

●陈源蒸

中文图书机读目录主题标引 采用后控规范的设想

一、问题的提出

1964年,我国出现了第1部中文主题词表——《航空科技资料主题表》。1980年3月,集中全国505个单位1378位图书情报工作者编辑的综合性《汉语主题词表》陆续问世。到目前为止虽已编制近百种主题词表,但真正使用的却寥若晨星。

早在1985年,我国就开始制作中文文献数据库,但进展也十分缓慢,至1990年底可供检索的数据才接近100万条。

中文图书机读目录试验已经多年,1990年北京图书馆编写的《中国机读目录通讯格式(CN-MARC)》已由书目文献出版社出版,北京图书馆同时已开始发行中文图书机读目录数据,不少图书馆也已按照这种格式编制中文图书机读目录,但步履维艰。图书在版编目(CIP)领导小组计划发行的CIP机读数据也处于类似的境地。

归结上述,原因都与规范化的情报检索语言有关:一是编制规范化的主题词表要花大量的人力财力和时间;二是规范化主题词表的词汇滞后现象十分严重;三是采用规范化情报检索语言对标引员和编目员要求很高,而其效率又很低(日均标引文献10篇左右);四是广大用户很难掌握主题词表要领,并利用它找到自己所需的

文献。基于这些原因,发达国家在联机检索系统中已出现采用自然语言标引和检索的明显趋势。

二、采用自然语言标引 与检索的优点

(一)减少编制词表的人力、物力和时间消耗,克服“词汇滞后”与“词汇浪费”现象。诚然,主题词表还是需要的,正如兰开斯特指出,叙词表并不因自然语言检索系统的出现而成为多余的东西,实际上,最需要叙词表的查找人员是用自然语言系统的查找人员。但这种词表不必事先编制,可以在文献标引过程中逐渐积累,因而不必花费大量人力、物力和时间去突击编表。同时,由于在文献标引过程中可随时选用新词,并根据其词义去建立词间关系,因而可避免“词汇滞后”现象。在已有的分类法与叙词表中,“词汇浪费”严重存在。大体上说,实际使用的比例在1/3到1/2之间,有一半以上的无用词汇白白浪费了编表的人力、物力和时间。采用自然语言标引和检索,这一问题自然就不存在。

(二)加快文献数据库的制作进度。使用规范语言,只有具备专业知识的人才能进行标引,而用自然语言标引,具有高中

文化程度的人就可以直接从文献题名中选词,只有少量文献题名不能表达文献内容的选词才需要少数专业人员去解决。标引效率可提高5~10倍,即人均标引量可达到50~100篇。由于大多数词汇直接选自文献本身,不用查表和概念转换,避免了主题语言标识的不一致。这样,必将大大降低文献数据库的制作成本,减少国家投资,同时也大大加快文献数据库的制作进度,对于书目数据库的建设将具有重大的现实意义。

(三) 适应计算机的应用与发展。兰开斯特认为,联机检索系统对自然语言查找特别有益。即使文献数据库不是用自然语言标引的,用户也可以用自然语言查询,并将其自然语言提问自动地与规范化词对比。而且愈接近自然语言,其检索结果,尤其是查准率就愈好。因为规范化的主题语言词汇数量有限,不能充分表达文献的专指内容,所以查准率较低。而自然语言可以选择任意的专指词,达到任意的专指度,所以查准率比较高。同时用户使用自然语言也比较方便,如果联机检索系统具有完善功能,用户能比较容易地编写出提问公式。

使用自然语言还利于向自动标引检索系统过渡。当文献题名及其文摘所包含的词汇足以反映文献内容时,即可进行自动标引与自动检索,只在文献题名及其文摘所包含的词汇不能或不能完全反映文献内容时,才需人工标引一些主题语言。如果编写文摘时注意避免后一种情况的发生,就可有效地采用无标引制作文献数据的方法。

(四) 便于统一规划,分工共建大型综合数据库。为解决科学技术相互渗透、文献数据源交叉重复的问题,建立综合性大型文献数据库是现代联机情报检索系统的一个特点。这样,就需要有统一的主题词

表。采用先控的规范化语言很难做到,《汉语主题词表》不能为各专业部门接受就是一个明显的例证。如果采用自然语言,就比较容易做到。各个专业可以先按各自选择的自然语言标引,把词表的综合管理放到后面去做,也就是采用“后控规范”的做法。

