浅谈非专利文献加工中生僻字的标引方法

专利检索咨询中心　武芳　张雨春

摘要：在医药类非专利文献数据加工过程中，尤其是在标引药品名称、中药名称和疗效时，常会遇到一些生僻字。如果这些生僻字标引不得当，则可能会影响到数据库整体检索的准确性和全面性，因此，规范与统一标引这些生僻字对于非专利文献数据加工有着重要的意义。本文归纳总结了文献中常见的几类生僻字，并针对其特点提出了不同的标引方法和建议。

关键词：医药类非专利文献；数据加工；生僻字

武　芳

助理研究员，北京工业大学材料学专业，主要从事医药类非专利文献数据加工工作。

张雨春

助理研究员，北京化工大学发酵工程专业，主要从事医药类非专利文献数据加工及质量控制工作。

生僻字是相对而言的。在药品名称、中药名称和疗效等词汇中，一些对于日常生活而言的所谓生僻字比较多见。这些信息是现阶段非专利医药类文献加工的重点内容，如果这些重要信息中生僻字标引不得当，则会影响非专利医药类文献数据加工的质量，也势必影响检索结果。因此，规范与统一标引生僻字的标引有着重要的意义。

检索中国期刊全文数据库（CNKI）时发现，在搜索文章题目时，CNKI 数据库中用拆字法和空格法对其中的生僻字进行处理。如图 1：

吲(口艹两)酰胺双盲法治疗高血压病的临床疗效观察	赵连友	第四军医大学学报	1989/01
吲(口艹两)酰胺治疗高血压的观察	程小	云南医药	1988/03
吲■酰胺治疗老年及老年前期高血压病临床疗效观察	施志明	中国老年学杂志	1990/04
藤黄属植物中笼状多异戊烯基　酮类化合物的研究进展	王丽莉	天然产物研究与开发	2011/04

图 1 CNKI 常用的生僻字处理方法

然而这些处理方法不能表述化合物名称的准确信息，给识别和理解带来困难，也不能用于检索。

对于期刊文献正文中的生僻字，除了上述方法外，作者 / 编辑还常采用拼字法、手工补写或是利用 window 自带的造字功能进行处理。例如：对于化合物“1,2,8- 三甲氧基 -6- 羟基吲酮”[1] 的处理，由于吲酮 的吲字，通过常规的输入法无法输入，作者将“口”、“山”两个字放在一起，通过调整两个字的间距和字体大小，把两个字拼成一个字来表示，即拼字法；或是用 window 自带造字功能，把自造字存为图片格式，粘贴到字句中间。而葶蓂的“蓂”，则采用手工补写的方法处理，如“基本方：葶蓂 30g，薏苡仁 20g…”[2]。以上方法可以直观清楚地识别出化合物以及中药的名称等信息，虽然为我们在数据加工工作中提供了一些借鉴和启示的作用，但仍存在着一些弊端，毕竟检索的过程是字符的匹配，而非检索图片的内容，无论是拆拼字的造字法还是手工补写法，在检索过程中还是会对信息的准确性造成影响。

为此，本文首先归纳总结了医药类文献中常见的几类生僻字，进而根据其各自的特点，从检索的角度出发，提出了规范与统一标引生僻字的可行方法和建议。

生僻字的分类：

1. 有些在医药领域中比较专业的词汇，在日常生活中比较少见，CNKI 对这些词语常采用拼字或手工补写的方法进行处理，如下图中的“葶蓂”[3] 和“蟾蜍”[4]。这些词目前可以使用常规的输入法输入。

均给以自拟二花葶蓂汤治疗，其基本组成为鸡冠花 30g，荠菜花 30g，葶蓂 20g，益智仁 10g，山

疗，治疗组在以上基础上加自拟方飞天蟾蜍汤（飞天蟾蜍 12 g、法半夏 4 g、枇杷叶 4 g、橘络 3 g、蝉蜕 3

图 2 生僻字的拼字造字法和手工补写

2. 另外还有一些医药领域中较专业的词汇，通过常用的输入法是不能输入的。例如癌瘤[5]、吲酮[6]、地尔硫䓬[7]、大黄䗪虫丸[8]。

处理方法及对策

对于第 1 类生僻字的处理，我们在数据加工标引过程中可以采用以下对策：

（1）在不知道“蟾蜍”、“葶蓂”、“癥瘕”这些词语的读音的情况下，可以借助 handinput 等笔画输入软件（图 3）、或五笔输入法输入。

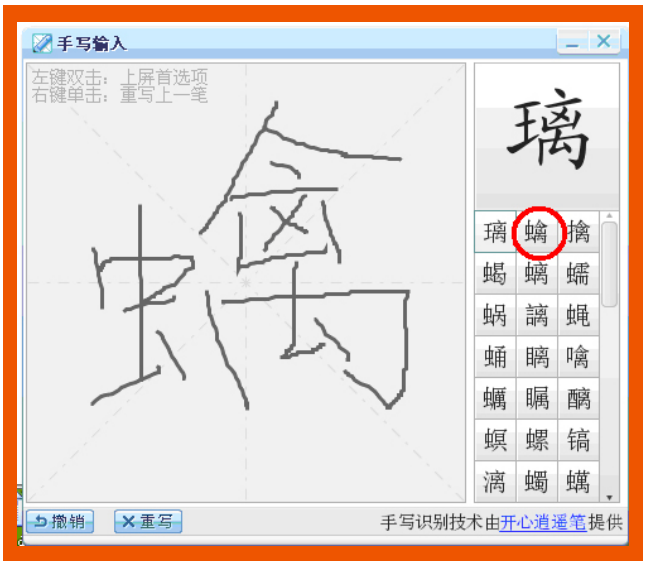

图 3 借助 handinput 笔画输入生僻字“蟾”

