ICS 01.140
A 14

WH

中华人民共和国文化行业标准

WH/T 91—2020

汉文古籍集外字描述规范

Guidelines for description of Gaiji

2020-09-01 发布　　2021-01-01 实施

中华人民共和国文化和旅游部　　发 布

目　　次

前　言

本标准按照 GB/T 1.1—2009 给出的规则起草。

请注意本标准的某些内容可能涉及专利。本文件的发布机构不承担识别这些专利的责任。

本标准由中华人民共和国文化和旅游部提出。

本标准由全国图书馆标准化技术委员会(SAC/TC 389)归口。

本标准起草单位:国家图书馆、天津图书馆、北京汉王数字科技有限公司。

本标准主要起草人:白帆、陈红彦、张毅、王昭、杜立功、赵依澍、周升川、肖禹、谢冬荣、萨仁高娃、李国庆、李志峰、潘慧敏、江世盛、刘正珍、王晓健、王战波。

汉文古籍集外字描述规范

1 范围

本标准规范了汉文古籍集外字描述的基本原则、集外字拆分流程和描述数据结构。

本标准适用于汉文古籍数字化过程中对汉字集外字描述。

2 术语和定义

下列术语和定义适用于本文件。

2.1

古籍 ancient Chinese books

主要指1911年以前(含1911年)在中国书写或印刷的书籍。

[GB/T 3792.7—2008,定义3.1]

2.2

字符集 character set

按一定的规则确定的包含汉字及有关基本图形字符的有序集合。

2.3

集外字 Gaiji

特定的字符集以外的汉字。

注：Gaiji为日语“外字”一词发音的英文转写。

2.4

表意文字描述符 ideographic description characters；IDC

描述文字各部件结构关系的符号,简称描述符。

2.5

表意文字描述序列 ideographic description sequence；IDS

表意文字描述符(2.4)和编码字符对表意文字字符作抽象描述的序列。

3 集外字描述基本原则

3.1 客观性

描述时应以保持字形原貌为基本前提。

3.2 唯一性

每一个被描述的集外字应具有唯一的描述方式。

3.3 可扩展性

在进行集外字描述时可依据项目具体需求进行扩展。

4 表意文字描述序列(IDS)语法规则

4.1 表意文字描述符(IDC)

IDC 共有 12 个,在 Unicode 中的编码为 U+2FF0 至 U+2FFB,其对应 Unicode 编码、描述和例字见表 1。

表 1 IDC 编码表

序号	Unicode 编码	IDC	IDC 描述	例字
1	U+2FF0	⿰	左右结构	離潔
2	U+2FF1	⿱	上下结构	粵弊
3	U+2FF2	⿲	左中右结构	弼欐
4	U+2FF3	⿳	上中下结构	曼勞
5	U+2FF4	⿴	全包围结构	圍圖
6	U+2FF5	⿵	上三包围结构	闢聞
7	U+2FF6	⿶	下三包围结构	凼函
8	U+2FF7	⿷	左三包围结构	匱匡
9	U+2FF8	⿸	左上包围结构	屑厝
10	U+2FF9	⿹	右上包围结构	匐忒
11	U+2FFA	⿺	左下包围结构	鴯題
12	U+2FFB	⿻	交叉结构	爽彧

4.2 表意文字描述序列(IDS)语法

4.2.1 IDS 由 IDC 和表意文字、部首、CJK 笔画、用户自定义编码和“?”(该字符编码为“U+FF1F”)组成。以下将表意文字、部首、CJK 笔画、用户自定义编码和“?”统称为部件。

4.2.2 IDC 包括二元操作符 10 个:⿰(U+2FF0)、⿱(U+2FF1)、⿴(U+2FF4)、⿵(U+2FF5)、⿶(U+2FF6)、⿷(U+2FF7)、⿸(U+2FF8)、⿹(U+2FF9)、⿺(U+2FFA)、⿻(U+2FFB)和三元操作符 2 个:⿲(U+2FF2)、⿳(U+2FF3)。

