网络出版时间：2018-12-05 09:15:27
网络出版地址：http://kns.cnki.net/kcms/detail/44.1306.G2.20181204.2346.004.html

人文数据库建设中人文学者何为

——以《全宋文》墓志铭亲属信息提取为例

陈佩辉

摘　要数据库在人文研究中发挥着越来越重要的作用，但也面临一些学者对其数据准确性和全面性的质疑。为了回应这一点，数据库建设中需要人文学者扮演更重要的角色。通过对哈佛大学中国历代人物传记资料库《全宋文》墓志铭亲属信息提取流程再造的考察，我们发现人文学者不仅在辨别文本信息上具有关键作用，由此保证数据准确性和全面性，而且对于促进新技术、新方法的应用方面也具有十分重要的作用，由此提高数据库建设的效率。因此在数据库建设中，人文学者既要承担文献辨析责任，也要积极地承担起技术责任。

关键词　人文数据库 人文学者 文献辨析 技术责任

引用本文格式　陈佩辉. 人文数据库建设中人文学者何为——以《全宋文》墓志铭亲属信息提取为例[J].图书馆论坛，2019

What the humanist can do in constructing Humanity database——Taking the extraction of kinship data from Epitaphs in *Quansongwen* as an Example

Chen Peihui

Abstract　Databases play a more and more important role in humanity studies, while some scholars question its accuracy and completeness. In order to deal with this problem, the database needs the humanists play a vital part in its construction process. After investigating the reengineering of the process of the extraction of kinship data from Epitaphs in *Quansongwen* in Chinese biographical database project at Harvard University, we find that the humanists not only could guarantee the accuracy and completeness of the historical data by analyzing the epitaphs exactly, but also could improve the efficiency of the database construction by facilitating the application of new technology and new methods. Thus, humanists not only have to bear the liability for the literature analysis, but also should be actively in shouldering the technical responsibility.

Keywords　Humanity database; Humanists; Literature analysis; Technical responsibility

人文资料电子数据库自诞生以来，受越来越多人文学者和爱好者好评的同时，也受到不少来自文献学、史学研究者的质疑与批评，这些质疑在数据库方面主要集中在数据的准确性与全面性上。王瑞来从版本学的角度对时下数据库中的一些书籍的版本选择提出批评，认为有些书目的版本选择有误，而这会给研究者带来很多误解，进而影响研究论文的质量[1]。包伟民以多年的深厚史学功底对数据库选取信息的整全性以及可行性提出质疑，认为提取数据时建设者的史学修养非常重要，如果不明了选取对象的多种表达方式就会出现各种漏选，基于数据库研究的准确性自然会令人质疑。此外，很多历史信息的关联会随着研究的不断深入而凸显，本来数据库是为了方便研究，而在甄别数据的过程中就已经完成了相关研究，那么是否还有必要建设数据库就成了新的问题。 两位学者都强调人文基本功的重要性，认为计算机不可能代替人文学者。其实这是对数字人文的一个常见误区，数字人文本质上依然是人文，计算机不可能替代人文研究者，除非人工智能可以像人一样思考。两位学者也并非完全拒斥数据库，而是保持一定的开放态度，包伟民认为面对数字人文新时代人文学者应该学会使用信息技术以促进史学的深入发展[2]。

面对数字人文对传统人文学的挑战，以两位学者为代表的人文学者实质上是基于历史文献信息的复杂性对数据库建设提出了挑战以及可能改进的方向。本文即是基于笔者在实际建设数据库过程中的经验，尝试探讨人文学者如何在数据库建设中发挥其主导作用①。

由此既可以解决来自数字人文时代对史学研究者的挑战，也可以回应历史文献研究者对数据库的质疑与挑战。

1 问题的发现

目前中国历代人物传记资料库（Chinese Biographical Database Project, 以下简称CBDB）在建设数据库时主要采用的方法是标记（tagging）即先利用正则表达式提取文献中的信息，再通过人工审核补充标记的方法进行的。在数据库建设中，人文学者负责审阅文本相关信息，找出古代汉语常见的表达规则，并在信息技术专家的指导下撰写简单的正则表达式，信息技术专家主要负责校正正则表达式和编程。在分工上，人文学者与信息技术专家基本是分开的，彼此不熟悉对方的专业领域。当研究对象的语言表述较为简单时，如此分工并不影响数据挖掘的质量，但当研究对象的语言表述较为复杂时，就很可能会影响信息提取的全面性和准确性。

