

数字方志文献外字基本属性研究

□ 翟喜奎 / 国家图书馆 北京 100081

摘要: 文章介绍了外字查重、外字属性编制、外字出处标引、研究成果应用等内容。研究成果已经作为国家提案向国际标准化组织ISO/IEC JTC1/SC2/WG2/IRG工作小组正式提交。

关键词: 数字图书馆, 中文信息处理, 汉字属性

DOI: 10.3772/j.issn.1673-2286.2010.07.009

1 引言

数字方志文献外字基本属性研究是2007年7月1日至2009年7月31日在国家图书馆立项的科研项目, 2009年4月17日项目通过了专家鉴定。国家图书馆从2004年开始对地方志文献进行全文数字化。一期处理地方志文献775种, 共524,732页, 平均每页404个汉字, 总计212,187,071个汉字。出现ISO10646-2003、UNICODE4.0字符集中没有涵盖的汉字共4866个(简称外字或字符集外字)。2005年, 二期处理地方志文献663种, 共400,349页, 平均每页403个汉字, 总计161,222,531个汉字, 又遇外字1104个。现在共有数字方志文献外字5970个。对数字方志文献外字的研究在国内外还是空白, 国际标准化组织ISO/IEC JTC1/SC2/WG2/IRG工作小组正在研究、收集字符集外字。

数字方志文献外字基本属性研究成果为IRG工作小组收集字符集外字提供了解决方案和处理依据,

最终目的是将研究成果列入国际标准字符集, 满足数字图书馆资源建设以及实际应用的需求。研究成果对国内外古籍、地方志文献的数字化建设都具有指导意义。

数字方志文献外字基本属性研究采用了调查法和文献法; 研究内容是数字方志文献出现的5970个外字; 目的是将外字列入到国际标准字符集。该研究首先对外字进行查重; 然后, 制作外字基本属性, 生成外字属性字表; 对外字出处进行标引, 制作外字出处字表和索引表。研制成果包含: 《数字方志文献外字基本属性研究研制报告》、《数字方志文献外字基本属性字表》、《数字方志文献外字结构描述基本属性字表》、《数字方志文献外字出处原文标引》、《数字方志文献外字出处索引表(外字编码顺序)》、《数字方志文献外字出处索引表(图像文件顺序)》。

《数字方志文献外字基本属性字表》、《数字方志文献外字结构描述基本属性字表》是以外字编码为序的文件, 为外字的聚类提

供汉字的基本属性信息; 《数字方志文献外字出处索引表(外字编码顺序)》是前两个成果的索引和检索工具。《数字方志文献外字出处原文标引》是以图像文件为序的文件, 标明外字在原文图像文件中的具体出处和位置; 《数字方志文献外字出处索引表(图像文件顺序)》是该成果的索引和检索工具。

2 外字查重

国家图书馆在“地方志文献”全文数字化过程中, 共补造字符集外字5970个。为了保证外字在ISO10646-2003、UNICODE4.0字符集中没有被涵盖, 需要进行查重。除此之外, 还要依据ISO/IEC JTC1/SC2/WG2/IRG工作组提出的扩C1 4.1版进行了查重。

数字方志文献5970个外字依据ISO10646-2003、UNICODE4.0和扩C1 4.1版字符集(共9万多汉字)进行查重。经过去重, 最后确定数字方志文献外字为4952个。其中

一期外字为3981个（去重885个汉字）；二期外字为971个（去重233个汉字）。去重率为17%。

3 外字属性编制

外字属性编制包括数字方志文献外字基本属性字表编制和数字方志文献外字结构描述属性字表编制两个内容。共有两个研究成果：

《数字方志文献外字基本属性字表》和《数字方志文献外字结构描述基本属性字表》。《数字方志文献外字基本属性字表》标明外字的基本属性；《数字方志文献外字结构描述基本属性字表》标明外字的结构描述的基本属性。

3.1 外字基本属性字表编制

数字方志文献外字基本属性字表收录了字形、《康熙字典》位置标引信息、康熙部首、部首外笔画数、部首外起笔、外字编码和《汉语大字典》位置标引信息，为外字的聚类提供汉字的基本属性信息。

编制内容：

（1）字形（Glyph）：表意文字的具体字型。

（2）康熙字典位置（kx index）：是汉字在《康熙字典》中的位置标引信息。其格式为xxxx.xx1，小数点前四位表示康熙页码，小数点后两位表示字位，最后一位是1，表示康熙字典无此字；如果最后一位是0，表示康熙字典有此字。

