-16、UTF-32 等，当前编码空间共启用 17 个平面，共计 111 万多个码位，归类为大字符集编码，可以处理生僻字。

GB 18030《信息技术　中文编码字符集》是强制性国家标准，是我国自主研制的编码字符集标准，其编码空间中共有 161 万多个码位。在 GB/T 13000 的码位空间内，GB 18030 的码位与 GB/T 13000 的码位一一映射，可以认为二者的表达能力是等价的。因此，GB18030 编码也归类为大字符集编码，能够处理生僻字。

GB/T 2312《信息交换用汉字编码字符集　基本集》仅收录了 6763 个汉字，无法处理生僻字，归类为小字符集编码。

《汉字扩展内码规范（GBK）》不是标准，而是一个技术规范指导性文件。该文件已于 2000 年废止。GBK 编码收录 21003 个汉字，无法处理生僻字,应立即停止使用,并尽快升级为 GB 18030 编码或 UTF-8、UTF-16、UTF-32 等编码。

4.2 码位的使用

对于 GB 18030 中已经收录的汉字，应当使用 GB 18030 中的码位（或 GB/T 13000 中的对应码位），对于 GB 18030 中未收录，但 GB/T 13000（或 ISO/IEC 10646）中已收录的字，应当使用 GB/T 13000（或 ISO/IEC 10646）

中规定的码位。

ISO/IEC 10646 中预留的码位（GB 18030 中已经占用的部分除外）是为未来增补字符使用，一般机构和个人不应占用。

ISO/IEC 10646 和 GB 18030 中均定义了用户自定义区（私用区，PUA），按照标准规定，用户自定义区可以由最终用户使用。由于 PUA 使用的自由性导致互不兼容，PUA 码位的使用应严格遵循最小范围原则。任何需要与外部做信息交换的场景都不应使用 PUA 码位。对于 GB 18030 或 ISO/IEC 10646 中已经收录的 PUA 字，应尽快升级使用其对应的正式码位。对于暂时未被收录的 PUA 字，应根据本文档附件的要求向全国信息技术标准化技术委员会反馈，以期尽快收录至国家标准，详见附件：《提交补充编码汉字的技术要求》。

当前 BMP 还有兼容区，除已被收录到 GB 18030 中的 12 个兼容区汉字以外，其他的兼容区汉字均不应使用。

康熙部首区（U+2F00～U+2FFF）和扩展部首区（U+2E80～U+2EFF）的汉字或部件与一般汉字同形时，原则上应当使用一般汉字。

第五章 信息处理产品支持生僻字指南

5.1 通用要求

5.1.1 编码字符集要求

由于 GB 18030 是强制性国家标准，也是我国现行有效的国家标准中收录中文字符最多的编码字符集标准。它与 ISO/IEC 10646 协同更新，其最新版本（包含修改单）收录了 ISO/IEC 10646 最新版本中的全部汉字，以及公安人口信息字库的汉字。因此，面向生僻字处理的软件产品应支持 GB 18030 的实现级别 3。

5.1.2 汉字字型要求

——字汇范围

汉字的字汇范围应达到 GB 18030 实现级别 3 的要求。

——字形规范

汉字字形应符合国家语言文字政策法规，并符合 GB/T 11460《信息技术 汉字字型要求和检测方法》的要求。

5.1.3 汉字输入要求

处理生僻字的汉字输入法产品，应能输入 GB 18030 实现级别 3 规定的所有字符，并符合相应的国家标准：

a）数字键盘输入应符合 GB/T 18031 的规定；

b）通用键盘输入应符合 GB/T 19246 的规定；

c）手写输入应符合 GB/T 18790 的规定；

d）语音输入应符合 GB/T 21023 的规定。

5.2 系统软件

除通用要求外，建议系统软件类产品满足如下要求。

a）操作系统：

- 至少提供一种可以编辑/浏览并打印按照 GB 18030 实现级别 3 的编码方式进行编码的文件的应用程序；
- 提供支持 GB 18030 实现级别 3 中的字符打印的打印机驱动程序；
- 至少提供一种符合 GB 18030 实现级别 3 部分相关字型标准的字库，可以是点阵字型、曲线字型或压缩字型，并可通过操作系统提供的字体解释引擎正常显示该字库中的图形字符；
- 至少提供一种可以输入 GB 18030 实现级别 3 中所有字符的输入方法。

b）数据库管理系统：

- 应能正确执行包含 GB 18030 实现级别 3 的编码字符的 SQL 脚本；
- 产品在用户界面使用自带的专用字库时，该字库应符合 GB 18030 实现级别 3 部分相关字型标准的要求；
- 产品在用户界面使用第三方字库时，应向用户提供明确的字库选用方法。当用户选用了符合 GB 18030 实现级别 3 部分相关字型标准要求的字库时，产品应能在所有功能界面正确显示字库中的图形字符；
- 产品在用户交互中使用自带的专用输入法时，该输入法应能正确输入 GB 18030 实现级别 3 内的所有字符；
- 产品在用户交互中允许使用第三方输入法时，当用户选用了能正确输入 GB 18030 实现级别 3 内的所有字符的输入法，产品应能在所有交互界面正确完成所有字符的输入。

c）驱动程序：

- 面向具有图形字符打印、显示、存储等输出功能的设备的驱动程序，如打印机驱动程序、显示器驱动程序等，不应影响 GB 18030 实现级别 3 中规定的编码字符的正确输出；
- 面向具有编码字符传输、交换等功能的设备的驱动程序，如调制解调器、证件阅读器等，不应影响 GB 18030 实现级别 3 中规定的编码字符的正确传输、交换。

5.3 支撑软件

除通用要求外，建议支撑软件类产品分别满足如下要求。

a）开发支撑软件：

- 带有编译功能的开发支撑软件类产品，应能正确完成对含有 GB 18030 实现级别 3 编码字符数据的程序文件的编译。GB 18030 编码字符数据可以是变量的值、函数中的参数、程序中的注解等；
- 产品在用户界面使用自带的专用字库时，该字库应符合 GB 18030 实现级别 3 部分相关字型标准的要求。
- 产品在用户界面使用第三方字库时，应向用户提供明确的字库选用方法。当用户选用了符合 GB 18030 实现级别 3 部分相关字型标准要求的字库时，产品应能在所有功能界面正确显示字库中的图形字符；
- 产品在用户交互时使用自带的专用输入法时，该输入法应能正确输入 GB 18030 实现级别 3 内的所有字符；
- 产品在用户交互时允许使用第三方输入法时，当用户选用了能正确输入 GB 18030 实现级别 3 内的所有字符的输入法，产品应能在所有交互界面正确完成所有字符的输入。

b）中间件：

- 应能正确处理按照 GB 18030 实现级别 3 进行编码的输入数据，并产生正确的输出数据。

c）浏览器、搜索引擎、虚拟化软件、大数据处理软件和人工智能软件等：

- 产品在用户界面使用自带的专用字库时，该字库应符合 GB 18030 实现级别 3 部分相关字型标准的要求；
- 产品在用户界面使用第三方字库时，应向用户提供明确的字库选用方法。当用户选用了符合 GB 18030 实现级别 3 部分相关字型标准要求的字库时，产品应能在所有功能界面正确显示字库中的图形字符；
- 产品在用户交互时使用自带的专用输入法时，该输入法应能正确输入 GB 18030 实现级别 3 内的所有字符；
- 产品在用户交互时允许使用第三方输入法时，当用户选用了能正确输入

GB 18030 实现级别 3 内的所有字符的输入法，产品应能在所有交互界面正确完成所有字符的输入。

5.4 应用软件

涉及生僻字信息处理的行业应用软件，应提供符合 GB 18030 实现级别 3 要求的字库。

此外，还应提供符合 GB 18030 要求的输入方式。

第六章 信息处理系统建设改造指南

6.1 通用原则

涉及居民个人信息处理的服务机构在信息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。

- **遵循标准**

支持国家标准 GB 18030，并鼓励兼容 GB/T 13000 的相应编码，字符集以二者的最新版本为准，使信息链路的各环节都能支持大字符集处理，避免出现小字符集的瓶颈。

- **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用户体验。

- **经济适用**

以满足用户实际需要为基础，配置实用的字库、输入法、接口设备、输出设备等。

- **兼容处理**

尽可能兼容处理涉及民生的各类信息系统中的生僻字问题，提升用户体验。

- **包容普惠**

充分考虑农村与偏远地区居民、老年人、残疾人、少数民族等群体的需求，如手机 APP 等针对老年人群体提供大字号的版本等。

- **接口统一**

非大字符集系统对超出支持范围的字符转义表示宜采用统一方案。

6.2 各环节改造要求

6.2.1 输入

6.2.1.1 应配备的输入法/输入设备

业务系统应配备符合第四章要求的输入法/输入设备。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。在市场上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（例如，从“姓名生僻字平台”上拷贝），以保证至少有一种方法可将生僻字录入到系统中；

b）手机 APP、Pad APP、智慧柜台等触摸屏应用因无物理键盘，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。

6.2.1.2 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）针对客户临柜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，提供经主管授权许可的补充手段；

b）针对老年人等群体建议提供大字号的输入界面、手写识别输入、语言识别等多种便民方式；

c）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；

d)对于手机 APP 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；

e）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；

f）部分少数民族同胞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,对应 UCS 编码 U+00B7)。考虑到常用字符集中实心“点”字符有多个（如“ˑ”U+02D9、“·”U+0387、“ ”U+1427、“ ”U+16EB、“•”U+2022、“·”U+2027、“·”U+2219、“ ”U+22C5、“ ”U+2E31、“▪”U+30FB、“ ”U+A78F、“．”U+FF0E、“･”U+FF65），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“点”时，自动转换成 U+00B7。

6.2.2 显示

6.2.2.1 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）能显示 GB 18030-2022 规定的全部汉字；

b）鼓励参考 ISO/IEC 10646 最新版本覆盖新增汉字（目前是 CJK 扩充 G、H 及其他 CJK 统一汉字区块尾部新增汉字等）。

6.2.2.2 特殊情况的处理

生僻字信息在显示时可能遇到的一些特殊情况及其处理方式列举如下：

a）由于单个字型文件字形数量的限制，宜通过操作系统的字体回退机制或者应用软件自行实现字体回退机制实现生僻字的显示；

b）由于一字多码和相似字形的客观存在，建议在姓名、地址等可能涉及生僻字的栏位旁边回显字符编码，以便在遇到疑难情形时可迅速判断问题成因；

c）在必须显示 PUA 编码汉字的情况下：

- 宜对 PUA 编码汉字字形与正式编码字形作出明显区分；
- 对于身份证姓名的 PUA 编码生僻字，由于用户可能使用包含不兼容人口信息字库 PUA 编码生僻字的输入法输入，宜采用和公安人口信息字库兼容的字库予以显示，供用户确认，避免输入不兼容的 PUA 编码生僻字引发后续投诉。

d）针对老年人等群体建议提供大字号的显示界面版本。

6.2.3 打印

6.2.3.1 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 6-1。

表 6-1 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。	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 GB 18030 实现级别 3 的汉字。	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。