笔者在做《全宋文》墓志铭亲属信息提取工作时就先按这一分工进行的。不过在工作过程发现了非常棘手的问题即计算机无法识别人名。与官名不同，人名所用的词汇量很大，无法仅仅根据词频与常见语法结构有效确定其为人名，在长辈亲属姓名提取中这个问题并不明显，因为有许多关键词如讳、曰、娶、氏和句尾标点等符号可以作为人名的提示信息，计算机可以根据这些特征提取人名。比如，墓志铭中常见的表达方式为“亲属关系+讳（曰）+名字”，那么相应的正则表达式(曾祖|祖父|祖等亲属关系)(諱|曰)([^，。]{0,5})[，。]就可非常简易地提取出相关信息。但是，在提取子孙等晚辈姓名时计算机无法识别人名的问题就会严重影响数据挖掘的准确性和效率，比如下面这则信息：

孫男二十人：長仲俶，右武衛大將軍、眉州刺史；次仲誘，右武衛大將軍、茂州刺史；次仲虺，右武衛大將軍、春州刺史；次仲螯，右監門衛大將軍；次仲沃、仲芮、仲雪、仲敔、仲靡，并右千牛衛將軍；次仲頎、仲霤、仲吟、仲疇、仲逢，并太子右監門率府率；次仲誥、仲諱、仲愷，并太子右內率府率；餘未命。[3]

无论用長、次还是分号“；”等作为提示词，都无法提取出所有的人名信息，其中至少一半的人名信息将被忽略，也就是说虽然发现了这些常见的表达结构，但是却无法运用简易的正则表达式准确的提取这些人名信息。

仔细审视这段文字所包含的信息，会发现大量的信息是关于地名和官名的信息，如果剔除或者替代这些信息后，那么人名是否就容易被识别出来呢？来看剔除了这些干扰信息之后的结果：

孫男二十人：長仲俶，、；次仲誘，、；次仲虺，、；次仲螯，；次仲沃、仲芮、仲雪、仲敔、仲靡，并；次仲頎、仲霤、仲吟、仲疇、仲逢，并；次仲誥、仲諱、仲愷，并；餘未命。

显然，这就比之前更容易被提取出来，如果再对个别词汇作进一步处理，就很容易提取出人名信息。

2 新方法的引入

上面这个例子说明人名信息虽然不是以有规律的方式呈现的，但人名周边的信息是以某种有规则的方式出现，如官名、地名、并、余。那么对有规则的表达就可以利用计算机的工具进行处理，就可以把段落中的官名地名信息替换掉，实质上与正则表达式提取信息有异曲同工之妙，前者反向凸显人名信息，先删除有规则表达然后再提取人名，而后者根据人名的有规则表达直接提取人名信息。然而，这种思路适合全部的墓志铭信息吗？

为验证这一方法是否对整个《全宋文》墓志铭有效，笔者进行随机抽样（抽取全体中的百分之十），并进行官名地名以及标点的替换。通过熟读关于孙男的墓志铭信息，得到“孫男.{0,2}人[，：][^女孫] .{0,50}”这一正则表达式[4]。在 visual studio code 中 run，在全部墓志铭中可以发现 787 個相关結果。检索到的信息有以下几种典型形式：

（1）孫男二人：長應運，登丙戌進士第，儒林郎、兩浙轉運司物料官，即亨之也；次應龍，

習舉子業[5]。

（2）孫男五人：汝直、汝敦、汝平、汝功、汝能、皆業進士[6]。

（3）孫男六人：曰夷仲，曰虞仲，曰於仲，曰南仲，曰武仲，曰延仲[7]。

（4）孫男二十人：長仲俶，右武衛大將軍、眉州刺史；次仲誘，右武衛大將軍、茂州刺史；次仲虺，右武衛大將軍、春州刺史；次仲鎣，右監門衛大將軍；次仲沃、仲芮、仲雪、仲敔、仲靡，并右千牛衛將軍；次仲頎、仲霤、仲吟、仲疇、仲逢，并太子右監門率府率；次仲誥、仲諱、仲愷，并太子右內率府率；餘未命。

