ICS 01.140
A 14

WH

中华人民共和国文化行业标准

WH/T 90—2020

汉文古籍文字认同描述规范

Unity description for Chinese character identification

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、汉王科技股份有限公司。

本标准主要起草人:王昭、陈红彦、谢冬荣、萨仁高娃、李国庆、潘慧敏、肖禹、张毅、白帆、杜立功、赵依澍、江世盛、孟晓静、王战波。

汉文古籍文字认同描述规范

1 范围

本标准规定了汉文古籍文字认同描述的元数据、文字认同规则描述以及文字认同实例描述的内容、结构及各要素的描述规则。

本标准适用于图书馆及相关机构开展汉文古籍数字化工作中对文字认同过程和结果进行描述。民国时期文献的文字认同可参考执行。

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

汉字 Chinese character

记录汉语的书写符号系统。汉字也被其他一些国家或民族用作为书写符号。

[GB/T 12200.1—90,定义 4.1.3.6]

2.3

字体 character style

同一**汉字**(2.2)由于各种原因(历史演变、书写、印刷等)而形成的各种不同体式。

例:印刷体(宋体、仿宋体、黑体、楷体等)、手写体、古体、繁体、简体、正体、异体、俗体、讹体等。

[GB/T 12200.2—94,定义 4.1.2.2]

2.4

文字 script

人类记录和传达语言的书写符号系统。

[GB/T 12200.1—90,定义 4.1.2.7]

2.5

字音 character pronunciation

字的读音。**汉字**(2.2)中有的一字一音,有的一字多音。

[GB/T 12200.2—94,定义 4.1.2.31]

2.6

语义 semantics

词或词组与它们的含义之间的关系。

[GB/T 12200.1—90,定义 4.1.2.12]

2.7

文字认同 Chinese character identification

同一文字的不同字体转换为同一字体的过程。

3 文字认同描述的基本原则

3.1 客观性原则

文字认同描述应符合文字的客观属性。

3.2 一致性原则

指定范围内(如同一项目等)文字认同描述方式应一致。

3.3 适用性原则

应根据项目实际需要,选取必要的、实用的要素进行文字认同描述。

3.4 灵活性原则

可依据需求选取本标准中部分内容进行文字认同描述。

3.5 可扩展性原则

文字认同描述可依据需求进行扩展。

4 文字认同描述数据

4.1 概述

文字认同描述数据用于汉文古籍数字化中文字认同的描述,由文字认同描述元数据、文字认同规则描述数据(包括文字认同规则数据和文字认同规则适用范围描述数据)和文字认同实例描述数据3部分组成。

4.2 文字认同描述元数据

4.2.1 文字认同描述元数据概述

文字认同描述元数据是定义和描述汉文古籍数字化中文字认同描述数据的数据,是对文字认同描述的总体性说明。

4.2.2 文字认同描述元数据元素

文字认同描述元数据元素包括:文字认同描述说明、文字认同描述范围、文字字符集、文字描述方式、文字位置描述、文字认同依据、文字认同规则说明、文字认同描述数据说明。可根据需要扩展元素。文字认同描述数据元数据的元素组成及其定义见表1。

表 1　文字认同描述元数据元素描述

元素名	英文对应词	定义	注释
文字认同描述说明	Description of identification	说明古籍数字化过程中文字认同描述情况	说明文字认同描述的目的、处理等情况，可用于数据交换与共享
文字认同描述范围	Range of identification	说明文字认同描述适用的范围	在指定范围内文字认同描述方式相同，如项目、册、卷、叶等
文字字符集	Character set	定义文字的字符集范围	自行确定字符集的范围，如 Unicode 字符基础集、通用规范汉字表等
文字描述方式	Character description	说明认同前和认同后文字的描述方式	通过文字描述可以识别、检索或匹配文字，如 Unicode 编码、集外字使用 IDS 描述等
文字位置描述	Location description	说明文字在文献中位置的描述方式	通过该描述可以定位到文字，自行确定描述方式，如项目_书号_册_叶_行_列、项目_书号_册_叶_坐标等
文字认同依据	Basis of identification	说明文字认同依据的规范或工具书	可自行确定，但规则间不能相互冲突
文字认同规则说明	Identification rules	文字认同规则相关的说明	文字认同规则的总体性说明，可用于文字认同数据交换与共享
文字认同描述数据说明	Identification data	文字认同描述数据相关的说明	文字认同描述数据中结构、内容等相关说明

4.3　文字认同规则描述数据

4.3.1　文字认同规则描述数据概述

文字认同规则描述数据是文字认同描述数据的组成部分，是对汉文古籍数字化中文字认同规则及适用范围的描述，由文字认同规则数据和文字认同规则适用范围描述数据两部分组成，可依据实际需求进行扩展。