6.2.3.2 关于打印机字库的建议

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，建议服务机构：

a）请生产厂商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，也可通过图形打印的方式确保生僻字被正确打印；

d）在需要使用程序来生成 OFD、PDF 等文件时，宜使用支持生僻字的字库，避免生僻字打印结果与客户信息不一致。

6.2.4 信息交换

6.2.4.1 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）原则上应支持 GB 18030 的汉字无损透传处理，同时兼容 ISO/IEC 10646（一般用 UTF-8 编码）；

b）原使用 GBK 编码的报文及文件交换宜升级为 GB 18030 编码，同时

兼容 ISO/IEC 10646（一般用 UTF-8 编码）；

c）转接系统在转接时，因输入、输出双方编码不同，需要做编码转换时，不应发生：

- 丢弃某些字符或转成替代符“?”、“□”等的有损转换；
- 报文丢弃或报错的情况。

6.2.4.2 特殊情况的处理

生僻字信息在交换时可能遇到的一些特殊情况及其处理方式列举如下：

a）如果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；

b）如需要交换的信息包含 PUA 编码汉字，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）当柜面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，在备注字段可使用拆字等方式描述生僻字。

d）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：

- 使业务系统支持一字多码互相认同的智能比较；
- 对于当前系统未改造尚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。

6.2.4.3 应注意的要点

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，因生僻字 GB 18030 的 4 字节编码或转义格式可能引发字符数统计问题影响对齐截取。另外生僻字转义格式可能导致超长；

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：如“弢”字的 GBK/GB 18030 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。建议采用不可见字符如 0x03 字符作为分隔符以避免冲突；

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，以免 XML 解析器解码错误（反例：头部的 encoding 编码为 GBK，内容采用的编码为 GB 18030。）；

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 GB 18030 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，如生僻字 U+20164 “𠘤”（𠘤亩心）在 JSON 中被表示为可读字符串"\uD840\uDD64"，应检测所用报文解析器是否能正确解析处理；

e）注意对于 UTF-8、UTF-16、UTF-32 编码以及 GB18030 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，文件使用时往往会报错；

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

6.2.5 内部处理

关于生僻字信息的内部处理，部分要点可参见 5.2.4 节。需要另外注意的是：

a）消除常见误区，汉字不全是 2 字节，非生僻字的 GBK 编码、UCS2 编码才是 2 字节，非生僻字和 CJK 扩充 A 区的生僻字这两部分的 UTF-8 编码是 3 字节，其它大部分生僻字是 4 字节（GB 18030、UTF-8、UTF-16 编码）。因此，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。因此，往往需要另行开发支持生僻字的函数/方法；

b）为支持生僻字，所用编程语言的编译宜采用 Unicode 等模式，避免采用原先的 MBCS 等模式；

c）考虑到姓名生僻字“一字多码”问题的长期存在，对于如开户客户或账户户名与收款人姓名等“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。

6.2.6 存储

6.2.6.1 一般要求

业务系统在汉字信息存储方面的一般要求包括：

a）数据库存储和查询应支持 GB 18030，同时兼容 UTF-8 编码；

b）文件存储宜采用 GB 18030，同时兼容 UTF-8 编码；

c）在不能改变存量数据库字符集设置的情况下（如成本过高），如果存量数据库采用了 GBK 编码，那么对超出 GBK 范围的生僻字，宜在应用系统层面用转义格式编码后，再写入数据库。此时：

- 从数据库读出数据时，宜将转义格式还原成汉字；
- 转义格式宜基于易于还原、占用空间小的 UCS 编码；
- 转义格式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。

6.2.6.2 其他注意要点

在存储方面还应注意的要点有：

a）字段设计应遵循《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，在信息系统设置中规定姓名数据项最大长度不少于 50 个字符（25 个汉字）[1]；

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；

c）部分少数民族同胞姓名的间隔符宜按 5.2.1.2 的规范形式存储，不规范的旧数据宜定期迁移。

6.2.6.3 常见数据库产品的处理要求

对常见的几种数据库产品类型，分别列举其针对生僻字信息存储的要求（下文所列 MySQL、DB2、Oracle 均指数据库产品名称）：

➢ 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

[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）中提到的“字符”实应为“字节”，信息系统设置中规定姓名数据项最大长度应不少于 50 个字节。

UTF-8 的字符。

➢ DB2 数据库

在大型主机系统中，CJK 扩充 B 区及以上扩充区、其他 CJK 统一汉字区块尾部新增汉字宜用转义格式处理；

开放平台 DB2 数据库宜升级支持 GB 18030 或 UTF-8 编码。

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

➢ 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。

➢ 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。

第七章 组织和个人生僻字处理指南

7.1 组织

建议涉及居民个人信息处理的服务机构分别从技术管理角度和服务管理角度建立面向生僻字处理需求的相关机制。

7.1.1 技术管理机制

面向生僻字处理的技术管理机制包括但不限于如下两类。

7.1.1.1 字符集管理

服务机构宜建立管理机制，管理本单位所开发、运营的信息系统及接口服务的字符集情况。

7.1.1.2 生僻字专业技术团队

服务机构宜设置生僻字专职人员或引入相关服务，负责设计和实施信息系统生僻字改造，在信息系统运营维护中提供专业技术支持，并负责开发生僻字处理技术操作手册供系统操作人员使用。

7.1.2 服务管理机制

面向生僻字处理的服务管理机制包括但不限于如下三类。

7.1.2.1 设置服务点生僻字处理专员

建议有条件的服务机构在线下各省/市的一级服务网点/站点设置生僻字处理专员，该专员应熟悉生僻字有关知识，熟练使用生僻字处理手册，并积极配合生僻字客户办理业务。

7.1.2.2 开展生僻字处理培训

建议服务机构开展生僻字处理培训。培训内容包括但不限于编码字符集、中文字符处理、生僻字处理、业务操作中生僻字处理等。培训可由单位生僻字专业技术团队计划和开展，按需求定期向员工开展培训。

7.1.2.3 建立生僻字专业知识库

建议服务机构将生僻字处理方法纳入知识库。定期收集、分析生僻字客户

的投诉以及处理过程，形成常见问题案例及处理指南。知识库中的生僻字相关案例、信息在各服务机构间共享。

7.2 个人用户

个人用户遇到生僻字处理问题时，建议通过如下步骤逐步排查具体情况和定位问题：

1）拼音输入法用户通过现代纸质辞书或线上辞书查找该生僻字的确切读音。如果辞书上已提供确切读音且尝试输入后得到需要的字，则问题解决；（注意不要随意拷贝线上辞书上用以显示的字头，可能会造成新的问题。）

2）如果输入拼音不能得到需要的字或没有查到确切的拼音，可以尝试用笔画等输入法输入；

3）在 https://name.vurls.cn/（姓名生僻字处理平台）、https://zi.tools/（字统网）、http://ccamc.org/（古今文字集成）等网站中，通过检索部件查到所需汉字的码位，也可从这几个网站上直接复制具体的字符；

- https://name.vurls.cn/的“户籍查字”可以通过部件检索到正式码的汉字，也可检索到 PUA 码的汉字，拷贝到别的业务系统中，虽然可能无法显示，但由于编码无误，如果继续在业务系统中操作，很多情况下可以顺利办理业务；
- https://zi.tools/的“组字”可以通过 IDS 检索到 GB/T 13000 所有已编码的字符，也包括部分尚未编码的条目（这部分不以字符的形式展现），若一般用户对专业的 IDS 不甚了解，也可以通过“搜字”进行简单查字；
- http://ccamc.org/的 IDS 查询也提供 IDS 检索，该网站提供所有已编码字在 GB 18030-2022 中的码位。

4）以上步骤均无法查到需要的字时，可先在“姓名生僻字处理平台”（https://name.vurls.cn）的“户籍查字”栏目中点击“登记生僻字”按钮登记生僻字，也可根据本文档附件《提交补充编码汉字的技术要求》的要求提交至全国信标委，以期尽快在国家标准和国际标准中收录。

第八章 测试评估

8.1 信息处理产品的标准符合性测试

8.1.1 编码字符集测试

8.1.1.1 测试方法

1）基于标准中相应实现级别的规定制备测试样本数据；

2）将样本数据输入至被测产品中；

3）验证输入是否成功；

4）执行产品后续功能，验证执行结果是否正常；

5）验证执行结果的输出是否正常（如：屏幕显示、打印结果、存储结果等）。

8.1.1.2 判定准则

1）当产品具备由用户自行录入信息的功能时，产品应可通过自身提供或第三方提供的输入方式，正确输入标准中定义的所有图形字符，否则为不合格产品；

2）当产品具备通过显示设备显示用户录入的相关信息时，产品应可通过自身提供或第三方提供的字库，正确显示标准中定义的所有图形字符，否则为不合格产品；

3）当产品具备打印用户录入信息的功能时，产品应可通过自身提供或第三方提供的字库，使用自身提供或第三方提供的打印机驱动程序及打印设备，正确打印标准中定义的所有图形字符，否则为不合格产品；

4）当产品具备将用户录入的信息存储为文件的功能时，存储完成后应能重新正确读取所有存储前的图形字符，否则为不合格产品；

5）当产品具备数据导入或读取功能时，应能正确导入或读取以标准中定义的编码格式进行编码的所有图形字符数据，否则为不合格产品；

6）当用户完成信息录入后，产品对数据执行内部处理（无须外部系统协作）时，应能正确返回预期结果，否则为不合格产品。

8.1.1.3 空码位的测试

空码位是指编码字符集标准中，给出了编码但并未给出对应的图形字符的编码位置。

空码位分为如下 2 种：

1）用户自定义区（私用区，PUA）

该类区域的编码空间在编码字符集标准中有明确的规定，这些编码空间是留给单机最终用户私有使用的，任何通用产品的开发、管理、运维方均不得占用，不得用于跨系统或跨机构信息交换领域。

GB 18030-2022 规定的用户自定义区包括：

- 双字节用户区：0xAAA1 至 0xAFFE、0xF8A1 至 0xFEFE、0xA140 至 0xA77E 和 0xA180 至 0xA7A0 共 1894 个码位；
- 四字节用户区：0xFD308130 至 0xFE39FE39 共 25200 个码位。

ISO/IEC 10646 规定的用户自定义区包括：

- BMP 平面的用户自定义区：U+E000 至 U+F8FF 共 6400 个码位；
- 辅助用户自定义区：U+F0000 至 U+10FFFF 共 131072 个码位。

2）保留区

该类区域的编码空间在编码字符集标准中有明确的规定，这些编码空间是留给标准化组织在未来的标准制修订过程中使用的，其他任何组织和个人均不得占用。如 GB 18030-2022 中明确规定：