显然，第三类是比较规则的句子，可以利用正则表达式直接提取，第一类和第三类经过后期程序编写也可以提取。第四类在之前例子中已经分析，是比较难以提取的，要经过比较复杂的编程才可以提取出人名。同一个正则表达式寻找到的结果在二次编程处理时却有着不同的难度，需要进行不同的处理。再看抽样的结果：

将新的正则表达式“孫男.{0,2}人([\w|\/]+([^孫女]{1,10})\/*)+”在 visual studio code 中 run,选择全字匹配和正则表达式，可以提取 87 個結果，约等于从全体中提取结果（787）的 10%，这与上面的正则表达式提取结果基本一致。不过再排除官名、地名后，剩余的信息为人名的可能性很高，数据的精确度得到了保证。上面四个句子在替换后转变为：

（1）孫男二人/wm/長應運/wsep/登丙戌/no_noc/第/wsep//no_noc//wsep/兩浙②/no_noc//wsep/即亨之也/wsep/次應龍/wsep/習舉子業

（2）孫男五人/wm/汝直/wsep/汝敦/wsep/汝平/wsep/汝功/wsep/汝能/wsep/皆/vno//no_noc/

（3）孫男六人/wm/曰夷仲/wsep/曰虞仲/wsep/曰於仲/wsep/曰南仲/wsep/曰武仲/wsep/曰延仲

（4）孫男二十人/wm/長仲俶/wsep//no_noc//wsep//ns//no_noc//wsep/次仲誘/wsep//no_noc//wsep//ns//no_noc//wsep/次仲虺/wsep//no_noc//wsep//ns//no_noc//wsep/次仲鎣/wsep//no_noc//wsep/次仲沃/wsep/仲芮/wsep/仲雪/wsep/仲敔/wsep/仲靡/wsep/并/no_noc//wsep/次仲頎/wsep/仲霤/wsep/仲吟/wsep/仲疇/wsep/仲逢/wsep/并/no_noc//wsep/次仲誥/wsep/仲諱/wsep/仲愷/wsep/并/no_noc//wsep/餘未命

虽然也出现了不是人名的其他词汇如“皆”“曰”“長”“次”，但都比较规则，很容易编写程序进行排除，下文会进一步讨论这一问题。显然，将官名地名信息替代后再提取信息会减少后期编程的复杂性，使整个提取流程变得更加有效率。

3 新流程探索中人文学者的作用

从这一思路出发，笔者尝试改进数据库建设流程并将其再造为更能准确提取信息的、人文学者能充分发挥其作用的数据库建设流程（本文提及的所有文本、表格和详细的 Python 代码以及各种输出结果已在网上发布，参见：https://github.com/cbdb-project/CBDB_Laxmi/tree/master/quan_song_wen.）。在这一尝试中，重点探讨人文学者如何在流程改进中承担人文责任和一定的技术责任。

根据笔者对《全宋文》墓志铭的观察，几乎所有特定亲属关系的表述都在同一个句子中表达，这样就可以通过提取关键句子来确定信息提取的范围。因此新流程的第一步是句子压缩，将《全宋文》墓志铭中的含有亲属关系的句子提取出来。这一步是信息技术专家提出的，但由于文本标点有时不规范，特定亲属信息并不在同一个句子中，会造成信息漏选，然而后面人工审核查漏补缺的流程能够保证信息的全面性，所以这一步并不会带来难以解决的困难。在这一步需要做一个包含所有《全宋文》墓志铭中出现的亲属关系的亲属关系表，并清楚地界定各个亲属关系的含义，这就需要人文学者发挥主要作用，因为历史知识基础决定了关系表的准确性和完备性。基于 CBDB 原有的亲属关系表，进行了补充和编码，建立了新的关系表。除了“祖”“考”“子”“男”③等直接表示亲属关系的词汇外，还需要将“娶”“配”“嫁”“归”“适”等表示嫁娶的词汇收入其中，因为这些词间接指向女性亲属如祖妣、妻子或者男性亲属如女婿、孙婿等。但在处理“配”“归”“适”这些词汇时要消除歧义，

