



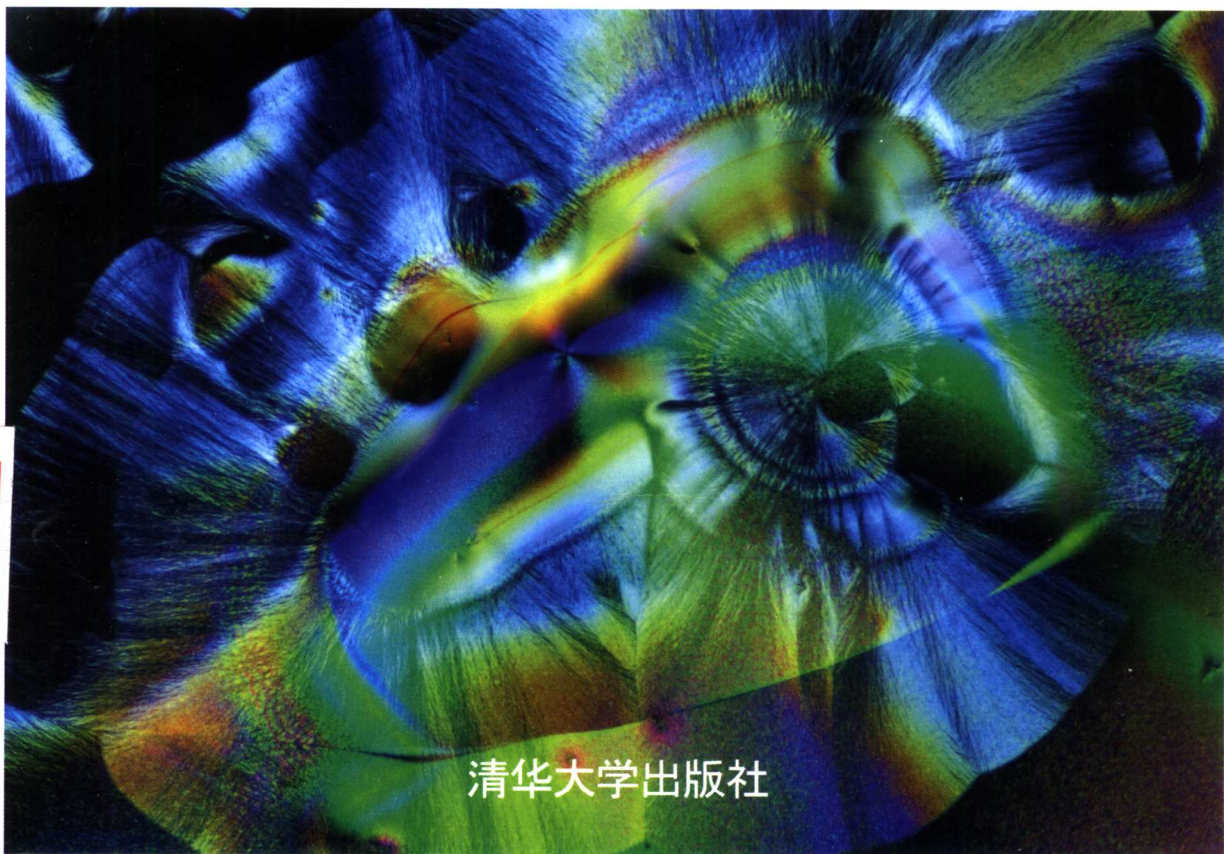
面向21世纪课程教材

信息管理与信息系统专业教材系列

信息组织

(第二版)

马张华 编著



清华大学出版社

面向 21 世纪课程教材
信息管理与信息系统专业教材系列

信 息 组 织

(第二版)

马 张 华 编 著

清华大学出版社
北 京

内 容 简 介

本书为信息管理专业系列教材之一,系统讲述了信息组织的基本理论方法,包括:信息组织原理,信息描述的理论和方法,分类法理论、典型类表以及分类实践,主题法理论、典型词表、主题标引实践以及自然语言标引与检索等。为了使本书能及时反映本领域研究进展和适合教学需要,新版对内容进行了较大幅度的更新和完善。

本书可作为高等院校信息管理专业、图书馆学、情报学专业教学的教材,也可供各种与信息资源组织有关的机构或部门,包括各类型文献机构、网络公司以及企业信息资源管理部门的工作人员学习参考之用,或供对信息资源组织感兴趣的读者阅读使用。