4.2.3 二元操作符后的部件限定为 2 个,三元操作符后的部件限定为 3 个。

4.2.4 IDS 支持嵌套机制,即:任意一个 IDS 可根据文字层次结构嵌入另一个 IDS 之中。

5 表意文字描述序列(IDS)的扩展

5.1 在 Unicode 中,除 12 个有编码的表意文字描述符(IDC)外,另有 4 个未编码 IDC,分别为:⿾(独体结构)⿽(右三包围结构)⿼(左右对角结构)⿿(右左对角结构)。在描述时可依据需要使用上述四个 IDC。为保证描述结果统一,在使用上述四个 IDC 时应优先采用以下编码:⿾(独体结构)编码为 FFFF0,⿽(右三包围结构)编码为 FFFF1,⿼(左右对角结构)编码为 FFFF2,⿿(右左对角结构)编码为 FFFF3。

5.2 为加强可操作性,保证描述结果的统一,在描述时应尽量使用已有的 IDS 进行表示,尽可能减少扩展。

5.3 必须进行扩展时，只能对 IDC、汉字部件和描述方式进行扩展，且扩展应符合 IDS 的基本规则。

5.4 扩展时，不应与现存的 IDS 产生冲突。

6 集外字拆分原则

IDS 的构建原则与应用需求直接相关。在具体操作中，项目可依据自身需求选择以字源为依据进行拆分或以字形为依据进行拆分，也可选择两者兼顾进行拆分。但在项目指定的范围内，拆分依据应统一。

7 集外字拆分流程

集外字拆分应依次按照以下流程进行：

1) 首先考虑集外字是否能被拆分为两个部件。如果集外字能拆分为两个部件，则按顺序优先选择二元操作符⿰(U+2FF0)、⿱(U+2FF1)、⿴(U+2FF4)、⿵(U+2FF5)、⿶(U+2FF6)、⿷(U+2FF7)、⿸(U+2FF8)、⿹(U+2FF9)、⿺(U+2FFA)或⿻(U+2FFB)来拆分。

   如果一个集外字既能按左右结构拆分，又能按上下结构拆分，应统一使用⿱(U+2FF1)进行拆分。

2) 如果集外字确实无法被拆分为两个部件，则考虑其是否能够被拆分为 3 个部件。如集外字能够被拆分为 3 个部件且符合上中下结构或左中右结构，则选择三元操作符⿲(U+2FF2)或⿳(U+2FF3)来拆分。

3) 如果集外字能够被拆分为 3 个部件，但不符合上中下结构或左中右结构，可以同时使用两个 IDC 进行拆分。

4) 如果集外字不能拆分为 3 个部件，而只能被拆分为 4 个及以上部件，可以同时使用 3 个以上 IDC 进行拆分。

5) 如集外字字形或结构独特，通过以上流程的处理仍无法依照 IDS 规则进行拆分，则放弃拆分。

8 汉字集外字描述数据的结构

汉字集外字描述数据中字段组成及字段著录要求见表 2。

表 2 汉字集外字描述数据必备字段

字段名	著录要求	字段属性
集外字 ID	著录集外字的序号。在同一项目内，集外字 ID 应有统一的命名规则，且每个集外字的 ID 号应唯一	必备字段
集外字位置	著录集外字在文献中所处的位置。在同一项目内，集外字位置的著录方式与格式应统一	必备字段
集外字字形	著录集外字的字形。可以为字形或图片	必备字段
集外字 IDS	著录集外字的 IDS。集外字 IDS 的构建应遵循本标准规定的拆分原则。如集外字为独体字或无法拆分的字，则本栏著录为“X”	必备字段

参 考 文 献

[1] 汉语大字典[M].成都:四川辞书出版社,2010.

[2] 通用规范汉字表[M].北京:语文出版社,2013.

[3] 蒋贤春.翟喜奎主编.中文文献全文版式还原与全文输入 XML 规范和应用指南[M].北京:国家图书馆出版社,2010.

[4] Ideographic Description Characters 表意文字描述符