

《中国档案主题词表》 研究与使用精要

张正强 编著



上海科学技术文献出版社

《中国档案主题词表》 研究与使用精要

张正强 编著

上海科学技术文献出版社

(沪)新登字 301 号

《中国档案主题词表》研究与使用精要

张正强

*

上海科学技术文献出版社出版发行

(上海市武康路 2 号 邮编 200031)

全国新华书店经销

上海科技文献出版社昆山联营厂印刷

开本 850×1168 1/32 印张 8.5 字数 228,000

1995 年 12 月第 1 版 1995 年 12 月第 1 次印刷

*

印数:1—2000

ISBN 7-5439-0871-9, Z·752

定价:23.00 元

《科技新书目》377-260

前 言

档案信息资源是一种难以估价的宝贵资源,随着我国社会主义现代化建设事业的发展和计算机的广泛运用,要求进一步开发档案信息资源,由此,利用主题词进行档案检索也日显迫切和需要。《中国档案主题词表》是与中华人民共和国国家标准《档案著录规则》相配套的主题检索语言,它的编制和出版,将为全国各档案部门开展档案主题标引、进行档案主题检索奠定基础,同时,也为在全国范围内逐步实现档案主题标引与检索的标准化和现代化提供了语言保证。

《中华人民共和国档案法》对全国各级各类档案机构所收藏档案的开放、利用作出了明确的规定。因此,在开放和利用档案中,实行主题词标引,是完善档案检索体系,更好地发挥计算机功能,进一步开发档案信息资源以满足用户需求的好方法。

自《中国档案主题词表》编制出版以来,笔者一直从事档案主题法的教学与研究,现不揣冒昧,编写了此书,供从事档案工作的同行参考。

在我国档案界,开展主题标引与检索的时间还很短,相应的有关档案主题词表的研究与档案主题标引规律的探讨也刚起步,本书正是在这种客观情况下编写的,由于时间仓促,加之有些问题(尤其是档案主题标引方面)过去在档案界探讨较少,书中疏漏和不足在所难免,敬请谅解和指正。

张正强

1994.8 写成于上海

目 录

第一章 《中国档案主题词表》的基本原理	(1)
1.1 档案主题检索中的语义控制	(1)
1.1.1 自然语言用于档案主题检索的局限性	(1)
1.1.2 档案主题检索的本质	(2)
1.1.3 经语义控制的检索语言——主题词语言	(3)
1.2 主题词语言实施语义控制的方法	(4)
1.2.1 对自然语言的单个语词进行控制	(4)
1.2.2 对自然语言的词类进行控制	(5)
1.2.3 对自然语言词汇进行系统控制	(6)
1.3 语义控制的评价指标	(7)
1.3.1 词汇等同率(入口率)	(7)
1.3.2 词汇参照度	(8)
1.3.3 词汇关联比	(9)
1.3.4 范畴和词族规模	(9)
1.4 语义控制与检索效率	(9)
1.4.1 检全率与检准率	(10)
1.4.2 检索速度、检索的方便性及检索的成本效益	(12)
1.4.3 语义控制与检索效率	(12)
第二章 《中国档案主题词表》的语义控制	(14)
2.1 《中档表》对单个语词的控制	(14)
2.1.1 词形控制	(14)
2.1.2 词义控制	(19)
2.2 《中档表》对词类的控制	(21)
2.3 《中档表》对词汇的系统控制	(22)
2.3.1 等同语义关系的控制	(22)

2.3.2 属分语义的控制	(23)
2.3.3 相关语义的控制	(26)
2.3.4 范畴语义的控制	(28)
第三章 《中国档案主题词表》的结构体系	(29)
3.1 《中档表》的宏观结构体系	(29)
3.2 《中档表》的微观结构	(30)
3.2.1 主表的主题词款目	(30)
3.2.2 范畴索引	(34)
3.2.3 主表拼音音节索引	(39)
3.2.4 词目首字笔划检字表	(39)
3.3 主表与范畴索引的联系	(40)
第四章 《中国档案主题词表》语义控制的评析	(43)
4.1 析《中档表》语义符号冗余现象	(43)
4.1.1 语义符号冗余现象分析	(43)
4.1.2 语义符号冗余现象对档案标引与检索的影响	(45)
4.1.3 对语义符号冗余现象的改进策略	(46)
4.2 析《中档表》等同语义关系的控制	(47)
4.2.1 《中档表》等同语义关系控制的作用	(47)
4.2.2 《中档表》等同语义关系控制的现状	(49)
4.2.3 《中档表》等同语义关系控制的改进策略	(52)
4.3 析《中档表》属分语义关系的控制	(53)
4.3.1 《中档表》属分语义关系控制的基本估价	(53)
4.3.2 《中档表》属分语义关系控制现状的具体分析	(54)
4.3.3 《中档表》属分语义关系控制的改进策略	(59)
4.4 析《中档表》相关语义关系的控制	(60)
4.4.1 《中档表》相关语义关系控制的统计分析	(60)
4.4.2 《中档表》提高相关语义关系控制比例的可能性	(62)
4.4.3 《中档表》在控制相关语义关系上应注意的问题	(66)
4.5 析《中档表》语义限定注释	(68)
4.5.1 《中档表》语义限定注释中存在的问题	(68)
4.5.2 《中档表》语义限定注释的改进策略	(75)