图 4 借助 baidu 搜索“薤”字

（2）借助 baidu 或 google 搜索引擎，上外网搜索该词语的读音，例如“薤”可以在 baidu 上搜索“草字头 解”来查找，然后通过常用输入法输入。

（3）借助现代汉语词典，通过“部首查字法”查找词语的读音，然后通过常用输入法输入。

对于第 2 类生僻字的分析：噩(è)、鼍(zhè)、斲(zhuó)等字是计算机常用汉字库没有收载的生僻字。这些生僻字往往通过常规的输入法打字不能解决问题，但如果采用手工补写，或是拼字法，都会出现用字不规范的现象，会影响到检索的准确性。因此有必要在标引过程中不断总结和积累，找到规范和统一这些生僻字的方法。为此提出以下标引建议：

（1）简体字和繁体字互换。可以用简体字替代繁体字，或是用去掉偏旁的字代替原字。如：月桂氮卓酮中用“卓”代替斲、用“山酮”代替𠰁酮。也可以用繁体字代替简体字。如：“櫼木”代替楔木；“牙齦痈”代替牙龈痈；磺胺甲噁唑中用“噁”代替噁。

（2）用已经公知的其他字代替。如用“癫痫”代替癲癎。虽然癎字有繁体字“癇”可以打出来，但是这是早期文献的书写方法，近年来的期刊文献中疾病名称一般都写为“癫痫”，而不用“癲癇”。

（3）用具有相同含义的形似字代替。大黄䗪虫丸中，处方中有“土鳖虫”这味药，据《辞海》记载，䗪虫，中药名，又称“土鳖”，可以直接用土鳖虫代替，或全拼输入法中有形似字“蟅”，通过 baidu 搜索“蟅”，该字也有“地鳖虫”的意思，建议用“蟅”代替䗪字。

（4）以上第 2 类生僻字中，中药名称和化合物名称在数据加工过程中，属于必须加工

的内容，因此需要规范与统一标引。但对于一些中医病名中的生僻字，有些情况下文章中给出了相对应的西医的疾病名称，此时应优先标引西医病名，可将中医病名中的生僻字归入非强制标引的内容。有些字现代汉语词典上有，但输入法打不出来，Baidu 搜索出来 copy 后变成一串代码，说明这个字属于扩充字库，例如**瘕瘤**的**瘕**(pēi)，中医指荨麻疹，此时若以形似字“瘕瘤”代替**瘕瘤**，并不合适，因为瘕（yīn）是喑的古字，其含义与**瘕**（pēi）的含义并不相同。在此情况下，如果文中有对该疾病名称的其他相应描述：“瘾疹即皮肤出现鲜红色或苍白色风团，时隐时现，故名。首见，于《内经素问》，此后历代文献有**瘕瘤**、风疹、赤白游风等名记载，俗称风疹块，现代医学称之为荨麻疹”[5]。则可以标引为“荨麻疹”。

总之，在医药期刊中还有不少生僻字，本文未能列举完全，但对于数据加工遇到生僻字，找到合适的处理方法，尽量准确反映原文给出的化合物名称、中药名称以及中医疗效等信息，是数据加工应该遵循的基本原则。因此，为保证非专利医药类文献加工信息的准确与规范，为了更好的服务于专利检索，我们应该尽可能规范与统一生僻字的标引。

（专利检索咨询中心　杨晓春　审校）

参考文献

1. 文荣荣；董秀华；段沅杏；李干鹏 獐牙菜的化学成分研究 云南民族大学学报（自然科学版）2010.02。

2. 刘运霞 葶藶渗湿汤治疗结节性红斑30例 河北中医 2001.06。

3. 高晓平 自拟二花葶藶汤治疗乳糜尿37例 甘肃中医 1997.03。

4. 吴巧燕；陈源 中西医结合治疗儿童喘息型支气管炎30例 江西中医药 2008.03。

5. 郭晓燕 瘾疹临床诊治体会 内蒙古中医药 1996.01。

6. 蒋富强；张雪梅；马云保；耿长安；江志勇；陈纪军 毛萼獐牙菜化学成分的研究 中国中药杂志 2011.16。

7. 周玉梅 地尔硫卓治疗慢性肺心病合并房性紊乱性心动过速疗效观察 中国社区医师（医学专业）2011.20。

8. 邹兰谷 大黄蟅虫丸的临床运用 江苏中医 1995.07。