三、采用自然语言标引 与检索应该、而且已经 解决的三个问题

过去认为,采用自然语言标引与检索,存在查全率低、用户负担重和磁盘空间大三个问题。现在,这3个问题都有了比较好的解决办法。

(一) 查全率问题。过去,对采用自然语言有一种误解,认为其词表是自然产生的,所以没有词间关系,因而对“同义”、“广义”、“狭义”等概念不能有效地处理,从而导致查全率比较低。在这里,有一个对采用自然语言后,怎样看待叙词表的问题。前面已说过,采用自然语言的系统更需要叙词表。与现在采用规范语言的系统相比,不同的只是变“先控规范”为“后控规范”。众所周知,词表编制容易维护难。任何离开计算机的词典管理系统,词表维护工作都只能是一句空话。如果与此相反,我们从宏观效益出发,组织一批人力长期从事主题词表的编制与维护工作,一方面定期对文献中新出现的主题词,进行“同义”、“异义”、“广义”、“狭义”等词间关系处理;另一方面注意收集其他词表编制工作的成果,将其中与本系统有关的部分吸收进来,那末,这种词表就不仅能减轻标引人员的劳动量,而且能提高用户的查全率。这是一种以少数人的专职劳动换取多数人效率提高的好办法,其宏观效益是极为明显的。

(二) 用户负担问题。与前面一个问题相关联, 由于系统中有了规范化的主题词管理字典, 用户构造检索策略已不再是一个问题。现在的软件设计技术已能要求系统开辟多个功能窗口, 为用户提供各种入口和多个可供选择的方案, 以及友好的人机会话界面 (对于新老用户还可以有不同的检索步骤), 从而使用户能轻松而满意地获得其查询的结果。

(三) 磁盘空间问题。用自然语言标引文献, 比用规范化语言标引文献, 所用的词量更多, 因而要占用更大的磁盘空间。随着计算机技术的发展, 一方面磁盘设备的性能价格比大大提高, 存贮空间已不再是数据处理系统的负担, 而且各种信息压缩技术可使同样的磁盘存贮更多的文献数据。何况经过测算, 使用自然语言标引文献所增加的磁盘容量并不是很多, 随着时间的发展可以稳定在一个常数上。

四、CN-MARC 主题标引 采用后控规范的设想

为了加快 CN-MARC 机读数据的制作进度, 建议其主题标引采用后控规范做法, 主要包括 3 个方面: 1、标经人员以抽取题名关键词为主, 再按文献内容标引若干自由词; 2、采用后组方式组配; 3、词表管理部门集中处理标引中出现的源源不断的新词。中文编目人员很容易接受这种做法, 关键是要有全面的系统设计, 注意解决好下面一些问题:

(一) 关于主题词的选取。目前, 主题词只能由人工选取。为此, 要有一定的选词规则, 不然会造成词汇混乱。首先是取词的范围, 特别是复合词的选取, 如“红玉苹果”、“蘑菇战术”是不可分的概念, 而“农业气象学”、“工业经济”就是可分的概念。其次有些概念需加注释, 如“北极星”可能

是讲星座的, 也可能是导弹的名称, 还可能是商标。再次虚词一般不取, 但名词中的虚词要具体考虑, 如《健康与长寿》中的“与”字可以不要, 但《红与黑》中的“与”字就应该取。如此等等, 都应由词表管理部门拟出抽词规则。大家共同遵守, 有了大量的词汇积累, 也就为自动标引创造了条件。

(二) 关于主题目录。在我国图书馆界, 中文图书历来没有主题目录。现在, 要增设主题目录, 很多问题需要解决。比如, 新书要设主题目录, 旧书要不要? 旧书补一套主题目录, 比新书还难。而且卡片式主题目录的维护工作很复杂, 随着词间关系的不断变化而应经常调整目录卡片的位置, 因此, 机检系统要, 卡片目录是否可以不要增加主题目录? 若只考虑机检, 标引工作就简明得多, 如《中国文学批评史》, 只要取“中国”、“文学”、“批评史”三个关键词即可, 在检索中去组配运用, 而不必做“中国文学——文学批评史”、“文学批评史——中国文学”这样的重复标引。