图书在版编目(CIP)数据

信息组织/马张华编著.—2版.—北京:清华大学出版社,2003
(面向21世纪课程教材.信息管理与信息系统专业教材系列)
ISBN 7-302-06652-3

I. 信… II. 马… III. 信息管理—高等学校—教材 IV. G203

中国版本图书馆CIP数据核字(2003)第037987号

出 版 者: 清华大学出版社(北京清华大学学研大厦,邮编 100084)

<http://www.tup.com.cn>

责任编辑:魏荣桥

印 刷 者:北京国马印刷厂

发 行 者:新华书店总店北京发行所

开 本:787×960 1/16 印张:21.5 字数:454千字

版 次:2003年7月第2版 2003年7月第1次印刷

书 号:ISBN 7-302-06652-3/F·533

印 数:0001~5000

定 价:35.00元

目 录

第一章 导言	(1)
第一节 信息组织的必要性.....	(1)
第二节 信息组织的形式和检索途径.....	(3)
第三节 信息组织的处理方式和研究内容.....	(7)
第二章 信息组织原理	(11)
第一节 对信息组织的要求	(11)
第二节 信息组织中的控制与规范	(13)
第三节 规范控制的依据	(18)
第三章 信息描述	(27)
第一节 信息描述的意义	(27)
第二节 信息资源描述项目、标识和描述级次.....	(29)
第三节 检索点和规范文档	(39)
第四节 计算机编码	(47)
第四章 信息描述工作	(55)
第一节 信息描述工作概述	(55)
第二节 信息描述方法	(64)
第五章 分类法	(77)
第一节 分类法概述	(77)
第二节 分类法结构剖析	(81)
第三节 类目体系的建立	(98)
第四节 网络分类法的结构特点.....	(109)
第六章 国内外主要分类法介绍	(117)
第一节 《杜威十进分类法》.....	(117)
第二节 《国际十进分类法》.....	(122)
第三节 《美国国会图书馆图书分类法》.....	(127)
第四节 《冒号分类法》.....	(131)
第五节 《中国图书馆分类法》.....	(137)

第六节 《雅虎》(Yahoo!)	143
第七章 分类标引与分类检索工具	148
第一节 分类标引概述	148
第二节 分类标引规则	162
第三节 分类排架与分类检索系统	173
第八章 主题法	185
第一节 主题法概述	185
第二节 叙词表的结构	190
第三节 叙词语言的词汇控制	206
第四节 叙词表的编制与维护	218
第九章 国内外主要主题词表介绍	225
第一节 《美国国会标题表》	225
第二节 《医学标题表》	232
第三节 《汉语主题词表》	238
第四节 《中国分类主题词表》	244
第五节 《社会科学检索词表》	249
第十章 主题标引和主题检索工具	255
第一节 主题标引概述	255
第二节 主题标引方法	258
第三节 主题标引规则	272
第四节 主题检索工具	289
第十一章 自然语言标引与检索	301
第一节 自然语言概述	301
第二节 关键词语言	303
第三节 文本检索	307
第四节 后控词表	313
第五节 自动标引	320
第六节 自然语言与控制语言性能及两者结合的趋势	329
主要参考文献	334
后记	336
第2版后记	337

第一章 导 言

第一节 信息组织的必要性

随着现代科学技术的发展,人类社会逐渐从信息社会向知识社会迈进,信息资源的开发利用日益受到广泛重视。目前信息资源的特点是数量大、种类多、来源广,特别是随着计算机的广泛使用和网络的逐步普及,现代信息环境所提供的文本复制的便利、自由发表的可能性,以及信息传输能力的加强,使得人们面对的信息资源数量急剧增加。但是社会信息量的增长并不意味着用户获取信息的增长,恰恰相反,无序的信息资源不仅无助于信息资源的使用,反而会加剧信息增长与使用的矛盾。人们生动地称这种情况是“信息超载,知识缺乏”。越来越多的人认识到,“原始信息本身并不能产生价值”,只有将其加以有效的组织,按特定的需要加以集中和揭示,才能产生价值。而要有效开发利用信息资源,必须采用相应的方法加以控制和处理,措施之一就是进行信息组织。

什么是信息组织?所谓信息组织,亦称为信息资源组织,是根据信息检索的需要,以文本及各种类型的信息资源为对象,通过对其内容特征等的分析、选择、标引、处理,使其成为有序化集合的过程。