4.3.2　文字认同规则数据

4.3.2.1　文字认同规则数据概述

文字认同规则数据是对汉文古籍数字化中文字认同所依据的认同规则进行描述，认同规则之间不能相互冲突。

4.3.2.2　文字认同规则数据字段

文字认同规则数据字段包括：规则 ID、认同前的文字、认同前的文字描述、认同后的文字、认同后的文字描述、认同条件、认同依据、操作方式、文字认同规则数据版本号、备注。可根据需要扩展著录内容。描述文字认同规则所需数据字段组成及其说明见表 2。

表 2　文字认同规则数据字段描述

字段名	字段说明	注释
规则 ID	文字认同规则数据的序号	指定范围内，编码方式一致且唯一
认同前的文字	著录认同前的文字	超出 4.2.2 中“文字字符集”范围的文字
认同前的文字描述	描述认同前的文字	依据 4.2.2 中“文字描述方式”著录，如存储文字的图、Unicode 编码、IDS 描述等
认同后的文字	著录认同后的文字	依据 4.2.2 中“文字认同依据”生成的文字，同一文字的不同字体认同结果唯一
认同后的文字描述	描述认同后的文字	依据 4.2.2 中“文字描述方式”著录，如 Unicode 编码等
认同条件	描述文字认同成立的限定性条件	自行确定著录格式，如无、字音、语义、词汇等
认同依据	著录文字认同依据的规范或相关工具书名称	依据 4.2.2 中“文字认同依据”确定，如通用规范汉字表(2013 版)、汉语大字典(第二版)、第一批异体字整理表等
操作方式	描述文字认同的处理方式	自行确定著录格式，如批处理、人工处理等
文字认同规则数据版本号	著录文字认同规则数据版本号	根据修改情况更新版本号
备注	其他相关说明	与文字认同规则数据有关的说明、备注

4.3.3　文字认同规则适用范围描述数据

4.3.3.1　文字认同规则适用范围描述数据概述

文字认同规则适用范围描述数据用于定义和描述文字认同规则适用的范围和对象，可根据需要自行确定适用范围。

4.3.3.2　文字认同规则适用范围描述数据字段

文字认同规则适用范围描述数据字段包括：适用范围 ID、适用内容、适用范围、适用规则、文字认同规则数据版本号、备注。可根据需要扩展著录内容。文字认同规则适用范围描述数据所需的字段组成及其说明见表 3。

表 3　文字认同规则适用范围描述数据字段描述

字段名	字段说明	注释
适用范围 ID	文字认同规则适用范围描述数据的序号	指定范围内，编码方式一致且唯一
适用内容	描述文字认同规则在文献中适用的对象	自行确定著录格式，如人名、地名、残字、题字等
适用范围	描述文字认同规则在文献中适用的范围	自行确定著录格式，如第 1-10 叶、卷五等
适用规则	著录适用的文字认同规则	著录 4.3.2.2 中“规则 ID”或集合，自行确定数据格式
文字认同规则数据版本号	著录文字认同依据的规则数据版本号	客观著录，格式同 4.3.2.2 中“文字认同规则数据版本号”
备注	其他相关说明	与文字认同规则适用范围描述数据有关的说明、备注

4.4 文字认同实例描述数据

4.4.1 文字认同实例描述数据概述

文字认同实例描述数据是文字认同描述数据的组成部分，是对汉文古籍数字化过程中文字认同的过程和结果进行描述，认同后的文字是依据文字认同规则描述数据（文字认同规则数据和文字认同规则适用范围描述数据）生成的，指定范围内文字认同结果唯一。

4.4.2 文字认同实例描述数据字段

文字认同实例描述数据字段包括：文字 ID、文字位置、认同前文字、认同前文字_描述、认同后文字、认同后文字_描述、规则 ID、适用范围 ID、文字认同规则数据版本号、备注。可根据需要扩展描述要素。描述文字认同实例数据字段组成及其说明见表 4。

表 4 文字认同实例描述数据字段描述

字段名	字段说明	注释
文字 ID	文字认同实例描述数据的序号	指定范围内，编码方式一致且唯一
文字位置	描述文字在文献中的位置	依据 4.2.2“文字位置描述”著录
认同前文字	著录认同前的文字	超出 4.2.2 中“文字字符集”范围的文字
认同前文字_描述	描述认同前的文字	依据 4.2.2 中“文字描述方式”著录，如存储文字的图、Unicode 编码、IDS 描述等
认同后文字	著录认同后的文字	依据文字认同规则描述数据生成的文字，同一文字的不同字体认同结果唯一
认同后文字_描述	描述认同后的文字	依据 4.2.2 中“文字描述方式”著录，如 Unicode 编码等
规则 ID	著录文字认同依据的“规则 ID”	客观著录，格式同 4.3.2.2 中“规则 ID”
适用范围 ID	著录文字认同依据的“适用范围 ID”	客观著录，格式同 4.3.3.2 中“适用范围 ID”
文字认同规则数据版本号	著录文字认同依据的“文字认同规则数据版本号”	客观著录，格式同 4.3.2.2 中“文字认同规则数据版本号”
备注	其他相关说明	与文字认同实例描述数据有关的说明、备注

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] 国家语言文字规范和标准选编[M].北京:中国标准出版社,1999.