(三) 关于主题词表的维护。全国需要有一个词表管理部门, 理想的做法是由北京图书馆承担, 如北京图书馆要做的事情较多, 也可由另一个单位承担。但国家对此应有投资, 且要有正式的规定。词汇来源主要为 CIP 机读数据、CN-MARC 机读数据和“回溯转换计划”所制作的大批机读数据。对书目记录中所标引的每一个关键词和自由词, 都按编制叙词表的要求, 包含词汇名称、拼音名称、英译名称、注释以及用、代、属、分、参诸项词间关系。与《中图法》和《科图法》分类号对应的词汇记录, 也应定期编入词库。词库中包括主表、附表和各种辅助索引。新词先做批量查重, 更新则重新进行编辑处理, 随着词汇量的增加, 可以采取分批分阶段的更新方法。词表不出印刷本, 主要

●王子舟

丛书起始辩证

丛书起于何时,历来众说纷纭。其中一种意见最为流行,即推南宋二俞编辑的《儒学警悟》为丛书之祖。其实,丛书之制在南北朝齐梁时代就已成熟,萧齐陆澄辑《地理书》一百四十九卷,梁代僧祐纂《法集》六十二卷,皆已创丛书之成法。不过,为说明这个问题,还须先从丛书定义入手,因为丛书起始的歧义与丛书概念的纷争有关。名不正则言不顺,正名方能追原。

一、丛书概念的界定

“丛”字有“聚集”、“众多”、“细碎丛脞”的意思。“丛书”二字连用,见于唐代韩愈《剥啄行》诗中“门以两版,丛书于间”,意为藏书众多。首次用作书名,为唐陆龟蒙《笠泽丛书》。陆氏将自己繁杂细碎的诗文作品汇聚成书,按其性质说还不是丛书,

属个人别集。《新唐书·艺文志》将其列入了集部别集类中。

宋元时期,随着雕版印刷术的普遍使用,图书出版量剧增,汇编多种著作作为一部大书的出版物大量出现。南宋嘉泰中俞氏兄弟的《儒学警悟》七集,收书六种四十一卷。咸淳中左圭的《百川学海》十集,收书一百种一百七十九卷。元季陶宗仪的《说郛》,其子书内容虽有删略,但汇集图书有千余种。丛书开始走上繁荣之路。延至明代,以“丛书”冠为书名的名实具备的丛书也灿然而出,如有程荣的《汉魏丛书》,胡文焕的《格致丛书》,钟人杰的《唐宋丛书》等。清代学者钱大昕谈到《百川学海》时,称其为“荟粹古人书,并为一部,而已意名之”^{〔1〕}。这种新形式图书——丛书的始祖。王鸣盛说“取前人零碎著述,难以单行者,汇刻为丛书。”^{〔2〕}纪

供计算机处理使用。

(四) 分类法与主题词表的关系。分类主题一体化已成为一种发展趋势,所以在主题词表的维护中要兼顾《中图法》与《科图法》的修订工作,如以分类法中的主要类目划分范畴索引,使之相互配合,以求整体效益,在指导思想,检索以主题词法为主,分类号码主要用于文献排架和编制索书号码,类目宜粗不宜细,号码不能过长,这些工作需有良好的软件支持。分类法要有印刷本,后面应附主题索

引,以方便大多数图书馆使用。

(五) 具体应用办法。一般来说,全国只有很少几个图书馆要使用全部机读词库,多数图书馆仅使用其中的某几个部分。抽取办法可参照编制专业主题词表的经验,既可按一定范畴和词族进行套录,也可以本馆所积累的词汇到词库中检索。这是一个比较复杂的问题,但只要大家同心协力,做到是可能的。

(作者单位:中科院文献情报中心。来稿时间:1991.2。编发者:丘峰。)

of book price as well as other reasons. In the meantime, the men-times service benefits of the money dropped from 0.867 man / yuan to 0.445 man / yuan. the author analyses the four main causes of the fall of funds benefits and puts forward five measures in order to raise them.