要正确了解信息组织的含义,必须首先了解数据和信息的含义及其区别。所谓数据,一般是指经过直接观察获得的对客观对象的记述。它是原生态、即未经过处理的,“以正式的、适合人类或自动方式交流、转换、处理的对于事实、概念、程序的表述”。信息则是经过人类处理的数据,它是供交流共享的知识,是“人们通过惯常方式赋予数据意义”得到的结果^{*}。从这个意义上说,我们日常所说许多信息系统,实际上并不是真正的信息系统,而只是数据系统,我们日常所说的信息处理,实际上也只是指数据处理。因此,在我们的讨论中,应注意将信息组织与数据组织相区分。

社会信息作为一种公共交流的知识,是通过不同的形式进行传播的,包括口头方式、传统的出版方式和电子方式等。各种信息媒体是信息存在的形式之一,同时也是进行交流的主要形式。以记录形式进行信息交流,能够克服面对面直接交流在时间上和空间上的局限,使得人类社会得以在继承已有经验的基础上向前发展,对人类社会的发展意义重大,是人类社会进步的重要条件。各种类型的信息资源,即是公共知识的集合,同时也是社会发展的智力资源。

* David P. Best. The Fourth Resources: Information and Its Management Aslib/Gower 1996. p44

在本书中,信息资源一词与文献的含义接近,通常是指一切记录形式存在的信息载体,只是比较而言,更侧重于对新型媒体形式的强调。信息资源范围广泛、种类繁多,就信息资源的基本类型而言,它可以是图书、期刊、报纸、论文、标准、档案等各种类型;就存在的形式而言,它可以是各种传统的印刷型资源,也可以是各种缩微型、视听型、机读型、光盘型资源,以及非实体资源——如存在于网络上的信息资源。信息组织就是以各种信息媒体形式为对象所进行的组织。就其处理的单元而言,信息组织存在着多种不同的层次:它可以直接以信息资源的存在形式为处理对象,例如,直接以一图书、期刊、网站等为处理单元;也可以以期刊、报刊、网站中的个体,如论文、新闻、网站中的构成部门等为处理单元;还可以选择其中的信息成分作为处理的单元,进行组织和揭示。

必须指出,信息组织的目的,是建立起信息资源收藏系统和检索工具,方便信息资源的开发利用。作为一种为了信息检索的需要对信息资源进行有序化组织的活动,它是与信息检索活动密切联系的。因此,要完整地理解信息组织的含义,还必须了解它在信息资源开发领域中的地位,特别是弄清信息组织与信息检索的关系。

从图 1-1 可以看出,任何对信息检索系统的完整论述,一般必须包括信息组织和检索两个方面。

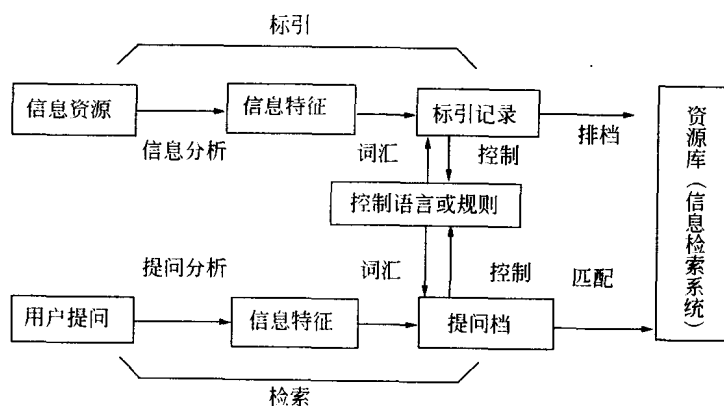


图 1-1 信息整序与信息检索语言

可以看出,在信息检索系统中,信息组织是依据信息资源的主题内容或特征,将无序的信息资源组织为有序集合的过程。信息组织是信息检索的基础,通常是根据检索的需要,以信息资源的一定单元为处理对象,对其特征加以记录,并根据需要进行必要标引以及相关处理,然后将这些信息资源记录或条目,按规定的方式输入系统,组成文献资源库或信息检索系统。为了保证处理能达到预定的目的,信息组织过程中对资源的处理通常应依据相应的、预先制定的规范,如作为描述记录资源特征依据的描述规范、作为资源主题内容标识依据的检索语言及相应的处理规则等。

信息检索则是根据检索需要,从检索系统中检出相应信息资源的过程。信息检索是

信息组织的目的和归宿,是信息组织的反向过程。在实施信息检索时,一般应首先依据该检索系统的特点,将检索提问转换成相应检索特征,再根据该系统提供的检索途径进行匹配查找,检出相应信息资源。在这一过程中,一般应注意按照系统使用的描述规范及控制语言的特点进行操作,以使检索获得较好的效果。