其他未占用的四字节码位为保留区，留待未来文件扩展使用。

➢ **判定准则**

空码位对应的显示/打印结果，可以是下列三种形式之一，同时鼓励采用第三种形式：

1）空白（占据一个字符位置）或 ISO/IEC 14496-22 中建议的形式；

2）方框或带×的方框（占据一个字符位置）；

3）显示在方框内的当前字符对应的编码（占据一个字符位置）。

8.1.1.4 一字多码问题的测试

一字多码问题，是指在标准化过程中造成的下列情况：

1）同一标准中同一图形字符出现在不同编码位置；

2）存在编码映射关系的标准之间，因标准修订不同步，导致原映射关

系变更至新码位；

3）某些机构早期通过占用 PUA 对一批急用图形字符分配了自定义编码，导致这些字符与现行标准编码不一致。

➢ 判定准则

1）对于同一标准中出现在不同编码位置的同一字形，产品应在每个码位均实现该字形；

2）对于原映射关系变更至新码位的字符，应采用现行有效标准的最新版中的定义，同时鼓励兼容原映射关系；

3）对于占用 PUA 编码的字形，应尽快转换为对应的现行标准编码。暂时无法完成转换的，应提供辅助功能以保障信息交换、处理等功能正常实现。涉及与外部系统信息交换的，应提供与现行标准编码兼容的处理功能。

8.1.1.5 对国际标准的支持

生僻字处理相关的国际标准主要是指 ISO/IEC 10646。我国的国家标准 GB/T 13000 等同采用了该标准，但由于标准化过程需要花费时间，所以 GB/T 13000 的最新版通常滞后于 ISO/IEC 10646 的最新版。

同时，GB 18030 与 ISO/IEC 10646 之间存在一一对应的码位映射关系，二者所收录的图形字符在码位上始终保持一致，但在收录时间上往往也会存在一些差别。

因此，产品在符合 GB 18030 的基础上，对 ISO/IEC 10646 的支持程度是评估产品生僻字处理能力的一项重要补充。

➢ 范围

ISO/IEC 10646 旨在定义世界上曾经出现过的所有文字的图形字符，与生僻字处理相关的部分仅涉及一部分子集，主要包括 CJK 统一汉字子集及 CJK 统一汉字诸扩充集（目前包括 CJK 统一汉字扩充 A 至 H）。对于信息处理产品生僻字处理能力的评估，应以这些子集为主。

➢ 判定准则

ISO/IEC 10646 标准符合性判定准则可参照 GB 18030 执行。

8.1.2 字型测试

产品涉及的字型，应符合 GB/T 11460 的要求，产品的检测按 GB/T 11460

中的相关要求执行。

适用于生僻字处理的字型产品，其字汇应至少达到 GB 18030-2022 实现级别 3 的要求。

8.1.3 输入法测试

- **测试方法**

采用被测产品，逐字输入标准中定义的所有汉字字符；

查看产品是否能正确完成所有字符的输入。

- **判定准则**

用于生僻字处理的输入法产品，能输入的字汇应达到 GB 18030-2022 实现级别 3 的要求。

除字汇外，输入法还应符合相关国家标准的要求。

8.2 信息处理系统生僻字处理能力评估

8.2.1 定义

信息处理系统以一个或多个信息处理产品为基础，除提供这些产品所具备的功能外，还应提供保障这些产品在处理生僻字时仍能正常提供所有功能。

涉及居民个人信息处理的信息系统的生僻字处理能力主要包括：

1）标准符合性；

2）易扩展性；

3）经济适用性；

4）兼容性；

5）易用性；

6）统一性。

8.2.2 标准符合性

1）国家标准符合性；

2）国际标准符合性。

8.2.3 易扩展性

1）技术方案是否可扩展；

2）技术方案是否安全可控。

8.2.4 经济适用性

1）完成升级改造所需的成本；

2）推广使用所需的成本。

8.2.5 兼容性

1）是否能在多种操作系统环境下运行；

2）是否能兼容常用的输入/输出设备；

3）是否能与多种外部系统正确交换信息（如：电子邮件、各种格式的文档、即时通讯、数据库文件等）；

4）是否具备识别非标编码数据的功能；

5）是否能兼容处理非标编码数据；

6）是否具备将非标编码数据转换为标准编码数据的功能；

7）是否具备“一字多码”问题的处理能力。

8.2.6 易用性

1）为内部服务人员（如：柜员等）进行培训以达到提供该项服务的难度；

2）为用户提供培训（如：在线帮助等）以使用户能够完成享受该项服务所必须执行的操作的难度；

3）用户为享受该项服务必须对客户端（如：手机等）进行设置等操作的难度；

4）充分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；

5）当用户无法自行完成相关操作时，是否有渠道得到帮助；

6）帮助是否及时；

7）帮助是否有效；

8）用户通过帮助仍无法自行完成时，是否能通过应急渠道得到妥善处理。

8.2.7 统一性

涉及现有标准中暂未收录的生僻字时，系统中不同模块所采用的解决方案应处处保持一致。

8.3 服务机构业务生僻字处理能力评估

8.3.1 定义

机构的生僻字处理能力主要包括：

1）技术管理机制

2）服务管理机制

8.3.2 技术管理机制

1）是否制定了管理机制，明确本机构所提供的信息系统及接口服务所应达到的生僻字处理能力；

2）是否制定了相关制度，明确本机构研发、运营的信息系统及接口服务在生僻字处理方面的采标要求；

3）机构是否为生僻字处理配备了专业的标准化团队；

4）机构是否为生僻字处理配备了专业的研发团队或引入技术支持与解决方案，技术团队是否具备足够的专业能力；

5）机构是否为生僻字处理配备了专业的运维团队，运维团队是否具备专业能力以解决系统运营中出现的各种生僻字处理问题。

8.3.3 服务管理机制

1）机构是否提供了有效的生僻字处理渠道，例如设置了线下生僻字处理专员；

2）生僻字问题处理的熟练程度；

3）机构是否制定了生僻字处理的相关制度或操作规程；

4）机构是否建立了生僻字处理业务培训机制并提供足够的培训；

5）机构是否建立了生僻字处理专业知识库，用以定期收集、分析生僻字客户的投诉以及处理过程，形成常见问题库及处理案例；

6）机构能否通过行业间、产业间的相关生僻字信息共享平台，与其他外部机构建立互动。

第九章 实用工具及资源

9.1 对本章所列资源的说明

本章提供的产品、工具相关内容由厂商提供，未经过第三方机构验证，请用户在选用时关注这些产品、工具的标准符合性程度。

9.2 软件产品

9.2.1 超大字符集字库

➢ “国标”系列超大字符集字库

“国标”系列超大字符集字库是中国电子技术标准化研究院依据强制性国家标准 GB 18030-2022《信息技术　中文编码字符集》开发的字型产品。研发过程中联合业界文字学专家和字形专家，开展了数万字的逐字考证、正形工作，确保字形的正确性和规范性。为我国汉字的信息技术应用提供字形范本，也为软硬件产品和字型产品的标准符合性检测提供字形依据。产品包括曲线字型和点阵字型产品两类：

1）曲线字型

“国标宋体-超大字符集”曲线字库，当前版本包含 87887 个汉字，字体风格为宋体，支持 GB 18030-2022 的实现级别 3（全集），解决信息系统不能显示生僻字字形的问题。除宋体字型外，国标系列曲线字库还包含“国标黑体”“国标仿宋”“国标楷体”“国标小标宋”等字体风格，可支持 GB 18030-2022 的实现级别 1 和 2，适用于党政公文/日常办公等应用场景。

2）点阵字型

支持 GB 18030-2022 实现级别 3 的超大字符集国标点阵字型共有三款：15×16 点阵、24×24 点阵宋体和 48×48 点阵宋体；同时还有从 11×12 点阵到 64×64 点阵不等的多种规格点阵字型产品可支持 GB 18030-2022 实现级别 1 和 2。

➢ 方正超大字库

方正超大字库是方正的一款经典产品。其第一版方正宋一体超大字库包含 7 万多汉字，是北大方正于 2002 年开发完成，并首家通过了由国家新闻出版总署、国家语言文字工作委员会全国印刷字体工作委员会联合主持的鉴定。方正超大字库最新版本包含 CJK 和康熙部首、包含国家标准 GB 18030-2022《信息技术　中文编码字符集》收录全部汉字 87887 个（CJK 统一汉字、CJK 统一汉字扩充 A、扩充 B、扩充 C、扩充 D、扩充 E、扩充 F），支持《通用规范汉字表》全部汉字。该产品开发之初北大方正邀请了多位文字学专家进行指导，对字库中的汉字进行了新笔形化，同时邀请多位字体设计专家对设计质量进行把关，保证字形、笔形规范、统一，符合国家语言文字工作委员会的标准。方正宋体超大字库同时也是 GB 18030-2022 标准文本的专用字库。该产品可以完美解决人名、地名缺字问题，解决专业辞书和古籍排印用字问题，已经在银行、邮电、保险、报业、信息传媒等行业得到广泛应用。

➢ 汉仪中黑 TS（汉仪）

汉仪中黑 TS，是一款传统经典风格黑体，最新版本参考国家标准 GB 18030-2022《信息技术　中文编码字符集》，收录全部汉字 87887 个。该字库分为 P1、P2 两个字体文件：P1 部分覆盖 GB 18030-2022 实现级别 2，其中包括 CJK 基本区、CJK 扩充 A 共 27570 汉字和 196 个《通用汉字规范表》汉字，并包含非汉字部分康熙部首 214 个和 CJK 补充部首 14 个。P2 部分包括除级别二之外全部汉字共 60121 汉字，其中包括 CJK 扩充 B 区 42675 个、扩充 C 区 4105 个、扩充 D 区 214 个、扩充 E 区 5654 个、扩充 F 区 7473 个。两个文件之和符合 GB 18030-2022 实现级别 3。

该字体字形以国家标准 GB/T 22321 字形为参考，笔画风格规范统一，视觉效果优质均衡。

➢ 知春宋体与知春等线体系列

北京卓玛之裕科技有限公司的知春宋体字库，在 Unicode 15.0 发布的当天即实现对 CJK 统一汉字近 10 万字的全覆盖，全面支持强制性国家标准

GB 18030-2022《信息技术　中文编码字符集》实现级别 3 要求的 87887 个汉字，全面支持金融行业标准 JR/T 0253-2022《金融服务　生僻字处理指南》中完整级汉字。知春宋体的精简版本曾提供给人民银行各地分支机构使用开展针对姓名生僻字的调研工作。

知春等线体系列同样全部覆盖 CJK 统一汉字近 10 万字，目前提供超细、特细、细、微细、标准、微粗、中等粗、粗等共计 8 个不同字重的版本。

➢ 遍黑体

这是一款在开源的思源黑体的基础上，进一步补充字数的字库，其设计风格沿用黑体，字形依据《通用规范汉字表》所规定的字形，更方便满足手机系统字库在我国使用上字数不足的问题，并在使用和再开发上维持开源和免费商用的特质。该字库分为两个字体文件，P1 包括扩充 A 区 49 个、扩充 B 区 6743 个、扩充 C 区 4153 个、扩充 D 区 222 个（完整）、扩充 E 区 5762 个（完整）、扩充 F 区 4868 个，P2 包括扩充 G 区 4939 个（完整）、扩充 H 区 4192 个（完整）。该字库仍在发展中，日后还将紧贴 GB/T 13000 和 GB 18030 的后续版本，继续增加新编码的字符。