表示嫁娶只是它们众多含义中的一个，它们在墓志铭中可能以嫁娶之外的含义出现，比如“适”有去往某地之义。因此在选取含有这些词汇的句子时，需要加上其他条件比如句子中必须同时包含“氏”“夫人”“女”“女孙”等词汇中至少一个时才能选取。如果不加限制条件那么自动提取的结果错误率会非常高，也会增加很多审核工作，由此可见人文学者在确定句子信息压缩方面有着非常重要的作用。当然，在处理古代史信息时，不能为了提高数据的精确性而牺牲数据的选取数量，这一点与社会学、经济学在处理数据时不同，因为后者往往拥有巨量数据以至于只能抽样选取，少选取一定数量并不影响分析的准确性，而古代史的数据尤其是宋代之前保存下来的信息并不多，因此要尽可能全部选取。第二步，字典词汇替换。将句子中的官名和地名分别替代为 no_noc 和 ns。与第一步类似，对 CBDB 中宋代的官名表和地名表进行了修改与补充，创建了新的官名表和地名表。必须指出，如果没有宋代的官名表和地名表，这项工作就会很受影响，整个新流程也将会变得不如之前方便、高效，这显示出数据累积的重要性。在编辑官名表和地名表时，尽可能不要出现仅有一个字的官名或地名，比如“令”、“守”等有歧义的词汇。但有些含义比较单一的词汇可以保留，比如表示通判的“倅”。需要特别指出的，有些地名或官名与亲属关系名称用词相同，这部分官名地名也需要考虑是否删除。比如“长子”这个地名，如果作为地名全部替换就会出现将大量的作为亲属关系名称的长子被替换为 ns，必然会影响亲属信息提取的准确性。此外表示亲属关系的“庶子”与表示官名的太子庶子，表示县名的“卢氏”与表示姓氏的“卢氏”虽然可能混淆，但这样的重合非常少，以致可以忽略。此外，还建立官职前常用委任词表（权，迁，授，赠等）。在建立各种表之后，就要运行相应的替换，在编写程序时要注意他们的替换顺序，一般以字数长短为优先级，长字符串优先替换，同时地名表要先于官名表。

第三步编辑正则表达式与编写程序。这里不讨论那些能够直接提取到信息的亲属信息，因为这不需要人文学者提出特殊的要求或建议。但在子孙、女、女婿等亲属信息提取上需要人文学者的参与才能有效的提取信息。还以前述四则信息为例，分析如何在人文学者引导下提取到有效信息。在地名和官名信息被替换后，发现还有一些干扰词汇，这些干扰词中有很多是共通的，大概有以下几类[④]：