4.6 析《中档表》语义涵义注释	(75)
4.6.1《中档表》语义涵义注释存在的问题	(75)
4.6.2《中档表》语义涵义注释的改进策略	(82)
4.7 析《中档表》范畴语义的控制	(83)
4.7.1《中档表》范畴索引中范畴划分的问题	(83)
4.7.2《中档表》范畴索引中涵义注释的问题	(84)
4.7.3《中档表》范畴索引中主题词归类问题	(85)
4.7.4《中档表》范畴语义控制的改进策略	(86)
4.8 析《中档表》无关联词的问题	(87)
4.9 析《中档表》词形控制的问题	(87)
4.10 析《中档表》并列词组作正式主题词的问题	(89)
4.11 析《中档表》有关表达事物易名的问题	(93)
4.12 析《中档表》简称词与全称词关系的问题	(96)
第五章 《中国档案主题词表》结构优化的探讨	(102)
5.1 建立《中档表》的轮排索引	(102)
5.1.1 建立《中档表》轮排索引的必要性	(102)
5.1.2 建立轮排索引对《中档表》的作用	(103)
5.1.3 轮排索引的编制	(107)
5.2 建立《中档表》的入口词表	(109)
5.2.1 建立《中档表》入口词表的必要性	(109)
5.2.2 入口词表的功能	(111)
5.2.3 入口词表的编制	(112)
5.3 实现《中档表》与《中档法》的一体化	(112)
5.3.1《中档表》与《中档法》一体化的意义	(112)
5.3.2《中档表》与《中档法》一体化的结构	(114)
5.4 析《中档表》款目主题词排序的优化	(114)
5.4.1《中档表》款目主题词排序上的问题	(114)
5.4.2《中档表》款目主题词排序结构的优化	(120)
5.5《中档表》词目首字笔划检字表的优化	(122)
5.5.1 词目首字笔划检字表的功能分析	(122)

5.5.2 词目首字笔划检字表的结构优化	(125)
5.6 《中档表》主表拼音音节索引的优化	(126)
5.6.1 主表拼音音节索引的功能分析	(126)
5.6.2 主表拼音音节的结构优化	(127)
5.7 主表版面安排及优化	(128)
5.7.1 主表版面安排的现状	(128)
5.7.2 主表版面安排的优化	(128)
5.8 范畴索引版面安排及优化	(130)
5.9 机读版《中档表》的开发	(130)
第六章 《中国档案主题词表》的使用	(132)
6.1 《中档表》使用的基本问题	(132)
6.1.1 标引的概念含义	(132)
6.1.2 档案主题的辨析	(134)
6.2 《中档表》的使用	(137)
6.2.1 《中档表》使用前的准备	(138)
6.2.2 审读档案	(139)
6.2.3 主题分析	(140)
6.2.4 主题分析质量的评价指标	(151)
6.3 主题概念转换	(153)
6.3.1 查表选词	(153)
6.3.2 主题概念转换	(159)
6.4 标引结果的著录与审核	(185)
6.4.1 标引结果的著录	(185)
6.4.2 标引结果的审核	(186)
6.5 主题概念转换的质量评价指标	(186)
第七章 档案主题标引规则	(189)
7.1 主题标引的基本规则	(189)
7.1.1 主题标引的基本要求	(189)
7.1.2 标引的优先顺序	(191)
7.1.3 标引的质量管理	(194)
7.2 各种主题因素的标引	(195)