注：该字库在 PUA（用户自定义区）放置了大量汉字，机构用户使用之前应先将其中 PUA 区的字形删除，以符合 GB 18030 关于自定义区的要求。

➢ BabelStone Han

这是一款由国际标准化组织和 Unicode 的参与专家开发的宋体风格的字库，其在文鼎科技的开源字体 AR PL SungtiL GB 的基础上发展而来，字形依据《通用规范汉字表》所规定的字形。该字库仅包含一个字体文件，是一款可免费商用的字库。其基本目标为覆盖 GB/T 13000（即国际标准 ISO/IEC 10646）上明确为在中国使用的字符。汉字基本区包括 20992 个（完整）、扩充 A 区包括 4593 个、扩充 B 包括 14025 个、扩充 C 包括 2197 个、扩充 D 区 222 个（完整）、扩充 E 区包括 3346 个、扩充 F 区包括 2975 个、扩充 G 区包括 3019 个、扩充 H 区包括 2266 个。该字库更新频繁，日后还将紧贴 GB/T 13000 和 GB 18030 的后续版本，继续增加新编码的字符。

注：该字库在 PUA（用户自定义区）放置了大量汉字，机构用户使用之前应先将

其中 PUA 区的字形删除，以符合 GB 18030 关于自定义区的要求。

9.2.2 输入法

➢ 搜狗输入法

腾讯搜狗输入法支持 Windows、Linux、Android、iOS、鸿蒙、RTOS、嵌入式 SDK、云输入等全部系统的输入方案，客户端活跃用户量 6 亿+。2022 年 11 月腾讯搜狗输入法发布“汉字守护计划”，率先发布支持 GB 18030-2022 全量汉字的安卓、iOS 手机输入法。

腾讯搜狗输入法在支持拼音、笔画、五笔、拆字、手写等输入方式基础上，为大字符集设计了专用的“生僻字键盘”。针对用户只识字形、不知读音的情况，腾讯搜狗输入法最新开发了“生僻字键盘”，通过“拆字”，实现笔画输入、拆分部首的拼音输入，以及部首拼音+笔画的混合输入 3 种便捷输入。

➢ 卓码输入法系列

卓码生僻字处理系列软件由中国科学院软件研究所和北京卓玛之裕科技有限公司合作研发，目前已实现对 Unicode 15.0 收录的全部近 10 万汉字（含扩充 G、扩充 H）的全覆盖，全面支持强制性国家标准 GB 18030-2022《信息技术　中文编码字符集》实现级别三要求的 87887 汉字，全面支持金融行业标准 JR/T 0253-2022《金融服务　生僻字处理指南》中完整级汉字。

1）卓码输入法

卓码输入法支持 Windows XP/7/8/10/11、麒麟 Linux、统信 UOS、红旗 Linux、方德桌面操作系统、macOS、Android、iOS、鸿蒙等操作系统平台，并针对使用 Windows 触摸屏的自助设备提供带软键盘的版本。

卓码输入法支持拼音、拆字、笔画、五笔、Unicode 直录、人名地名生僻字专用码等多种录入方法。卓码输入法以拼音和拆字拼音为主的录入方式，最符合普通用户当前使用输入法的习惯，配合精心设计的录入提示，真正做到了录入生僻字零门槛。卓码输入法以创新的方式附带汉字属性小字典，将汉字的拼音、编码、IDS、《通用规范汉字表》中的分级和编号等一系列丰富的属性信息呈现给用户，给用户更多的用字指引。

卓码输入法在解决生僻字录入问题的同时，兼顾日常使用的需要，收录

了大量生僻字人名、地名词语、金融词汇。

2）卓码生僻字显形工具软件

在 Windows 平台上，与卓码输入法配套的卓码生僻字显形工具软件从操作系统底层解决生僻字的显示问题，能够让大量应用软件不需修改源代码绑定特定生僻字字库，在使用系统缺省字库的情况下即可显示生僻字，真正做到“无感显字”。同时，针对 PUA 区的监测可有效避用户有意或无意使用其他应用软件录入与二代身份证不兼容的 PUA 字。

3）卓码查字

在 Windows 操作系统上，还提供卓码查字软件，支持拼音、拼音+总笔画数、部首、部首+总笔画数、拆字、拆字拼音、五笔、Unicode 直录、笔画（拼音笔画）、笔画（数字笔画）等共计 10 种查字方法，可快速找到需要的生僻字。

4）卓码云输入法

面向 Web/H5、Android、iOS、小程序端，卓码还提供云输入法和云字库，支持拼音、拆字、拆字拼音、五笔、笔画等多种查字方法。并提供基于服务器端的重量级部署和纯移动端的轻量级部署两种方案。

5）其他

除此之外，卓码还针对生僻字开发了卓码五笔输入法、适用于排队叫号机的生僻字补音组件（语音合成）、输入法生僻字支持能力自动检测软件等一系列产品。

➢ 方正超大字库输入法

方正超大字库输入法最早是解决字典辞书类书籍出版时超大字库的输入问题，随着国家标准 GB 18030 不断扩充，方正紧跟国家标准持续迭代，目前方正超大字库输入法最新版本已支持 8 万多汉字的显示、录入，全面支持国家标准 GB 18030-2022 全部汉字 87887 个（CJK 统一汉字、CJK 统一汉字扩充 A、扩充 B、扩充 C、扩充 D、扩充 E、扩充 F）。该产品根据用户使用场景，提供了 Windows 平台、信创平台（统信、麒麟以及红旗）、云平台和移动平台（Android 和 iOS）等多平台输入解决方案，可以完美解决人名、地名生僻字输入问题。

1）Windows 平台输入解决方案

Windows 平台提供独立的客户端程序，采用方正新典码输入技术，能直观地输入汉字，不需要记忆任何字根或编码，便于普通用户的学习和掌握。提供部首、笔顺和拼音三种相结合的检索方式，用户可以根据自己熟悉方式自由选择。该方案有效解决 Windows 平台超大字库字符的显示和录入问题。

2）信创平台输入解决方案

信创平台输入解决方案采用通用输入法架构，提供笔顺和全拼和拆分部件拼音三种检索方式。适配统信、麒麟和红旗操作系统。该方案有效解决信创平台超大字库字符的显示和输入问题。

3）云平台&移动平台输入解决方案

云平台和移动平台采用 Web Font 技术，通过构造一个输入键盘，用户在键盘上点击笔顺信息（横竖撇点折），动态显示候选结果，筛选出目标生僻字，完成目标字符输入。该方案有效解决云平台和移动平台超大字库字符的显示和输入问题。

该产品主要用于出版、政务以及公共服务等领域。

➢ 百度输入法

百度输入法是百度推出的一款人工智能输入工具，支持拼音、五笔、手写、语音、笔画等多种输入方式，具备智能调频、智能联想、智能纠错、智能预测、智能混输等算法，依托百度十多年搜索经验积淀，实现词库丰富、出词精准、联想智能，给用户带来极致的输入效率和体验。

百度输入法已实现全平台覆盖，支持 Windows、Linux、麒麟、UOS、中科方德、macOS、Android、iOS 等操作系统以及云输入法。

在生僻字系统升级方面，百度输入法提供拼音、手写、拆字（例如：通过“日木”输入杲，通过“龙天”输入奘）、笔画等多种方式录入生僻字，通过智能匹配算法降低用户学习成本、轻松录入生僻字。百度输入法各个平台的产品以及云输入法均已支持生僻字录入，结合百度大字库以及一系列生僻字服务，可以为客户提供全场景的生僻字解决方案，助力企业信息化升级，保障公民数字生活权益。

9.3 相关工具

9.3.1 字符转码工具

1）iconv 是 Linux 平台经典的编码字符集转换工具，支持全球绝大多数的编码字符集。

2）ICU 原是 IBM 开发的一个编码字符集转换工具和函数库，现由 Unicode 维护，会及时跟进 ISO/IEC 10646 新版本，支持多种操作系统平台。

3）Windows 平台上，MultiByteToWideChar 系列的 API 实现编码字符集之间的转换。

4）Java 编程语言中，String 类的 getBytes 方法可以得到对应编码字符集的字节数据。

5）Python 的字符串相关的 encode 和 decode 函数可实现编码字符集的转换。codecs 提供了对文件读写的自动编码转换功能。

需要注意的是，GB 18030-2022 调整了个别码位与国际标准 ISO/IEC 10646 的映射关系，各个编程语言尤其是低版本的函数库一般都未及时更新到与最新标准一致，有些库的编码转换功能对 GBK 编码的转换不支持自定义区的字。

9.3.2 字符转义工具

一般是相关的编程语言中提供 escape/encode/decode 之类的函数实现对字符的转义。针对生僻字的转义是在 GBK/EBCDIC 环境下存储生僻字而催生出的需求，目前有中信方案（中信银行）和招行方案（招商银行），中信方案是“`Hxxxxx”，其中，连续的 x 表示 5 位 UCS 编码，招行方案是“[U+xxxx]”或“[U+xxxxx]”。中信的方案已经开源，招行方案由供应商实现。由于转义处理会增加字符串长度，在使用转义处理的系统中，需要考虑对定长报文和数据库栏位宽度的影响。

9.3.3 公共组件

- 字符编码诊断

“姓名生僻字处理平台”中的“编码转汉字”栏目提供汉字与编码互转功能。（https://name.vurls.cn/ui/index/uchar/）

“姓名生僻字处理平台”中的“藏宝阁”提供 HexDisplayer 下载，可在 Windows 平台上实现汉字与 UCS 码的互相转换。

“姓名生僻字处理平台”中的“藏宝阁”提供“码字互转”APP，可在 Android 平台上实现汉字与 UCS 码的互相转换。

- **字库环境诊断**

“姓名生僻字处理平台”中的“设备文化程度检测”页面（链接：https://name.vurls.cn/ui/index/zktest/）提供了部分生僻字的实例及对应的正确显示的截图，可验证浏览器对生僻字的显示能力。

“姓名生僻字处理平台”中的“生僻字样例”页面（链接：https://name.vurls.cn/ui/index/sample/）提供了数百个 CJK 各个编码区块的生僻字实例。

9.4 论文文献

[1]马良有:正则表达式作用于汉字姓名生僻字方法初探[J].金融电子化,2021(10): 54-56.

[2]纪熙东、李言平：姓名生僻字的银行标准化实践和建议[A]. 中国标准化协会、郑州市人民政府.第十六届中国标准化论坛论文集[C].中国标准化协会、郑州市人民政府:中国标准化协会,2019:6.

[3]艾卓码：姓名生僻字应用的困境与对策[J].信息技术与标准化,2021(10): 77-82.

[4]杨健：一种基于编码转换的超 GBK 字符集解决方案[J].长江信息通信,2022,35(06):209-211.

[5]马征：银行信息系统生僻字问题探究[J].金融发展研究,2021(02):88-89.

[6]刘建军、杜晓、杨眉等：基础地理信息生僻字系统的设计与实现[J].地理信息世界,2016,23(1):133-135.