从信息组织和检索的过程可以看出,信息组织与信息检索是密切联系的。一方面,信息组织是信息检索的基础和前提,只有进行信息资源处理和有序化组织,才可以利用它有效地进行检索查找;另一方面,信息检索则是信息组织的出发点和归宿,是信息组织的真正原因。两者既有区别,又是相互依存、相互作用、互为因果的。因此要使信息组织取得较好的效果,必须充分了解系统的检索需要,根据信息资源的特点和用户的使用要求加以实施。反之,信息检索必须在信息组织已有的基础上进行,离开信息存储和处理的条件,就无法有效地检索。

信息组织和检索中使用的、用以描述检索特征的控制语言及规则,则是根据文献或信息资源组织和揭示的需要制订的组织规范。包括信息资源描述规范、情报检索语言等。前者如文献的编目规则及各种元数据描述或记录的规范等,后者则是一种作为主题标引依据的人工语言系统,包括各种分类表或标题表、叙词表,以及分类标引和主题标引的规则系统等,一般用来对所使用的词汇及其相关性进行规定,因此从本质上讲是知识组织的一种形式,目的是将知识结构引入信息资源组织,是文献存储和检索的语言保障。信息标引选择和组织的规范是保证信息组织过程中的处理质量以及一致性的重要依据,是信息组织和检索的重要工具。也有一些系统对信息资源标引或处理时不依据词表,但这类系统一般仍需要遵守一定的标引和处理规范。

第二节 信息组织的形式和检索途径

信息组织的目的是组织各种类型的信息资源检索工具,同时也在必要时对文献单位的信息资源进行有序化组织。

对信息资源进行组织,是指按照一定的方式将信息资源本身直接纳入信息资源集合。这是信息组织的任务之一。这一组织对传统文本文献的使用尤其重要,是文献单位信息组织的重要内容。

文献组织一般有两种基本排列方式,其一为固定排列法,其二按一定意义的次序排列。固定排列法或按照文献的形式,或根据到馆先后,或使用文献本身固有的号码,如专利号、标准号、报告号等排列文献资源;一些不适合按信息资源特征动态组织的资源类型,一般均采用固定排列的方法。这类文献资源包括档案、博物馆的收藏品等,此外如光盘、录音带、缩微品等,也大多按照固定排列的方式收藏。这类资源使用时通常必须先依据检索工具查出其所在的位置。

按一定意义的次序排列包括按题名顺序排列、按责任者排列等,但使用最多的是按分类方式排列。由于以分类方式组织文献资源,可以提供一种直接按照内容关系对文献资

源进行浏览检索的途径,为用户的使用提供了方便,因此一般文献单位均采用分类排列的方法。但在采用这一方法进行文献库组织时,对于同时具有两种以上内容的文献,通常只能按其中的一种排列,例如,一同时论述物理学和化学的著作,如按物理学归类排列,就无法同时按化学进行组织;此外,为了方便读者使用,某些特定形式的文献如工具书、特定对象的文献如儿童读物、盲人读物等往往需要单独组织,这些情况都会造成同一内容文献的分散。至于期刊、报纸中文献的内容,就更无法在文献组织中揭示了。因此在进行分类排列时一般应根据整体情况加以规划,同时通过检索工具予以揭示。

与信息资源本身的组织不同,检索工具一般以文献记录作为文献的替代物,可以方便地按统一的方式,从多个角度、多种层次组织检索系统。对于检索工具的类型,可以有多种不同的区分方法。按照检索工具的对象和特点,可以分为以下类型:

① 文献目录。通常是以各种媒体的文献单元为对象,对其进行记录、报道与揭示的工具,包括文献收藏目录和书目。前者以一个或多个文献单位收藏的文献为对象加以描述和揭示;后者则以某一领域、地域或时限内出版的文献为对象进行组织,通过对文献压缩、浓缩与整序,为用户提供较为全面的文献信息。

② 索引。是以文献集合或文献中包括的文献单元或信息内容为其揭示对象的检索工具,包括图书索引、期刊论文索引、报纸索引、专利索引、标准索引、档案索引、会议文献索引、语词索引、书后索引等。它一般可以给用户提供某一内容、特征的查找线索,其揭示能深入到文献所包含的信息单元,将其与目录结合,可以构成不同层次文献信息的有效揭示体系。