表示次序：長，次，幼，曰，季曰，伯曰，仲曰，叔曰，長即，也，次即。

表示科举：貢，等，第，及第，中第，中舉，舉子，登，科。

表示就任官职的动词：今，今以，今爲，授，事，都，轄，新，知，舊，監，倉，庫，起，終，故，前，後，左，右。

表示行政区：州，軍，路，郡，縣，府。

表示地名：江淮，兩浙，寺。

表示官职：尉，某官，官，税務，支鹽。

表示社会身份：士族，士人。

提示职业的词汇：俱，業，業，習。

提示两人以上的词汇：皆，並，并，餘，俱，竝。

表示人生过程的词汇：未，未冠，未仕，未官，未命，先歿，先亡，先公，早夭，早亡，早世，夭，卒，幼，尚，尚幼，未名，前卒，先卒，蚤卒，俱有，早，早卒，喪，早喪。

表示仕宦：未仕，未銓，左銓，司户，户部，户。

其他固定搭配：一，一人，二，二人，一尚，二尚，三尚，三，三人，三曰，四人等。

由于计算机无法识别表示名字的词汇，所以需要把这些词汇在编程中进行批量处理，这一点和前面的替换在原理上是一致的。把这些词汇删除之后，再看这四个句子的情况：

（1）孫男二人/wm/應運/wsep/丙戌/no_noc//wsep//no_noc//wsep//no_noc//wsep/即亨之/wsep/應龍/wsep/

（2）孫男五人/wm/汝直/wsep/汝敦/wsep/汝平/wsep/汝功/wsep/汝能/wsep//vno//no_noc/

（3）孫男六人/wm/夷仲/wsep/虞仲/wsep/於仲/wsep/南仲/wsep/武仲/wsep/延仲

（4）孫男二十人/wm/仲俶/wsep//no_noc//wsep//ns//no_noc//wsep/仲誘/wsep//no_noc//wsep//ns//no_noc//wsep/仲虺/wsep//no_noc//wsep//ns//no_noc//wsep/仲螯/wsep//no_noc//wsep/仲沃/wsep/仲芮/wsep/仲雪/wsep/仲敔/wsep/仲靡/wsep//no_noc//wsep/仲頎/wsep/仲霤/wsep/仲吟/wsep/仲醻/wsep/仲逢/wsep//no_noc//wsep/仲誥/wsep/仲諶/wsep/仲慥/wsep//no_noc//wsep/

在删除干扰词汇后，除了第一个句子中的“丙戌”被误认为名字外（“即亨之”由于超过两个字而不会被认为名字，因为这里的名字略去了姓，而古代的人名极少出现一个姓氏后面加三个字的情形。），所有其他剩下的词汇都是人名。而正则表达式的功能就不再是精确提取人名，而是在程序中被当作亲属信息的表述结构提示了。如果说在前两步中人文学者主要承担文献辨析责任，那么在这一点上，人文学者承担起了更多的技术责任，改变了技术的作用方式和组合方式。这也是整个流程改造的非常重要的两个点之一，甚至是最重要的一个点，因为在官名地名替换后，计算机不能准确定位人名的问题依旧存在，运用正则表达式提取人名依然问题重重，只有将所有的干扰信息尽可能删除后，才能保证人名提取的准确性。这部分内容与词典替换有相似之处，但并非完全相同，不可并为一表，因为有些内容一旦被标准化替换，就会出现各种具体语境中的替换错误。那么这些表示子孙的正则表达式相比于未替换或未删除干扰信息就简化了很多，最主要的表达式如下：

[^女]*(曾孫|孫)([^男女]{0,3}?)人(\/.+)

[^女]*(孫子|子)([^男女]{0,3}?)人(\/.+)

(子|孙|孫|曾孫)男(.{0,3}?)人(\/.+)

[^子孫孙]男(.{0,3}?)人.{0,50}

(子|孫|曾孫|曾孫男|男|孫男|子男|孙男|孙)[一二三四五六七八九十]/

(子|孫|曾孫|曾孫男|男|孫男|子男|孙男|孙|女|婿)曰[^/].{0,50}

生[一二三四五六七八九十][男|子][^女孫].{0,150}

[^第][一二三四五六七八九十][男|子][^女孫].{0,150}?/wm/

以上几个正则表达式帮助我们找到 4000 条左右的子孙信息，如果按照每个信息有两到三个人名，那么将会得到 10000 条左右的亲属信息。关于女儿女婿的信息也可以通过类似的方法提取，不再赘述。

第四步，人工审核。与此前审核需要通看全文查漏补缺不同，在压缩句子时就已经将众多的信息删除，方便人工审核时快速找到信息。在输出结果形式上，采用 Excel 表格，便于批量删除，大大提高了人工审核阶段的效率。由于能在 Excel 表格中直观地看到大量的数据，也便于发现存在问题的规律，有利于反馈给信息技术专家关于正则表达式的修正方案或者干扰词汇的删除工作，这一点技术创新也是人文学者提出并由信息技术专家实现的。但考虑到宋代历史信息比较稀缺，继续采用文本人工标记的方法，双重审核，以保证数据的全面性和准确性。

4 直接提取的重新探索

前面提到之所以采用压缩替代的方法进行人名提取，是因为目前的分词算法还无法有效识别人名。造成这种现象的原因可能是人名的规律还没有足够的认识，也可能是人名的命名本身并无规律，或者带有命名规律的名字占整个名字的比例比较小。

那么命名有规律可循吗？答案是肯定的。商代诸王的命名中有很多天干地支词汇，尤其是天干（甲、乙、丙、丁、戊、己、庚、辛、壬、癸）[8]。面对南北朝人名中大量“之”的现象，陈寅恪指出，清代考据大师钱大昕等人的误读并认为名字中“之”等字的道教信仰背景，因此王羲之的儿子可以叫王献之，并且“之”字在书写中可以省略也不用避讳[9]。同样，和尚道士的法号名字更有规律，如和尚的姓自释道安以后很多为“释”。白惇仁对中国