示例：𠄎0950.041，表示该字按照康熙字典的排序，应在第950页第4个字后面（即康熙部首的“网”部，因为该字部首外笔画是

13画，所以排在“网”部13画字的最后一个字之后，即950页第4个字之后）。在康熙字典中没有此字，所以最后一位是1。

（3）康熙部首（Radical）：代码按照国际标准ISO/IEC 10646的规定，如7F51表示康熙部首的“网”部。

（4）部首外笔画数（Stroke count）：除部首以外的笔画总数。

（5）部首外起笔（First stroke）：除部首以外的第一笔笔形，横为1；竖为2；撇为3；点为4；折为5。

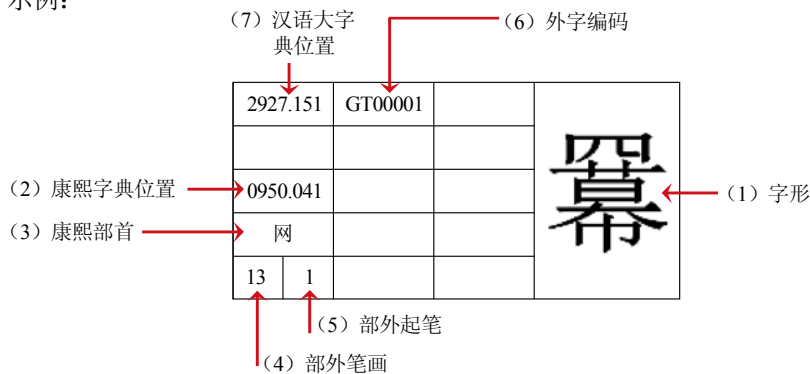
（6）外字编码（Glyph Code）：是汉字来源与编号，起止码为GT00001—GT04952（GT表示国图，00001—03981是数字方志一期发现的系统外字序号；

03981—04952是数字方志二期发现的系统外字序号）。

（7）汉语大字典位置（HanyuDaZidian index）：是汉字在《汉语大字典》中的位置标引信息。格式为xxxx.xx1。小数点前四位表示汉语大字典页码；小数点后两位表示字位；最后一位：1表示汉语大字典无此字，0表示汉语大字典有此字。

示例：𠄎2927.151，表示该字按照汉语大字典的排序，应在第2927页第15个字后面（即汉语大字典部首的“网”部，因为该字部首外笔画是13画，所以排在“网”部13画字的最后一个字之后，即2927页第15个字之后）。在汉语大字典中没有此字，所以最后一位是1。

示例：



3.2 外字汉字结构描述属性编制

数字方志文献外字结构描述属性字表收录了字形、外字编码、汉字结构描述、图像文件位置、图像文件名，为外字的聚类提供汉字的基本属性信息。

编制内容：

（1）字形（Glyph）。

（2）外字编码（Glyph

Code）。

（3）汉字结构描述（IDS）：汉字结构描述是依据汉字的结构特征用熟悉的汉字或部件描述一个完整的汉字。

IDS的示例如表1。

（4）图像文件位置：对外字在原文中的位置进行标引。前4位为图书种号；接下来3位为图书卷号；再接下来3位为具体页号；最后1位或2位为外字的具体行号。

表1 IDS的示例表

字符名称	DC数目	DC的相关位置	IDS示例	IDS示例显现
左右结构	2			母
上下结构	2			余
左中右结构	3			衍
上中下结构	3			咎
全包围结构	2			囧
从上包围结构	2			閑
从下包围结构	2			出
从左包围结构	2			匜
从左上包围结构	2			廌
从右上包围结构	2			匉
从左下包围结构	2			逄
相嵌结构	2			巫

图像文件位置：图书种号-卷号-页号-外字所在行号。

(5) 图像文件名 (Image File Name)：指明具体的外字出处位置信息；是外字所在位置的具体页图像。

图像文件名由10位数字和文件扩展名jpg组成。

具体形式为：xxxxxxxxxx.jpg。

10位数字具体再分：前4位为图书种号；中间3位为图书卷号；最后3位为具体页号。

图像文件的组成是：图书种号+卷号+页号+文件扩展名。

4 外字出处标引

外字出处标引包括数字方志文献外字出处原文标引、数字方志文献外字出处书目标引和外字出处索引编制三个内容。共有三个成果：

《数字方志文献外字出处原文标引》、《数字方志文献外字出处索引表（外字编码顺序）》、《数字方志文献外字出处索引表（图像文件顺序）》。《数字方志文献外字出处原文标引》标明外字在原文图像文件中的具体出处和位置；《数字方志文献外字出处索引表（外字编码顺序）》标明文献的具体书

目、题名、索书号和原文图像文件等出处信息，是外字编码顺序的检索工具；《数字方志文献外字出处索引表（图像文件顺序）》是图像文件顺序的检索工具。

4.1 外字出处原文标引

数字方志文献外字出处原文标引，指出了外字在图像文件中原文的具体位置和出处。数字方志文献外字出处原文标引收录了字形、外字编码、EUDC(16)、EUDC(10)、图像文件名、行号、4415幅原文图像文件等信息，在图像文件中用红圈将外字进行标引，为外字提供了原文出处信息。

标引内容：

- (1) 字形 (Glyph)。
- (2) 外字编码 (GlyphCode)。
- (3) EUDC(16)：外字十六进制编码，编码范围：E000—FFFF。
- (4) EUDC(10)：外字十进制编码，编码范围：57344—65535。
- (5) 图像文件名 (Image File Name)。