7.2.1 主体因素的标引	(195)
7.2.2 通用因素的标引	(197)
7.2.3 位置因素的标引	(197)
7.2.4 时间因素的标引	(198)
7.2.5 文件类型因素的标引	(198)
7.3 各种主题类型的标引	(199)
7.3.1 单主题与多主题的标引	(199)
7.3.2 整体主题与局部主题的标引	(203)
7.3.3 显性主题与隐性主题的标引	(205)
7.3.4 专业主题与相关主题的标引	(206)
7.4 各种档案类型的标引	(207)
7.4.1 转发类档案的标引	(207)
7.4.2 带有附件类档案的标引	(208)
7.4.3 请示与批复类档案的标引	(210)
7.4.4 贯彻、实施、学习类档案材料的标引	(211)
7.4.5 会议类档案的标引	(212)
7.4.6 计划、总结类档案的标引	(213)
7.5 各种内容性质档案的标引	(215)
7.5.1 党务类档案的标引	(215)
7.5.2 劳动行政工资类档案的标引	(219)
7.5.3 税务类档案的标引	(222)
7.5.4 行政区划类档案的标引	(227)
7.5.5 基建类档案的标引	(231)
7.5.6 教育类档案的标引	(234)
第八章 主题检索工具的组织	(240)
8.1 编制主题检索工具的意义	(240)
8.2 主题检索工具的组织	(241)
8.2.1 主题条目编制法	(241)
8.2.2 主题标目的结构	(245)
8.2.3 主题检索工具的排列规则	(247)
8.2.4 主题指引卡编制法	(249)

8.2.5 主题参照卡片和说明卡片编制法	(250)
8.2.6 主题检索工具管理和维护	(252)
8.2.7 编制主题检索工具的基本要求	(253)
第九章 建立计算机主题检索系统.....	(254)
9.1 计算机检索系统的原理、结构和功能	(254)
9.2 建立计算机主题检索系统的主题标引	(255)

第一章《中国档案主题词表》的基本原理

1.1 档案主题检索中的语义控制

1.1.1 自然语言用于档案主题检索的局限性

在档案主题检索中,档案检索人员要使用语词来表达所需的档案主题信息,而档案标引人员也要使用语词来表达档案主题概念,在这种档案主题检索中,检索人员和标引人员所使用的语词可以是自然语言,也可以是采用以自然语言为基础的、经过语义控制而形成的主题检索语言。自然语言是指人们在日常交流中所采用的书面语言或口头语言,它是一种不受控制的语言,这种语言应用于档案检索就会出现以下一些现象:

(1) 语词的同义现象。即同一个概念语义可以用多个语词来表达,多词一义,如“文件题名”与“文件标题”、“全宗”与“芬特”、“自行车”与“脚踏车”等。如果对自然语言中的这种同义现象不加控制,这样在档案主题检索中,就有可能出现标引人员用了表达同一语义的A词,而检索人员却用了表达该语义的B词,以至造成标引人员和检索人员对相同主题概念的表达用词不一致,从而导致档案主题检索的检全率下降。

(2) 语词的多义现象。即一个语词可以表达多个概念,具有多种语义,如“疲劳”既可以指人体生理上的疲劳,也可以指材料的疲劳,象这种现象如果脱离了该词的具体上下文语境,就很难判别其语义的指称。在档案主题检索中,如果标引人员和检索人员使用同一个词,即标引人员使用同一个词的A词义,而检

索人员却使用了该词的 B 词义,这样就会造成标引人员和检索人员对同一语词在语义上使用的不一致,从而导致档案主题检索的检准率下降。

(3) 语义关系无显示现象。即自然语言无法清楚地显示概念语义之间的关系,如“灌木”、“下木”和“乔木”这几个词之间具有什么样的语义关系,这是显示不出来的。自然语言中的这种现象的存在,给标引人员和检索人员正确选词以及给检索人员扩大检索范围、缩小检索范围和改变检索范围带来了困难,从而使档案主题检索效率下降。

此外,自然语言中还存在着语义含糊,词汇量巨大等现象。由于自然语言存在着以上种种不利于档案主题检索的现象,因此,在档案主题检索中不采取任何语义控制手段而直接、单纯地使用自然语言,很难达到较高的检索效率,很难满足档案用户的多种检索要求,同时,也很难把它们加工成计算机能够处理的形式,用计算机管理。

1.1.2 档案主题检索的本质

档案主题检索从广义来讲,就是指档案主题信息的存贮并依据用户对档案主题信息的需求查检出相关档案材料的过程;从狭义上讲,仅仅是指依据用户对档案主题信息的需求查检出相关档案材料的过程。档案主题检索不论其广义还是狭义,从其本质上来说,就是档案信息的交流与传递,这种交流与传递和日常人们用自然语言进行信息交流与传递一样,也会存在着交流中的语言障碍。如果自然语言不加控制地应用于档案主题检索,那么,其所存在的同义现象、多义现象、语义关系无显示现象以及语义含糊现象等等就是档案主题检索中的语言障碍。显然,在档案主题检索中存在着这样一些语言障碍而要进行有效的档案信息的交流与传递是不可能的,所以,从档案主题检索的本质上来看,要提高检索效率,就必须对应用于检索的自然语言实施语义控制,以