[7]李运富：论汉字数量的统计原则，郑州大学汉字文明研究中心（原载《辞书研究》2001 年第 1 期）

[8]尉迟治平：再论中文汉字字符集[J].语言研究,2020,40(01):78-89.

[9]保研险泡汤，不能让生僻字背锅[J].发明与创新(大科技),2019(12):43.

[10]范亚茹：文化传承视角下“生僻字”流行的再认识[J].汉字文化,2019(18): 38-39.

[11]吴汉江：同音字替代:消除地名生僻字的有效途径[J].中国地名,2017(09):21-22.

[12]雷天戈：生僻字勾起多少文化记忆[J].共产党员(河北),2016(08):47-48.

[13]徐剑锋：生僻字考验职能部门服务意识[N]. 西江日报,2015-08-06(F02).

[14]王旭东、金敖生：一个生僻字险让企业停产[N]. 浙江日报,2009-12-20 (002).

[15]丁芸：我国公民姓名用字中的生僻字分析[J].现代商贸工业,2009,21(18):240-241.

[16]许寿椿：汉字的技术性——机械化时代和信息化时代的比较[J].汉字文化,2009(02):80-84.

[17]邱莉芹：关于人名用字中生僻字使用情况的调查与分析[J].术语标准化与信息技术,2008(02):22-24.

[18]郑燕萍：名字使用生僻字的文化动因[J].汉字文化,2007(05):63-65.

[19]尉迟治平、汤勤：论中文字符集、字库及输入法的研制[J].语言研究,2006(03):63-66.

[20]戎明昌、程佳凌、郑蕾：粤 19 生僻字地名要改名引发争论[N]. 南方日报,2006-08-03(A06).

[21]许寿椿：新世纪,呼唤汉字的完整解决方案[J].汉字文化,2003(03):21-22.

第十章 实施案例

在信息系统生僻字处理和解决的问题上，地方政府、各级行业主管部门、行业应用单位以及软硬件研发生产厂商纷纷进行了积极的探索。本章征集了各类机构和单位在支持生僻字的升级改造方面所取得的经验，为未来计划开展信息系统生僻字治理的机构提供参考。

根据不同的改造范围，将实施案例划分为四个类别，分别是：

- 公共服务机构的系统内改造
- 大规模跨系统的公共服务升级改造
- 用字单位解决内部用字需求
- 通用的软硬件产品/服务的升级改造

下面按类别列举有关案例。

10.1 公共服务机构的系统内改造案例

➢ 案例一：中信银行系统改造

（一）问题描述

1）生僻字在业务处理层后端存储、链路传输的问题：2020 年之前，中信银行核心系统的主机使用 EBCDIC 编码，该编码字符集所收字汇相当于 GB 18030-2000，不支持 CJK 扩充 B 及之后的汉字。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 GBK 编码，不支持生僻字，部分系统升级为 GB 18030 或 UTF-8 编码的成本也很高。

2）生僻字在渠道层前端的显示、输入、打印的问题：当时柜面系统使用字库和输入法是操作系统自带的字库和输入法，人口信息字库 PUA 编码生僻字和 CJK 扩充 C 及之后的汉字不能显示也难以输入；柜面存折打印机也只支持 GB 18030-2000 的 27533 个汉字。网上银行、手机银行的字库依赖客户端。于是大量生僻字在各渠道无法显示，也难以输入。

（二）技术方案

1)采用通用的自定义编码扩展转义格式在小字符集编码(如 GBK、EBCDIC)中表示生僻字，开发通用 SDK（C 语言组件提供通用的静态库、普通动态库和 JNI 动态库及 Java 语言组件等)实现大字符集编码(如 GB18030、UTF-8、UTF-16 等)转小字符集编码时生僻字变为转义格式，字符集编码小转大时生僻字转义格式进行还原。升级方案不改变整体数据架构、技术架构，避免底层产品字符集升级而导致成本过高问题，外购系统或难以升级字符集的系统可保留小字符集编码不变，以保护系统资产投资。

2）兼容处理人口信息字库 PUA 与 UCS 正式码“一字多码”导致的身份证联网核查失败或转账校验账户户名与报文收款人姓名不符的问题，提升客户体验。

3）通过 ESB、通用文件传输平台作为交易报文转码、批量文件转换枢纽，减少整体系统改造的工作量与关联耦合。

4)采用云字库/云输入法与本地字库/输入法相结合的方式支持全面客渠道生僻字显示和输入。采购支持新国标的存折打印机实现柜面生僻字打印。

（三）实施策略

1）演进策略

①试点改造柜面与核心，保证生僻字客户可开户。国外主机核心采用 EBCDIC 的转义格式表示生僻字，核心下移后还原为 UTF-8。

②改造手机银行及相关链路的系统，保证生僻字客户方便使用常用及基本的金融服务。

③其它系统按关联系统的互相影响程度，分期分批分组改造各类系统。

2）新增策略

①新建、重构类系统，统一采用 UTF-8 编码标准。

②使用拆迁者模式，基于原有的业务，一次性将旧系统的数据、功能迁移到新系统上。

3）存量策略

结合特点分类处理，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 GB 18030 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，采用转义方式来表示生僻字。

4）保障策略

设置生僻字开关，避免上线回退风险，保持业务连续性。

（四）改造效果

中信银行在人民银行科技司的指导下，于 2020 年 4 月完成了全行系统生僻字改造。目前，生僻字在核心系统、柜面及手机银行等全渠道畅行无碍，为各行各业解决生僻字问题提供了切实可行的“中信方案”。

2021 年 9 月，中信银行又与北京金融科技产业联盟合作将全套生僻字解决方案、工具和软件进行开源，与全行业共享成果，推动生僻字社会难题的解决。

（五）生僻字改造后的现状以及未来 GB 18030-2022 的挑战

渠道层前端系统中，人工柜面安装了与户籍机关同版本的人口信息字库与输入法，实现了 8 万多字的显示与输入，网银、手机银行等电子渠道通过云字库/云输入法调用后台大字库系统，也是与户籍机关同版本的人口信息字库。柜面新采购的存折打印机已能支持 GB 18030-2005 的 7 万多字的打印。与 GB 18030-2022 发布后要求的 87887 个汉字还有一些差距。

交易链路中的架构各层各类系统间接口，已分别通过升级 UTF-8 或 GB18030 编码，或在 GBK 编码中使用转义方式表示生僻字等三种方法支持生僻字的传输。转义与链路传输支持的编码都已涵盖 GB 18030-2022 的编码范围。

在业务处理层后端系统中，数据库存储或文件存储也已分别通过升级 UTF-8 或 GB18030 编码，或在 GBK 编码中使用转义方式表示生僻字等三种方法支持生僻字，在后端系统的转码和内部处理中，涉及姓名比对、联网核查的两个功能，通过对照表实现了人口信息字库 PUA 与 UCS 正式码“一字多码”的兼容处理。GB 18030-2022 发布后，生僻字存储相关的部分无需升级即已支持其编码范围，但“一字多码”的兼容处理对照表还缺少 GB 18030-2022 相对人口信息字库增收的生僻字。

（六）当前及后续工作

GB 18030-2022 已于 2023 年 8 月 1 日正式实施，我们正进行如下的工作来实现 GB 18030-2022 的支持：

1）联系厂家更新渠道层人工柜面的人口信息字库与输入法，并升级大字库系统。

2）联系厂家更新渠道层存折打印机，或新采购一批支持 GB 18030-2022 的

打印机。

3）联系厂家获取更新的“一字多码”对照表，升级涉及姓名比对、联网核查的相关业务处理层系统的转码组件文件。

➢ 案例二：招商银行系统改造

（一）问题描述

由于招商银行使用的主机系统使用基于 EBCDIC 编码 DBCS 字符集，实际可以使用的编码仅能符合 GB 18030-2000。旧有的编码仅有 GB 18030 双字节部分的汉字，可以与主机映射。对于 GBK 和扩充 A 区共有的 52 个 GB/T 13000 双码字，如【龚】：U+E863、U+4DAE 与 GB 18030 中的 0xFE9F，客户、财务、他行来的业务信息，可能是其中一个码，经常无法入账，受到的投诉很多。

（二）技术方案与实施策略

1）由于在主机迁移上云过程中，所以不安排与主机相关的生僻字改造，而聚焦于投诉较多的历史原因造成 GB/T 13000 双码字不能兼容的问题上，兼容客户和柜员用 PUA 和正式码的输入，按 GB 18030 的要求在前端、后端统一处理成标准正式码（U+4DAE 与 0xFE9F）。

2)对于能映射到 GB 18030 双字节部分对应的 PUA 字(U+E000～U+E765)，顺便同时加以实现。

3）在新建系统和数据库时，凡是涉及客户姓名、地址的，统一要求升级到 UTF-8 进行存储和交换。

4）其它问题待主机上云后一并进行 GBK 升级 UTF-8 改造。

（三）实施步骤

1）本行开户时，完成实名联网核查后，优化以正式码存 52 个双码字。

2）本行代发代扣、还款时，52 个双码字，无论客户输入哪个码，前端先规范成正式码，进行后续业务处理。

3）本行转账，52 个双码字，无论客户输入哪个码，前端先规范成正式码，进行后续业务处理。

4）人行、网联、银联等他行提回转账，52 个双码字，无论客户输入哪个码，前置先规范成正式码，再送主机入账处理。

5）主机上云后，对有正式码的姓名，参考前 4 条的做法，继续加以优化。

6）宣传引导客户、账务，尽量遵从标准，使用正式码处理相关业务。

➢ 案例三：粤康码对生僻字的处理办法

疫情期间，一些姓名中有生僻字的群众遭遇了获取不到健康码的窘境。广东省粤康码在遇到生僻字受困群众反应获取不到健康码时，会向当事人手机发送如下一段提示：

【粤省心】姓名里有生僻字无法登录粤省事和使用粤康码操作如下：【第一步】进入生僻字造字链接 https://name.vurls.cn/ui/index/rkxx/，在搜索栏，将生僻字的偏旁和部首拆分输入，点击提交。【第二步】复制 PUA 码字短信通知用户在人脸识别时复制上去这个字进行识别。请勿直接通过微信或企业微信发送，会被转译导致无法识别。同时，优先建议使用 APP 登录，避免使用微信小程序时，被微信转译导致无法登录。

粤康码的做法虽然没能从根本上解决生僻字问题，但能够在不给群众增加明显负担的情况下，实实在在地解决群众困难，仍然值得推荐。

10.2 大规模、跨系统场景的生僻字问题解决

➢ 案例：上海市公共服务领域生僻字改造

（一）背景介绍

目前上海市在电子政务和公共服务领域各主要信息系统的生僻字处理方式选择就存在较大差异：

户口和居民身份证：以公安部门全国统一标准制发的户口和居民身份证是公民最重要的身份证明类证件，2004 年以来，为解决因姓名中含有生僻字造成计算机系统无法录入的问题，公安部门建立了公安人口信息专用字库（人口信息字符集），有效解决了绝大多数姓名中含有生僻字的公民正常办理户口登记和居民身份证的问题。