Public libraries——China

Expenditure——Benefit analysis

G259.252

The Deep Development of Periodical Informatoin in Libraries / Lin Pingzhong, Gu Dewei & Wang Zhiping / Bulletin of the Library Science in China / China Society of the Library Science.—1991, 17 (4) .—66~68

Periodicals are one of the main types of library documentations collection. According to library statistics, the number of periodical information accounts for over 65% of the total documentary information. In order to accomplish the modernization of periodical management and to strengthen the deep development of periodical information, deep processing of periodical literature should thereby be brought about, e.g. the compilation of lineal and quadratic documentation. Experiments have showed that the amount of information got by same research at the same time from his direct reading of the quadratic and cubic documentation is 10 to 15 times of that of his reading of the lineal documentation. For the purpose of improving the compiling quality of the cubic documentation, computer-based deep development of periodical information should be carried on as soon as possible. The computer-based periodical management system qualifies the functions of the "hot-point" of forecasting subject study and the establishment of subject bellwether. In

addition, the two computer-aided systems can also be used for the systematic demonstration of problems for study and the preparations of citation analysis. 3 illustrations.

Computer-based periodical management system——Functions

Periodicals——Deep development

G255.2

An Approach to the Adoption of After-Control Standard in Subject Indexing of Chinese Books-MARC / Chen Yuanzheng / Bulletin of the Library Science in China / China Society of Library Science.—1991, 17 (4) .—69~72

It has been for years since the experiment in Chinese Books-MARC was made. The format for the Chinese Books-MARC (CN-MARC) drawn up by the National Library of China had already been published by the Catalogue and Document Publishing House in 1990. Meanwhile, the distribution of the CN-MARC data has already been taken place by the Library. The author of this article, in view of existing problems, suggests replacing the precontrol standard of indexing and retrieval with the after-control one. The actual practices taken are: 1. First place should be given by indexing makers to the extraction of key words of the subjects; 2. The form of coordinate indexing should be adopted; 3. The steady stream of new words appeared in the course of indexing should dealt with

collectively by special agencies. The author also put forward five other jobs that should be done well, such as, the formulation of rules and regulation for selecting words.

CN-MARC——Indexing

After-Control standard——Approaches

G254-39

The Origin of Series: Differentiation and Demonstration / Wang Zizhou / Bulletin of the Library Science in China / China Society of Library Science.—1991.17 (4) .—72~76

Those who hold a viewpoint much in vogue about the beginning of Chinese series believe that the earliest one is none other than the book "A Keen Comprehension of Confucianism" (Ru Xue Jing Wu) (1202) collected by Yu Dingsun and Yu Jing of the Southern Song Dynasty. Actually, the system of series had already ripened into maturity during Qi and Liang of the Northern and Southern Dynasties. The "Books of Geography" (Di Li Shu) collected by Lu Cheng of the Southern Xiao Qi Dynasty (479-502) and the "Collection of Laws" (Fa Ji) compiled by Seng You of the Southern Liang Dynasty (502-557) all had achieved in the proper form of series. In actual practice of compiling and publishing, classifying and cataloging, it is a very common phenomenon to treat different books of same subject as series. This shows that people hold an identical view of understanding series, and there is no difference between the broad sense and the narrow sense of it. To raise the question of the narrow sense of series is to add to the confusion of the concept of series. And it is also not proper to cite the book "A Keen Comprehension of Confucianism" as the earliest series according to the narrow sense of it.

Series——China

History of books——Studies

G256.1

An Afterword of Collating "Han Fei-Zi" / Zhang Jue / Bulletin of the Library Science in China / China Society of the Library Science.—1991, 17 (4) .—77~82

It is highly commendable for collators of ancient books to collate the original copy of the original edition. If the second-hand material is used, one will make mistakes without knowing it. The harmful practice in the collation circles of replacing the firsthand material by the secondhand one is worthwhile to pay attention to. The article has evaluated the gains and losses of some editions of "Han Fei-Zi", e.g. Gu Guangqi's "Han fei-Zi Shi Wu" (Mistakes Seen in Han Fei-Zi); Chen Qiyong's "Han Fei-Zi Ji Shi" (A Variorum of "Han Fei-Zi") and Wang Xianshen's "Han Fei-Zi Ji Jie" (A Variorum of "Han Fei-Zi").

Rare books——Studies

"Han Fei-Zi"——Collations

G256.3

(周钰萍、张木早译, 周钰萍校)