消除档案主题信息交流与传递过程中的语言障碍。

1.1.3 经语义控制的检索语言——主题词语言

在检索语言的范围内,检索语言可分为分类语言和主题语言。主题语言是指以自然语言的字符为字符,以经语义控制的名词术语为基本词汇,以语义间的逻辑关系作为语法和词法,用一组词语作为文献检索标识的这样一类检索语言。主题语言又分为标题词语言、单元词语言和主题词语言,其中的主题词语言就是通常所说的叙词语言。主题语言的特点是,词汇都是以自然语言为基础的受控词汇,其具体体现就是收录于各种词表中的标题词、单元词或主题词。主题型的词表根据类型分别称为:标题表、单元词表和叙词表。由于在我国通常把叙词表称为“主题词表”,因此,叙词语言又被称作为主题词语言。《中国档案主题词表》中所集中的全部词汇就是经语义控制的受控词汇,其所形成的检索语言就是主题词语言。

主题词语言是为适应于主题检索中所要求的语义控制而产生的一种检索语言,这种语言与自然语言相比具有如下几个特点:

(1) 语义的唯一性。在主题词语言中,通过语义控制,一个语词只能表达一个概念语义,一个概念语义只能用一个语词表达,即概念语义与语词一一对应,从而消除了自然语言中所存在的语词同义和多义现象。

(2) 语义的明确性。主题词语言中的语词,经过语义控制,其所指称的范围是明确的,因而消除了自然语言中语词涵义的模糊现象。

(3) 语义的系统性。主题词语言中的词汇有一定的层次结构,能清晰地显示概念语义间的等同关系、属分关系及相关关系,从而在词汇的整体性上揭示了语义系统中概念语义的区别与联系。

主题词语言与自然语言相比,除了具有以上特点外,还具有词汇的简明性、稳定性等特点。由于主题词语言具有以上这些特

点,因此,当标引人员使用主题词语言时,用来表达一个主题的语词是从事先编好的主题词表中选出来的,从而使主题标引更加统一;而对检索人员来说,如果表达某一主题需求的语词也是从规定的主题词表中选出来的话,那么这些词用来表达相应的主题需求就比较容易。据此,通过主题词语言把标引人员表达主题的用语(即标引用语)和检索人员表达主题的用语(即检索用语)统一起来了,从而减少了主题检索中的语言障碍,使主题检索能够得以顺利地实现。

1.2 主题词语言实施语义控制的方法

主题词语言是在自然语言的基础上,经过语义控制而形成的一种适合于主题检索的人工语言,《中国档案主题词表》只是主题词语言中的一个语种,但其所进行的语义控制也遵循主题词语言语义控制的原理、规律和方法。在主题词语言中,所谓语义控制就是把自然语言加工成主题词语言的过程。

主题词语言所采用的语义控制的方法很多,但归结起来主要包括三个方面:

- (1) 对自然语言的单个语词进行控制;
- (2) 对自然语言的词类进行控制;
- (3) 对自然语言的词汇进行系统控制。

对自然语言进行语义控制的目的是为了规范自然语言,使其适合于主题检索的需要,因此,语义控制的过程又往往被称之为规范化处理,而对自然语言所实施的语义控制的方法则被称之为规范化措施。

1.2.1 对自然语言的单个语词进行控制

对自然语言单个语词进行控制的方式主要有两类:一类是对单个语词的词形进行控制,一类是对单个语词的词义进行控制。对

自然语言的单个语词进行控制，其作用就是使其语义在表达上具有唯一性和准确性。

(1) 对单个语词的词形进行控制。包括对自然语言中一词多形的控制，不同书写形式的控制，词序的确定，以及同义词的控制。词形是语义的符号，对词形进行控制，就是规定作为检索用的语词其一个语义只能与一个词形符号相对应，而不能与多个词形符号相对应，如简体、繁体、异体等。这样就可以使标引人员和检索人员在表达同一概念语义时都能使用同一词形符号，保证词形使用的规范化，减少因词形使用上的无序而造成的语言障碍。

(2) 对单个语词的词义进行控制。包括对自然语言中多义词的控制和词义含糊的词的控制。词义控制就是规定作为检索用的语词其一个词形符号只能用来表达一个概念语义，而且必须明确，不能用一个词形符号同时来表达多个概念语义，如果一个词形符号可以表达多个概念语义，就必须给予单义化，使其唯一；如果一个词形符号所表达的概念语义不明确，就必须加以注释或定义，使其明确，从而使标引人员和检索人员在单个词的语义的使用上规范化，减少因语义使用上的无序而造成的语言障碍。