社保卡相关部门：在上海市，社保卡由上海市社会保障卡服务中心（以下简称社保卡中心）制发和管理，在申领和使用过程中需要与人社局进行信息交互。社保卡最大的使用场景——就医过程中，社保卡的信息还需要经过医院的 HIS 系统与医保中心的系统进行交互。也就是说，社保卡中心、人社局、医保局以及医院联动支撑了社保卡的就医场景，因此在对待含有生僻字的姓名时处理方式应一致。目前相关单位统一使用一个 21 世纪初的自造字软件支持生僻字的录入和显示，确保就医信息流闭环内生僻字处理规则的一致性，但该软件早已不再更新。

其他部门：选择方式较为多样，以采用同音常见字替代、拆字组合替代或者拼音替代为主。

随着近几年上海市大力推广“一网通办”，居民经实名认证后在线上线下享受公共服务的场景越来越多，而居民的实名认证基本都是依赖于居民身份证上的信息，但是由于公安人口信息专用字库并不是公开的技术标准，造成公安部以外的部门的信息系统无法处理公安人口信息专用字库专有的生僻字，一些居民的身份证上明明打印上了生僻字，但是在持身份证办理其他部门公共服务事项时还是困难重重，引起较大的不解和不满。此外，由于不同的条线部门在生僻字处理上方法不一，也造成“一网通办”建设过程中数据治理难题，对于部分含有生僻字的姓名数据，在信息比对和交互时会出现非业务因素的问题，造成这些居民的部分业务无法顺利办成，给这些居民带来不便。

为填平这一历史和技术原因造成的“数字鸿沟”，解决人民群众“急难愁盼”问

题，真正落实“人民群众人民建，人民城市为人民”的为民发展理念，2021 年 9 月，上海市经济和信息化委员会、上海市大数据中心联合发出《关于开展信息系统生僻字问题改造的通知》，拟针对居民日常生活常见服务场景中涉及的信息系统进行统一技术路线的生僻字支持改造，争取达到让姓名中含有生僻字的居民也能在“一网通办”时代顺畅享受各种公共服务。

（二）总体思路

由于身份证是居民的法定身份证明证件，因此本次生僻字改造的技术路线原则上向身份证靠拢，确保含有生僻字的姓名从源头上统一处理规则。

落到改造的技术思路上，就是利用公安人口信息专用字库针对各个信息系统的改造工作，具体又分为需求侧和供给侧两部分，其中供给侧由市大数据中心牵头，通过购买服务方式在市电子政务云上提供基于公安人口信息专用字库的 PaaS 服务（以下简称生僻字 PaaS 服务），通过 API 和 SDK 方式对外提供生僻字的“显示、输入”功能；需求侧是全市各个系统调用生僻字 PaaS 服务，在本系统内实现生僻字的显示、输入、打印等功能，并确保外部数据交互过程中生僻字处理的规则统一性，对于部分因网络原因无法调用生僻字 PaaS 服务的系统，可酌情采购公安人口信息专用字库的本地版字库或者私有 PaaS 服务。

本次改造的需求端涉及政府部门、医院、药店等医疗机构、部分银行、保险等金融机构、医院、药店等医疗机构等，初步统计涉及系统 50 余个。

（三）典型经验

1）PaaS 服务使用

生僻字 PaaS 服务采用 API 和 SDK 方式提供功能，对于基于浏览器的系统或者移动端 APP，都可以通过内嵌生僻字 PaaS 服务的 SDK 实现对生僻字的支持。以基于浏览器 B/S/S 架构的系统为例，系统在整合生僻字 PaaS 服务的 SDK 后，可以加载生僻字 PaaS 服务的 JavaScript 文件，在页面上调用其包含的特定组件，可以实现生僻字的显示和输入，其中输入的时候是借助一个生僻字专用输入法的 iframe 层实现，效果如图 10-1 所示：

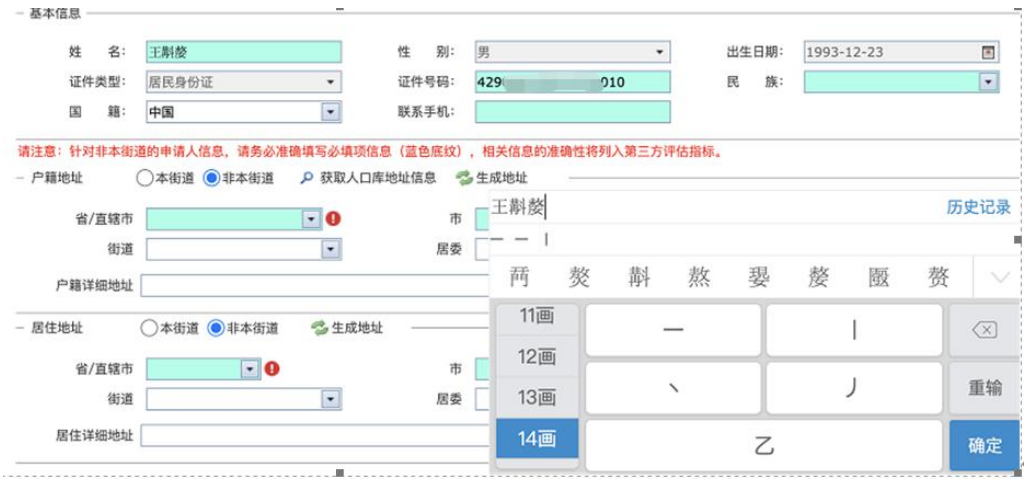


图 10-1　用专用输入法来输入生僻字

组件在显示生僻字或者打印生僻字时会根据内容的编码判定是否属于生僻字，如属于生僻字会向服务端请求该字的字体文件下载并用于显示，基于这种模式即便用户端没有相应字体也不影响显示。

对于传统的 C/S/S 系统或是因网络原因无法调用生僻字 PaaS 服务的，可以在服务端部署公安人口信息专用字库的本地版字库，并自行实现生僻字显示和录入（主要是字体文件的下载）。

2）系统改造

各个系统的改造基本分为前端和后端，前端需要在所有存在录入或者显示姓名的位置，按照生僻字 PaaS 服务的规范调用 SDK 相关方法，替换原生的 input 或者 label 标签，实现生僻字的录入或者显示。后端主要解决生僻字的持久化保存问题，生僻字的保存需要数据库支持 UTF-8 或者 UTF-16 的编码字符集，对于采用 GBK 甚至更小范围字符集的数据库，需要采用一些特殊的方法：

- 利用数据库特性，例如 Oracle 的国家字符集。一般我们说 Oracle 数据库采用 GBK 字符集，是指 Oracle 的 NLS_CHARACTERSET 参数值为 ZHS16GBK，也就是说设置为 char 或者 varchar 的字段只能记录长度为 2 字节范围在 GBK 内的汉字，如果是超过两字节的汉字（大部分生僻字）会被截断从而变成乱码。但是利用 Oracle 的国家字符集特性，即将 NLS_NCHAR_CHARACTERSET 参数设置为 AL16UTF16，那么设置为 nchar 或者 nvarchar 的字段就可以存储两字节以上的汉字。采用国家字符集特性后应用代码所需调整较小。
- 将生僻字转义为编码字符串进行存储，这个方案具有普适性，但在存储

和读取时有一个转码的环节，对应用代码以及性能影响较大。

系统中存量的生僻字也是改造的一个难点，这里一般分为两种处理方式：

- 对于无法识别出生僻字的存量数据（以前采用同音常见字或者拼音标识的，基本无法还原出原有的生僻字），一般只能通过日常业务或者与权威数据比对逐步更新。对于这类存量数据，在对外提供数据交互时，应尽量避免使用姓名作为查询条件，以免出现本系统内存量生僻字不符合统一的生僻字处理规则而造成数据交互失败的情况。
- 对于可以识别出原有的生僻字的存量数据，可以因地制宜制定解决方案。以上海的社保卡存量数据处理为例，由于社保卡相关单位之前使用一个统一的生僻字软件，因此是可以全量提取存量数据中含有生僻字的姓名（约 2 万余人），针对这些人，可以通知其在系统改造完成后前来换卡。

10.3 用字单位解决内部用字需求的办法

➢ 案例：文化艺术出版社

（一）问题描述

在出版行业文化艺术类的日常工作中，经常会遇到人名、地名、非遗等方面的生僻字，还需要考虑和既有字库的字体风格相协调。在传统的计算机排版到线下印刷厂打印，可以考虑使用图片或其他方式进行替代，但到了复合媒体时代，操作流程要求纸书的排版及其相关的数字资源不应进行过多的重复劳动。纸书的计算机排版工程文件所包含的数字化文本也应有效利用到数字资源中。过去的办法便不能满足新的需求。另一方面，图书销售和推广当前还广泛利用微信公众号。主要阅读场合为手机端的微信公众号在字库上完全依赖于操作系统的系统级字库，而这些系统级字库尚不能满足所有的生僻字需求，另外微信公众号对标点符号和视觉断行的处理都未能达到基本的出版要求。

（二）技术方案与实施策略

1）供复合媒体二维码使用的自主开发网站。MySQL 数据库编码选用 utf8mb4，网页前端（html、xml、php 等）编码选用 gb18030 或 UTF-8。在编辑过程中，遇到生僻字时，通过国际标准化组织表意文字小组 IWDS co-editor 提供的 IDS 数据（https://zi.tools/）与 ISO/IEC 10646（UCS）配套的 RSIndex 文档查找正式码位。在测试各系统的显示情况后，对任一操作系统中回退到.notdef 的生僻字或不符合《通用规范汉字表》与《印刷通用汉字字形表》所规定写法规则的字形添加特定字形到专用的 woff 字库中，若未编码的生僻字暂时使用 IDS+'ccmp' 技术处理，并通过 css 实现网页的全平台显示。

2）微信公众号。微信公众号并不允许用户写入 woff。当遇到任一系统中回退到.notdef 的生僻字或不符合《通用规范汉字表》与《印刷通用汉字字形表》所规定写法规则的字形时，通过 SVG font 嵌入页面。

3）搜集未编码生僻字形成数据表，通过适当渠道提交至以后的 IRG WS 中。对 IRG UCV 中要求需要进行认同的生僻字，考虑提交至 IVD。当编码确认且稳定后，一并更新数据表、字库、数据库和网页等。

（三）实施步骤

1）优先把握当前正在做的数字资源，并尽量做到纸书排版工程文件与数字

资源文本内容共享。

2）逐步回溯检查已经发布的数字资源，作出必要的更新。

3）探讨进一步完善电子书数字资源对生僻字的支持。

（四）改造案例

一本关于非遗传承的图书中，提及僤家人的音乐与舞蹈，复合媒体二维码使用的数字资源需要展示题目、介绍、音频名称等。“僤”字的 UCS 码位为 U+313BC，GB 18030 码位为 0x9A388534，另经查发现山人（Sunman）、人口信息和某权威出版社的 PUA 都包含这个字，但在制作网页时，已获得正式码，则决定只使用正式码，而不是其他。将 U+313BC 所对应的字形加入到专用的 woff 字库中，通过 css 指定在网页上显示。