及周边地区的命名现象进行了考察，认为先秦时期命名中就有五行相生思想的渗入，在宋代有进一步扩散的趋势。依据五行进行排行，不一定与常见的排行字重合比如明代宗室的命名[10]。朱孟臻在《宋代姓名文化研究》中也考察了宋代的数字名现象[11]。这些研究充分说明了古代命名具有一定的规律，而信息技术专家对人名命名规律的忽视可能会造成其分词算法仅仅根据人名出现的规律比如词频或者常见的语法结构（时间、官名、地名等组合）来进行[12]。

张海鸥考察宋代的名字说现象，认为由于宋儒复兴古礼，因此对于命名非常重视[13]。受前人研究的启发，笔者对近万个子孙辈名字的进行了初步考察，初步发现宋人名字背后的几个常见现象。第一类是偏旁部首相同或者同用一个行列的名字，这一类现象最多，比如王安石的兄弟辈（安道、安石、安世、安礼、安上）与子辈（雱、雱、旁、旟、防、旂、旂、放）。第二类是引用四书五经，主要是周易卦名以及儒家圣贤名字以及常用修身用语。比如苏才翁的孙辈（之颜、之闵、之冉，之孟、之偃、之友、之恂、之悌、之邵、之杨、之南、之烈、之点。）和曾孙辈（开、宪、洁、商、若、赤、仕）大部分取法孔门弟子名字，且孔门弟子都是孔子的第一代弟子，一定意义上破除了尊卑之差。第三类是按照五行命名，比如邵潜（水旁，五行属水）之子辈（材、梂、梃、槫，木旁，五行属木）与孙辈（勳、然、熊、熹、薦、谯、点、羔，灬旁，五行属火，这里必须指出勳简体字为勋，一旦简化就无法发现其中的五行属性），朱熹的父亲为朱松也属于此类现象。第四类，名字中包含伯仲叔季等序列词。上述命名规则只是笔者略微观察而得，再次证明运用信息技术获得大量人名信息有助于人文学者开展其专业研究。另一方面，虽然发现的规则尚未穷尽，但对于信息技术专家进行分词研究具有重要的意义，不能再仅仅依据人名出现的规律进行人名的分词研究，还应该根据命名规律，比如汉字的偏旁部首以及其五行属性或者它们的组合类型来进行研究，甚至要结合汉字的笔画数来发现和运用名字背后的规则。那么信息技术专家能否按照这一思路重新进行分词，进而较为有效地完成人名的直接提取，目前尚无结果，笔者将会与相关信息技术专家合作进行进一步的考察，相关结果会另撰一文。再次需要强调的是，有些在信息技术专家看来很难高效解决的问题，经过人文学者的重新思考，能够发现新思路新方法，为更高效地开展人文方面的数据挖掘奠定方法论基础。

5 结语

人文学者从文本出发尝试与信息技术专家共同进行数据挖掘，不断向信息技术专家提出问题，在信息技术专家的帮助下学会数字思考，进而尝试提出可能的解决思路。在这一合作过程中，发现无论哪种标记方法本质上都是利用自然语言表述中的固有规则来提取所需要的信息。经过句子压缩，节省了时间和成本，高效地利用正则表达式标记所需要的信息。正则表达式也不再仅限于标记一些表述结构比较简单的信息，扩大了其应用的范围。同时在某种程度上，改变了其本来的意义，即不再是作为准确信息提取的工具而是成为提示信息出现的工具，与编程相配合，高效地解决了复杂表述中的信息标记问题。在人名分词方面，人文学者也提出了新的方法与思路，对开展进一步的数据挖掘也具有重要意义。当然，如果没有信息技术专家，人文学者的想法也不可能得以实现，二者是不可或缺的。

笔者还发现从数据准确性的角度看数据库建设应当以人文学者为中心，从数据库建设效率的角度来看，数据库建设当以信息技术专家的为中心，但人文学者也发挥着重要作用。历史人文数据库的特点又决定了人文学者的思考在数据库建设过程中更加重要。在合作中，不断改进数据挖掘流程，最终产生出新的方法，提高了数据库建设的效率和准确性。由此可见新流程的探索，既离不开信息技术专家的思考也离不开人文学者的不断思考与发问。人文学者与信息技术专家在建设数据库过程中虽有分工，但更应该彼此合作，互相熟悉，加强交流与沟通，尤其是作为研究者和数据库建设参与者的人文学者更应熟悉信息技