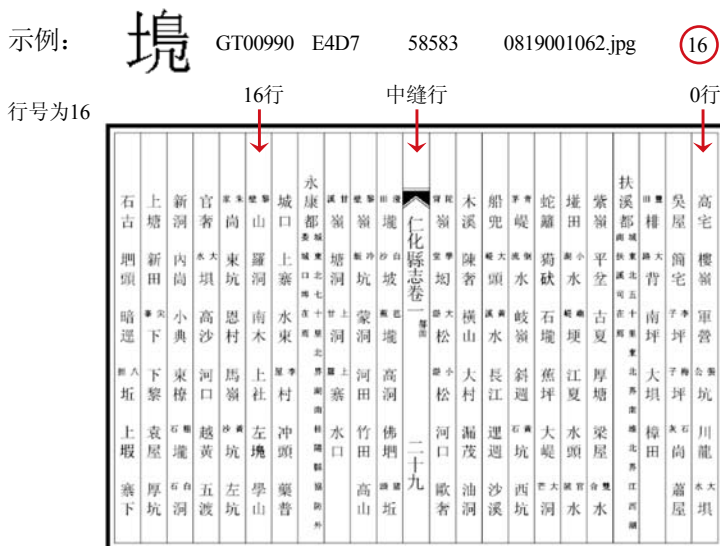
(6) 行号 (Column Number)：指示外字所在的具体竖行位置。

以稿纸竖行为一行，一行中可能包括多个小行（只算一行）。行是从右向左数，第一行为0，依次数起（中缝行不算）。

(7) 在4415幅原文图像文件中用红圈将外字进行标引。

4.2 外字出处书目标引

数字方志文献外字出处书目标引，指出了外字在国家图书馆书目系统中的MARC记录的具体位置和出处。数字方志文献外字出处书目标引收录了字形、外字编码、书目



记录系统号、索书号、题名、题名汉语拼音等信息，为外字提供了汉字原文所在书目位置和出处信息。

标引内容:

- (1) 字形 (Glyph)。
- (2) 外字编码 (Glyph Code)。
- (3) 书目记录系统号 (System Number): 在国家图书馆自动化系统中是书目记录的唯一编号。用此号可以直接调出书目记录。

(4) 索书号 (Call Number): 在图书馆查找图书的号码。

(5) 题名 (Title): 外字出处的具体书名。

(6) 题名汉语拼音 (Title Pin Yin): 题名的汉语拼音。

4.3 外字出处索引编制

数字方志文献外字出处索引表 (外字编码顺序和图像文件顺序) 指出外字在图像文件中原文的具体位置和出处以及在书目数据 (MARC) 中的具体位置和出处。数字方志文献外字出处索引表为外字编码顺序和图像文件顺序的文件提供检索工具。

5 研究成果应用

数字方志文献外字基本属性研究成果: 《数字方志文献外字基本属性字表》、《数字方志文献外

字出处索引表》在2007年ISO/IEC JTC1/SC2/WG2/IRG 第29次工作小组会议上由国家图书馆委托中国电子技术标准化研究所正式提交。

网址为: <http://appsrv.cse.cuhk.edu.hk/~irg/irg/irg29/IRG29.htm>

文件名为: IRGN1366, 具体文件名是IRGN1366中IRGN1366_Appendix1.pdf和IRGN1366_Appendix2.pdf。

《数字方志文献外字结构描述基本属性字表》在2008年ISO/IEC JTC1/SC2/WG2/IRG 第30次工作小组会议上由国家图书馆正式提交。

网址为: <http://appsrv.cse.cuhk.edu.hk/~irg/irg/irg30/IRG30.htm>

文件名为: IRGN1467 (AP1, AP2, AP3)。

研究成果应用于数字图书馆古籍全文数据库建设, 满足了数字图书馆资源建设以及实际应用的需求。研究成果对国内外古籍、地方志文献的数字化建设都具有指导意义。

数字方志文献外字基本属性研究虽然结束了, 但是还需要跟踪ISO/IEC JTC1/SC2/WG2/IRG 工作小组的动向, 最终目的是将研究成果列入国际标准字符集, 满足数字图书馆资源建设以及实际应用的需求。

参考文献

- [1] 汉语大字典编辑委员会. 汉语大字典[M]. 武汉: 湖北辞书出版社, 成都: 四川辞书出版社, 2006.
- [2] 张玉书, 等. 康熙字典[M]. 北京: 中华书局, 1958, 2006重印.

作者简介

翟喜奎, 国家图书馆副研究馆员, 主要从事图书馆自动化、中文信息处理、数字图书馆等技术工作。研究方向: 图书馆自动化、中文信息处理、数字图书馆技术研究。通讯地址: 北京中关村南大街33号 国家图书馆业务管理处 100081。E-mail: zhaixk@nlc.gov.cn

The Research of Waizi Basic Attributes in Digital Local Chronicles

Zhai Xikui / National Library of China, Beijing, 100081

Abstract: The paper introduced the duplicates finding, property formulation, reference indexing of Waizi, and research findings application. The research findings have been officially delivered to ISO/IEC JTC1/SC2/WG2/IRG working group as the national proposal.

Keywords: Digital library, Chinese information processing, Hanzi attributes

(收稿日期: 2009-10-02)