10.4 软件产品支持生僻字的解决方案

➢ 案例一：腾讯搜狗输入法

（一）问题描述

由于互联网手机应用依赖系统字库，如果系统不支持大字符集，应用内无法直接调用显示，在系统不支持大字符集的情况下支持应用内大字符集显示需要新的技术方案。

（二）技术方案与实施策略

解决方案是通过合作大字符集字体，通过内置，完成应用内大字符集显示，有如下两个方案：

方案一：客户端直接内置大字符集字体包（30MB），或可采用引导用户后下载的方式。

方案二：云字体方式，按照调用在客户端显示，节省本地存储空间。云字体方式技术方案：应用检测用户上屏的内容是否包含生僻字，如果包含生僻字，请求云字体服务生成包含生僻字的云字体库并加载云字体库，刷新生僻字候选，生僻字可正常展示，点击候选内容完成输入上屏。

由于云字体受网络依赖较大，针对快速输入场景有一定体验损失，腾讯搜狗输入法采用方案一客户端直接内置大字符集字体包（30MB）。

（三）安卓/iOS 手机版搜狗输入法实施步骤

1）根据设备端汉字的支持情况，筛选出端上不支持汉字的 UCS 编码，使用汉仪或其他字体全量汉字集生成字体包。

2）引导用户下载字体包。

3）在需要文本展示的地方判断要展示的问题端上是否支持，若不支持则加载使用下载的字体包渲染。

➢ 案例二：永中 Office

（一）解决政务/日常办公的生僻字处理需求

作为国产 Office 办公软件的引领者，永中 Office 与国内字体库厂商进行战略合作，解决了 Office 办公软件用户对常用字、生僻字输入、编辑、打印体验方面的问题。

陇南市政府协同办公系统项目中采用了永中 Office 产品。安装永中 Office 办公软件时，将字体库同步进行安装。永中 Office 设置项中，设置“将字体嵌入文件”，在进行文件保存时，可以将内容中所涉及的字体库与文件共同保存，即能解决生僻字无法录入、无法显示的问题。

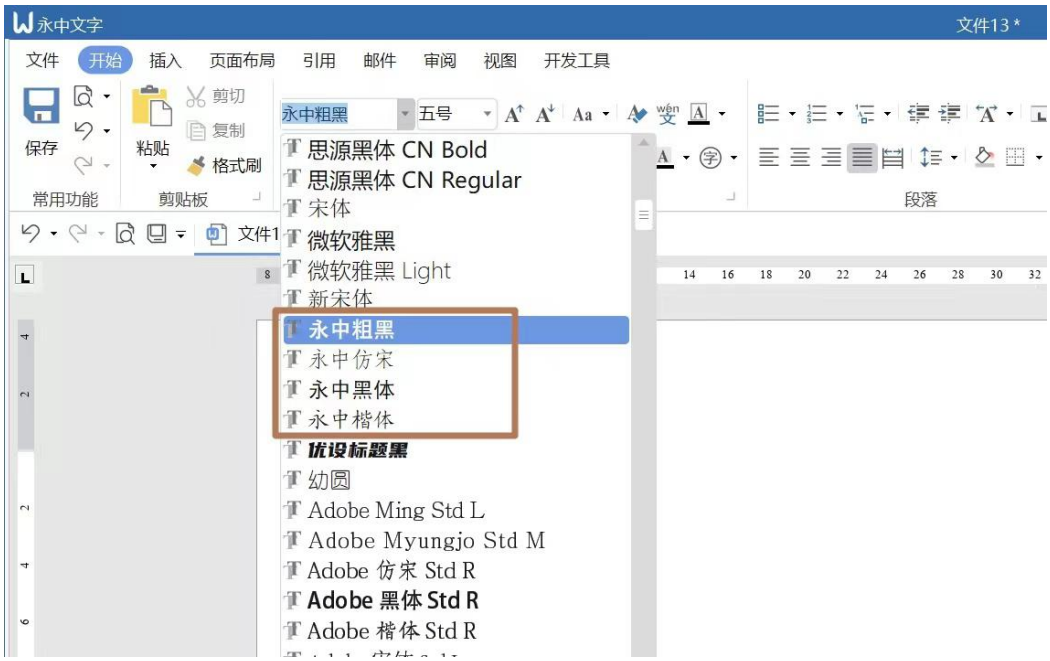


图 10-2　软件字体库下拉菜单

（二）利用版式文件解决生僻字显示问题

永中软件作为国产办公软件产品开发和服务提供商，拥有 Office 办公软件、OFD 版式办公软件、文档转换软件、数字档案管理系统等产品。在某市公安局数字档案项目中，通过永中数字档案管理系统直接调用 Office 软件进行文档的编辑，通过文档转换软件将流式文件转换为版式文件进行归档，流式文件转换为版式文件时，为解决生僻字无法正常显示的问题，永中软件利用自己本身产品提供的功能将字体库嵌入到 OFD 版式文件中，使问题得到解决。

➢ 案例三：支付宝生僻字实践案例

（一）生僻字改造前的问题

姓名中含有生僻字的用户在支付宝端遇到的困难存在于注册、认证、绑卡以及开放授权业务（例如健康码、社保、公积金、余额宝等）。问题包括：

- 前端输入时系统无法输入和验证；
- 即使通过粘贴完成输入展示仍然不正常；
- 证件审核中 OCR 识别对生僻字支持不到位，智能程度不足；
- 姓名含生僻字的用户普遍使用替代字方案，但是不同行也不同关联方方案不同。由此存在快捷绑卡无法完成关联认证，无法绑定多张银行卡，刷脸支付开放认证等多项功能受到影响。

这些问题的来源包括支付宝内部系统的问题也包括外部依存生态环境和交互环境问题。

（二）技术方案

为此，支付宝建立了生僻字解决方案，尝试寻找低成本解决问题的路径。主要包括：

输入端和显示端建设专门支持生僻字的组件，为其提供专用入口。

1）建立生僻字多编码字库。总结内外部接入机构的生僻字存在的情况，提供生僻字多种编码形式的映射，包括 UCS 正式码、各类替代字（如拼音、拆字、繁简异体字 UCS 正式码、ncr 编码等）等。为各业务接入生僻字提供基础，降低升级成本。

2）内部编码升级。支付宝在编码、传输、调用、数据库连接方面统一成标准格式。

3）Oceanbase 数据库全面支持 GB 18030-2005，目前按照 GB 18030-2022 要求正在升级，后续全面支持 GB 18030-2022。

（三）实施策略

该方案在实施中整体采用分批推进的策略。

1）从常见生僻字开始分批推进。从存量客户中梳理出常见生僻字，按照使用频率分批进入“生僻字多编码字库”。

2）前端优先解决常见生僻字输入和显示问题。

从“身份认证”环节开始，支付宝首先上线了独立组件“生僻字键盘”，输入拼音只显示可能对应的生僻字，解决普通输入法经常打不出生僻字的问题，方便用户快速找到并打出正确名字，从而通过身份认证、正常使用支付宝的线上服务。

在显示环节，支付宝联合阿里巴巴普惠体“生僻字计划”共同设计开发“生僻字”专用字体，配合“生僻字键盘”解决生僻字的输入和显示问题。

3）分场景分环节逐步接入。

根据总结的一字多码情况，技术实现跨机构联网核查。生僻字用户通过支付宝身份认证后，就可以正常绑卡、刷脸支付，并分场景分环节逐步接入，逐步实现查看公积金、医保卡、交管等系列政务服务。

4）客服进一步保障服务体验。识别生僻字用户，为之提供补充服务。

5）用户声音推动产品持续改进。收集生僻字用户的反馈，对生僻字解决方案中应用问题及时解决的同时推动产品进一步优化，不断丰富和完善功能。

附件：提交补充编码汉字的技术要求

为更好、更快地以标准化手段解决信息系统中地生僻字处理问题，凡国家标准未曾收录的汉字，可提交有关机构，申请补入相关国家标准。

一、待编码汉字的范围

凡 GB 18030《信息技术　中文编码字符集》和 GB/T 13000《信息技术　通用多八位编码字符集（UCS）》（idt.ISO/IEC 10646）均未收录的汉字，可提出编码。

待编码汉字应已经在一定范围内使用，且字形、字义、字音稳定。

用户自创汉字一般不予接受。在特定行业、领域获得认可、通行的新出现汉字可以提交，但应提供相关机构予以认可的证明。如科技用字，应提供国家语言文字主管部门、行业主管部门或有关专业机构（如全国科学技术名词审定委员会）予以认可的证明。

二、提交申请的单位

提交申请的单位应为政府机构或国有、民营企事业单位。不接受个人申请。

三、接受待编码汉字申请的单位

接受待编码汉字申请的单位是全国信息技术标准化技术委员会字符集与编码分技术委员会。

联系电话：010-64102866（秘书处）。

四、申请资料的技术要求

提交待编码汉字，请点击进入以下网址，在线填写《补充编码汉字申请表》：https://www.wjx.cn/vm/rBErS9m.aspx#

填写《补充编码汉字申请表》时的具体所需资料如下：

1、字形（必选）

待编码汉字的字形风格应为宋体。

如果字形与 GB 18030 或 GB/T 13000 所收字形相似，应附相关资料，说明其与相似字形的异同及其编码的必要性。

图片宜采用 jpeg 格式

2、字型（可选）

TrueType 或 OpenType 格式，宋体。

3、使用情况的文字说明（可选）

应提供下列信息，以证明对该汉字编码的必要性。信息包括但不限于：汉字的来源、使用者和使用场景的实例。例如：人名、地名用字的来源，在身份证系统和户口所在地的使用情况等。

文字须可以编辑，宜采用.doc 格式。

图片证据（4）或汉字属性信息（5）足以说明使用情况的，不必提交单独的文字说明。

4、使用情况的图片证据（必选）

文字说明之外，应附必要的图片作为证据。例如：身份证、户口簿或其它合法、有效的文件或证件，或使用场景的现场照片。涉及国家、单位、个人秘密或隐私的内容可以隐去。

图片应清晰，其中的文字应易于辨认。

图片宜采用 jpeg 格式。

5、汉字属性信息（必选）

汉字属性应包括下列信息，并以.xlsx 的文档格式提交：

1）康熙部首。待编码汉字均以《康熙字典》规定的 214 个部首归类。康熙部首在 GB/T 13000 中的代码、序号、字形见附表 1，24 个康熙部首的简化字形及其序号见附表 2。

附表 1 康熙部首的序号和编码位置

代码	序号	字形	代码	序号	字形	代码	序号	字形
2F00	1.0	一	2F48	73.0	曰	2F90	145.0	衣
2F01	2.0	丨	2F49	74.0	月	2F91	146.0	襾
2F02	3.0	丶	2F4A	75.0	木	2F92	147.0	見
2F03	4.0	丿	2F4B	76.0	欠	2F93	148.0	角
2F04	5.0	乙	2F4C	77.0	止	2F94	149.0	言
2F05	6.0	亅	2F4D	78.0	歹	2F95	150.0	谷
2F06	7.0	二	2F4E	79.0	殳	2F96	151.0	豆
2F07	8.0	亠	2F4F	80.0	毋	2F97	152.0	豕
2F08	9.0	人	2F50	81.0	比	2F98	153.0	豸
2F09	10.0	儿	2F51	82.0	毛	2F99	154.0	貝