术，由此才能使比较成熟的信息技术为建设数据库服务。当前数字人文的本质还是人文，人文学者应该更加积极的面向信息技术，向信息技术专家学习，在数据库建设中承担起技术责任，为技术的改进做出应有贡献。

致　谢　本文在撰写过程中得到美国哈佛大学中国历代人物传记资料库（CBDB）项目经理马季先生和王宏甦先生的指导，谨致谢忱！

注　释

①已有学者对此问题进行更加全面和理论化的论述，本文侧重于人文学者在参与数据库建设中的具体思考与反思。参见：王宏甦.跨学科合作中的人文学者.第九届上海国际图书馆论坛论文，2018.

②由于宋代地名表中没有两浙，在这里就没有被替换。又因这样的例子很少，没有必要添入地名表进行替换，下文在词典部分会进一步讨论这一问题。

③有必要指出，考、子、男等除表达亲属关系之外还有其他意思，尤其是子，在古代还有个常用含义是先生老师的尊称，由此引申，一些伟大的思想家在后世就以姓氏加上“子”的形式被尊称，如孔子、孟子、老子和庄子。因此还要做一个仔细的审查与排除，将这些专有名词删除。还有一类常见的亲属关系表述是两个不同亲属连称，如父母、父子、男女、兄弟，它们所在的句子往往并没有亲属的人名信息，因此也要对其进行排查与删除，以减少后期审查工作的强度。最后一种常见的是亲属关系言说某些话，比如“母曰：”而非“母曰+名字”，这类表述也不包含需要提取的信息，也需要将其排除在外。为此，需要建立了排除与删除表以供后期编程参考。

④这些词汇首先根据正则表达式提取的文本进行总结，然后运行编程得出结果再进行进一步的修正。

⑤伯仲叔季四个字的每个字都不能单独删除，因为古代中国的名字中不少会出现伯仲叔季，比如第三和第四则信息就出现了“仲”，且出现了 23 次之多。但“伯曰”“仲曰”表示长幼次序的可以删除而不影响结果。

参考文献

[1]王瑞来.警惕数据库[J].史学月刊，2018（9）：21-26.

[2]包伟民.数字人文及其对历史学的挑战[J].史学月刊，2018（9）：5-12.

[3]卷一一六 0[M]//全宋文：第 0 五三册.上海：上海辞书出版社；合肥：安徽教育出版社，2006：297.

[4]Goyvaerts J, Levithan S. Regular expressions cookbook[M]. 2nd Edition. Sebastopol, CA: O’Reilly Media, Inc. .2012.

[5]卷七四四三[M]//全宋文：第三二四册.122.

[6]卷五七八[M]//全宋文：第 0 二七册.204.

[7]卷七五七[M]//全宋文：第 0 三五册.390.

[8].王国维.殷卜辞中所见先公先王考、续考[M]//观堂集林卷 9.

[9]陈寅恪.崔浩与寇谦之[M]//金明馆丛稿初编.北京：三联书店，2001：121-122.

[10]白惇仁.东亚诸邦族谱行辈命名考[M]//第二届亚洲族谱学术研讨会会议记录.台北：联经出版事业公司，1985：181-233.

[11]朱孟臻.宋代姓名文化研究[D]. 宁波：宁波大学，2016.

[12]Chao-Lin Liu, Chih-Kai Huang, Hongsu Wang, Peter K. Bol. Toward Algorithmic Discovery of Biographical Information in Local Gazetteers of Ancient China[C]// 第 29 届亚太地区语言讯息与计算国际研讨会会议论文. http://www.aclweb.org/anthology/Y/Y15/Y15-1011.pdf

[13]张海鸥. 宋代的名字说与名字文化[J].中山大学学报（社会科学版），2013（5）：16-30.

作者简介: 陈佩辉，北京大学哲学系中国哲学专业博士生，chenpeihui@pku.edu.cn。

收稿日期: 2018-09-25。