1.2.2 对自然语言的词类进行控制

主题词语言是以自然语言中的语词作为基本词汇的，但从自然语言语词所表达的语义性质上区分，可以区分出很多词类，如汉语中就包括名词、动词、形容词、副词、感叹词、介词等多种词类。自然语言中的这些词类并不都适合于主题检索，因此，在选择自然语言中的语词作为主题词语言的基本词汇时，对词类必须有所限制。主题词语言在选择词类时，一般都遵循这一规则，即主要选用能表达文献主题概念的名词和名词性词组，根据需要，也选用少量的形容词、接头词等。

在汉语中，名词按其结构一般分为单纯词与合成词。单纯词是指只含有一个词素的词，如火、鸟、码等；合成词是包括两个

以上词素的词，如戏剧、铁路、电流等。名词性词组按结构可分为五种：

- (1) 联合式：如，法规条例、学工学农；
- (2) 偏正式：如，轻工业、重工业；
- (3) 主谓式：如，电子干扰、工资调整；
- (4) 复杂词组：如，欧洲文艺复兴运动；
- (5) 固定词组：如，戊戌维新、文化大革命。

名词性词组中可以包含有介词、连词或数词，如以工代赈、五四式手枪。除名词和名词性词组以及少量的形容词以外，其他词类在主题词语言中一般都不得使用。词类控制可以从语义上排除主题词语言中与表达文献主题无关的词类，使标引人员和检索人员都能使用相同的词类来表达主题及主题需求，从而保证了主题检索中词类使用的规范化，由此可减少因词类使用上的无序而造成的语言障碍。

1.2.3 对自然语言词汇进行系统控制

词汇是语词的集合，对词汇进行系统控制就是将词汇间的语义关系加以系统地揭示出来。对自然语言词汇进行系统控制，其控制方式主要有两类：一类是对词汇词形的系统控制，一类是对语汇语义的系统控制。其控制的目的是为了建立起一个适合于主题检索的多维的语义系统，使主题词语言诸词汇之间的相互联系、相互区别得到清楚地显示。

(1) 对词汇词形的系统控制。这主要是利用词汇中具有相同词素的形体可以字面成族的原理，使词汇中具有相同字面的词按其词形集中在一起，从而显示词汇之间在语义上的某种联系、关系与区别。这样，使标引人员和检索人员在运用主题词语言词汇时，就有一个可资比较的语义环境，在这一语义环境中，标引人员和检索人员易于正确地识别和选择词汇，从而达到规范标引人员和检索人员对主题词语言使用的目的。

(2) 对词汇语义的系统控制。这主要包括等同语义的控制、属分语义的控制、相关语义的控制以及范畴语义的控制。这一类语义控制,除范畴语义控制外,都是利用语义关系的参照符号来规范词汇整体的语义关系的。通过参照符号,词汇之间的等同语义关系、属分语义关系和相关语义关系得到明确的显示并形成一个纵横有序的语义网络,利用这一语义网络在标引和检索中查找词汇、选择词汇和识别词汇更为容易,语言交流的障碍更少。

范畴语义控制是利用范畴表的形式来进行整体词汇控制的,它把同属于一个范畴的语词集中在一起,从而显示词汇语义之间的相关性。对词汇语义进行范畴控制可以使词汇在检索中的运用更符合查询人员的需要,如查词、选词、区别词汇语义等,因而也是规范主题词语言的使用,减少语言障碍的重要措施。

1.3 语义控制的评价指标

各类语义控制方法的实施,其控制水平如何评价,这就涉及到语义控制的评价指标问题。语义控制的评价指标主要有:词汇等同率、词汇参照度、词汇关联比以及范畴和词族规模等这几项。

1.3.1 词汇等同率(入口率)

所谓词汇等同率是指主题词语言中正式主题词数与非正式主题词(入口词)数的比率,词汇等同率是评价等同语义关系控制强弱的重要指标。

词汇等同率高,说明等同语义关系显示充分,正式主题词所控制的非正式词多,标引人员和检索人员通过非正式主题词及等同语义关系的指引,转换成正式主题词的途径就多,从自然语言进入主题词语言的入口词就多,这样在时间上和空间上就可以提高标引人员和检索人员使用主题词表达主题及其需求的一致性,减少主题检索中以主题词语言作为交流语言的障碍。而词汇等同