代码	序号	字形	代码	序号	字形	代码	序号	字形
2F0A	11.0	入	2F52	83.0	氏	2F9A	155.0	赤
2F0B	12.0	八	2F53	84.0	气	2F9B	156.0	走
2F0C	13.0	冂	2F54	85.0	水	2F9C	157.0	足
2F0D	14.0	冖	2F55	86.0	火	2F9D	158.0	身
2F0E	15.0	冫	2F56	87.0	爪	2F9E	159.0	車
2F0F	16.0	几	2F57	88.0	父	2F9F	160.0	辛
2F10	17.0	凵	2F58	89.0	爻	2FA0	161.0	辰
2F11	18.0	刀	2F59	90.0	爿	2FA1	162.0	辵
2F12	19.0	力	2F5A	91.0	片	2FA2	163.0	邑
2F13	20.0	勹	2F5B	92.0	牙	2FA3	164.0	酉
2F14	21.0	匕	2F5C	93.0	牛	2FA4	165.0	釆
2F15	22.0	匚	2F5D	94.0	犬	2FA5	166.0	里
2F16	23.0	匸	2F5E	95.0	玄	2FA6	167.0	金
2F17	24.0	十	2F5F	96.0	玉	2FA7	168.0	長
2F18	25.0	卜	2F60	97.0	瓜	2FA8	169.0	門
2F19	26.0	卩	2F61	98.0	瓦	2FA9	170.0	阜
2F1A	27.0	厂	2F62	99.0	甘	2FAA	171.0	隶
2F1B	28.0	厶	2F63	100.0	生	2FAB	172.0	隹
2F1C	29.0	又	2F64	101.0	用	2FAC	173.0	雨
2F1D	30.0	口	2F65	102.0	田	2FAD	174.0	青
2F1E	31.0	口	2F66	103.0	疋	2FAE	175.0	非
2F1F	32.0	土	2F67	104.0	疒	2FAF	176.0	面
2F20	33.0	士	2F68	105.0	癶	2FB0	177.0	革
2F21	34.0	夂	2F69	106.0	白	2FB1	178.0	韋
2F22	35.0	夊	2F6A	107.0	皮	2FB2	179.0	韭
2F23	36.0	夕	2F6B	108.0	皿	2FB3	180.0	音
2F24	37.0	大	2F6C	109.0	目	2FB4	181.0	頁
2F25	38.0	女	2F6D	110.0	矛	2FB5	182.0	風
2F26	39.0	子	2F6E	111.0	矢	2FB6	183.0	飛
2F27	40.0	宀	2F6F	112.0	石	2FB7	184.0	食
2F28	41.0	寸	2F70	113.0	示	2FB8	185.0	首
2F29	42.0	小	2F71	114.0	内	2FB9	186.0	香
2F2A	43.0	尢	2F72	115.0	禾	2FBA	187.0	馬
2F2B	44.0	尸	2F73	116.0	穴	2FBB	188.0	骨
2F2C	45.0	屮	2F74	117.0	立	2FBC	189.0	高
2F2D	46.0	山	2F75	118.0	竹	2FBD	190.0	髟
2F2E	47.0	巛	2F76	119.0	米	2FBE	191.0	鬥
2F2F	48.0	工	2F77	120.0	糸	2FBF	192.0	鬯
2F30	49.0	己	2F78	121.0	缶	2FC0	193.0	鬲
2F31	50.0	巾	2F79	122.0	网	2FC1	194.0	鬼
2F32	51.0	干	2F7A	123.0	羊	2FC2	195.0	魚
2F33	52.0	幺	2F7B	124.0	羽	2FC3	196.0	鳥
2F34	53.0	广	2F7C	125.0	老	2FC4	197.0	鹵
2F35	54.0	廴	2F7D	126.0	而	2FC5	198.0	鹿
2F36	55.0	廾	2F7E	127.0	耒	2FC6	199.0	麥
2F37	56.0	弋	2F7F	128.0	耳	2FC7	200.0	麻

代码	序号	字形	代码	序号	字形	代码	序号	字形
2F38	57.0	弓	2F80	129.0	聿	2FC8	201.0	黄
2F39	58.0	彐	2F81	130.0	肉	2FC9	202.0	黍
2F3A	59.0	彡	2F82	131.0	臣	2FCA	203.0	黑
2F3B	60.0	彳	2F83	132.0	自	2FCB	204.0	黹
2F3C	61.0	心	2F84	133.0	至	2FCC	205.0	黽
2F3D	62.0	戈	2F85	134.0	臼	2FCD	206.0	鼎
2F3E	63.0	戶	2F86	135.0	舌	2FCE	207.0	鼓
2F3F	64.0	手	2F87	136.0	舛	2FCF	208.0	鼠
2F40	65.0	支	2F88	137.0	舟	2FD0	209.0	鼻
2F41	66.0	攴	2F89	138.0	艮	2FD1	210.0	齊
2F42	67.0	文	2F8A	139.0	色	2FD2	211.0	齒
2F43	68.0	斗	2F8B	140.0	艸	2FD3	212.0	龍
2F44	69.0	斤	2F8C	141.0	虍	2FD4	213.0	龜
2F45	70.0	方	2F8D	142.0	虫	2FD5	214.0	龠
2F46	71.0	无	2F8E	143.0	血			
2F47	72.0	日	2F8F	144.0	行			

附表 2　简化字形的康熙部首序号和字形

非简化字形的序号	非简化字形	简化字形的序号	简化字形
90.0	爿	090.1	丬
120.0	糸	120.1	纟
147.0	見	147.1	见
149.0	言	149.1	讠
154.0	貝	154.1	贝
159.0	車	159.1	车
167.0	金	167.1	钅
168.0	長	168.1	长
169.0	門	169.1	门
178.0	韋	178.1	韦
181.0	頁	181.1	页
182.0	風	182.1	风
183.0	飛	183.1	飞
184.0	食	184.1	饣
187.0	馬	187.1	马
195.0	魚	195.1	鱼
196.0	鳥	196.1	鸟
197.0	鹵	197.1	卤
199.0	麥	199.1	麦
205.0	黽	205.1	黾
210.0	齊	210.1	齐
211.0	齒	211.1	齿
212.0	龍	212.1	龙
213.0	龜	213.1	龟

2）部首外笔画数。

3）部首外起笔标记。按照笔形，顺序为横（一）、直（丨）、撇（丿）、点（丶）折（乙，包括𠂆乚等笔形），对应数字分别为 1、2、3、4、5。例如：“部”的部首外起笔为“丶”，标记为“4”；“侯”的部首外起笔为“𠂆”，标记为“5”。特殊地，当整个字即为部首本身或其变体，标记为“0”。

4）总笔画数。

5）表意文字描述序列（IDS）。通过组合使用表意文字描述符和已经编码的汉字，对待编码汉字的字形进行准确或近似的描述。这种组合，称为表意文字表述序列（IDS）。需要使用的表意文字描述符及其在 GB/T 13000 中的代码见附表 3。

附表 3 表意文字描述符

序号	表意文字描述符	代码	功能描述
1	⿳	303E	相似而不等，用于借某字表达外字。
2	⿰	2FF0	左右结构
3	⿱	2FF1	上下结构
4	⿲	2FF2	左中右结构
5	⿳	2FF3	上中下结构
6	⿴	2FF4	全包围结构
7	⿵	2FF5	向下包围结构
8	⿶	2FF6	向上包围结构
9	⿷	2FF7	向右包围结构
10	⿸	2FF8	向右下包围结构
11	⿹	2FF9	向左下包围结构
12	⿺	2FFA	向右上包围结构
13	⿻	2FFB	嵌套结构
14	⿼	2FFC	向左包围结构
15	⿽	2FFD	向左上包围结构
16	⿾	2FFE	左右翻转结构
17	⿿	2FFF	上下颠倒结构
18	⿳	31EF	笔画减少结构

6）读音：优先提供汉语普通话读音，以汉语拼音表示。若有其他汉语方言或民族语言及其方言的读音也可一并提供，以该方言或民族语言及其方言最常见的拉丁化方式或 IPA 表示。

7）临时替代用字：在汉字信息处理应用中代替本字使用的汉字，如同音字、近似音字或者形近字。

8）来源。若出自书面文献材料，则应提供该字所在页的完整页面与该文献

清晰的文献信息。

9）提交者。

注：1-4 部分以《康熙字典》的规则为准，部分字可允许使用两套部首和部外笔画等信息。

五、申请资料样例

为帮助读者理解本技术要求，下面给出 2 个待编码汉字申请资料的实例。包括相关的字形图片、汉字属性信息和使用情况的图片证据。

序号	部首序号	部首外笔画数	部首外起笔	总笔画数	IDS	已编码的近似字		读音	临时替代用字	来源	提交者
						字形	代码				
1	32.0	5	3	8	⿰钅土			tuó	沱	江苏省宜兴县地名录	中华字库工程
2	167.1	10	3	15	⿰钅荣			héng	鑅	中国文学编年史·清前中期卷上	中华字库工程



标准名称	汉 语 拼 音	备　　注
乐沱渚	Lètuózhǔ	原地形似骆驼，得名骆驼墅，后人称望墅垄，现改为乐沱渚。乐沱渚大队驻地。
朱家圩	Zhūjiāwéi	
万渔大队	Wànyú Dàduì	原名万圩大队，1981年更名。
湾　里	Wānlǐ	村后有一大湾。万渔大队驻地。
渔庠圩	Yúhúwéi	地势低洼多鱼。
木乃桥	Mùnǎiqiáo	原有一独木桥，惯称“没奈桥”（木乃、没奈同音）。
新舍大队	Xīnshè Dàduì	取新街和舍顶圩二村首字命名。
舍顶圩	Shèdǐngwéi	在圩顶头建舍取名。新舍大队驻地。
祠山殿	Císhāndiàn	村旁有一祠山庙得名。又叫新街上。
洋埝大队	Yángniàn Dàduì	以洋埝村命名。
洋　埝	Yángniàn	村后是马公荡，一片汪洋，又是雁常栖之处，故名（埝、雁同音）。洋埝大队驻地。
大白圩	Dàbáiwéi	地势低洼，常受水灾，惯称大白圩。

范道公社（125）

徐锴庆（1758—1802）生。鎌庆初名嵩，字朗斋，金匮人。乾隆五十一年举人。历官湖北武昌通判，黄梅、崇阳知县，蕲州知州。著有《玉山阁古文选》四卷、《诗选》八卷。事迹见王芑孙《署湖北蕲州知州徐君墓志铭》（《惕甫未定稿》卷一三）、《清史列传》杨芳灿传附。