③ 机读数据库。是由文献资源为对象的机读记录的有组织的集合。这是一种依托现代计算机技术,以机读形式建立的检索系统。按照数据的特点,机读数据库可以分为参考数据库和源数据库两类。参考数据库中,书目数据库通常收录文献目录或索引数据,是手工文摘索引刊物和文献目录的机读形式;指南数据库存储有关机构、人物、计划、活动的简要描述,具有为有关资源查找提供线索的作用。源数据库包括事实数据库、数值数据库、全文数据库、图像数据库、音像数据库等多种类型,一般均直接收入相应的原始资料。机读数据库,特别是源数据库收入的数据往往比较完备,可以利用计算机的数据处理能力进行检索,一般同时提供多种查找方法,并可通过联机方式检索使用,受到广泛的欢迎。

④ 网络搜索引擎。这是一种以网络信息资源为对象的检索系统,实际上是数据库的一种特殊形式。与一般的数据库相比,网络搜索引擎涉及的资源类型种类更多,动态性更强,形式更为多样。它不仅可以处理各种类型网络文献、网站、个人主页等网上资源,而且还可以连接包括聊天室、BBS、网上即时播报等动态资源类型以及各种电子形式的数据库,通过超文本链接的方式,访问各种类型的信息资源。

这样,根据对信息资源中不同对象和层次揭示上的需要,上述信息资源检索工具构成了以不同检索需要为目标的、形式多样的、完备的系统。在上述的各种形式中,后面的类型往往是对在前类型的深化或发展,例如,索引是对目录的一种深化和补充;机读数据库

包括了目录和索引,同时又增加了新的信息资源组织类型和形式,并将其与计算机的数据处理能力结合,是对传统检索工具的一次推进;而网络搜索引擎则是对前几种形式的又一次综合和发展,作为一种网络环境下发展起来的检索工具,其范围更广,包括的资源种类更多,可供检索的对象也更丰富。

按照依据的标识特征,检索工具可以分为以下两种基本类型:

① 一类以信息资源的外部特征为检索依据,提供从资源形式出发进行检索的途径。常见的这类标识包括:责任者名、题名、机构名、出版地、出版社、标准书号、专利号、档案号、文献登录号等。按照这类标识进行组织,一般可以通过已知的资源特征获得明确的查找结果。有的特征,如标准书号、专利号、档案号等,可以据以查出对应的资源对象;其他特征,如责任者名、题名等,通常可将范围迅速缩小到便于查检的数量;此外,还可以通过数种成分的组配进行查找,如通过出版者与出版时间、题名与出版者等的组配,对特定信息资源进行查找。

此外,还有一种根据引证关系建立的检索途径。这种方法将文献写作过程中的引文作为组织检索系统的依据,供用户从引证关系入手对文献进行检索。这一途径也可以通过文献的外部特征直接进行编制。

② 另一类则以表达信息资源主题内容的标识为依据,提供从内容角度进行检索的途径。主要有分类法和主题法两种。分类法以表达信息资源内容的标记符号为标识,按照依据主题之间关系进行系统组织;主题法直接以表达文献内容的语词为标识,依据主题字顺等方式提供检索途径。使用主题内容途径检索,不必预先了解信息资源的具体特征,可以直接从所需要的主题内容入手进行查找。例如,在主题检索系统中,可以直接使用“信息组织”一词在检索系统中进行查找,检出论述该主题内容的有关文献。同时还可以根据“信息组织”这一概念与其他主题词的联系,进行相关内容文献的查找和使用。在数据库中,从内容角度的检索还可以利用资源记录的全部数据进行操作。

可以看出,上述两类检索途径都是十分有价值的,前一类途径可从已知文献数据的角度进行查找,后一类途径则可以根据需要使用,从主题内容出发对信息资源进行检索。比较而言,前一种途径的特点是可以迅速、准确地查找特定文献,缺点是必须预先掌握确切的数据,否则就无法进行,此外,也很难在此基础上扩大查找相关文献。后一种途径的优点是,不必事先了解相应的数据,用户不仅可以根据需要使用直接检索特定主题内容的文献,而且还可以通过系统中主题内容之间的联系,扩大或缩小检索范围,进行相关资源的查找,对于信息资源的开发利用,帮助更大;但在查找确定的对象时,则不如前一种方法直接和准确,因此两者是相互补充的。为了满足用户的不同需要,各种检索工具在对文献信息进行多层次揭示的同时,一般均提供多种检索途径,方便用户从不同角度检索查找。

按照标识记录的方式,检索系统则可以分为资源单元方式和标识单元方式两种。

资源单元方式以信息资源为中心记录资源,组织检索工具。这类检索系统通常以资源的描述记录代替资源进行组织,作为排检依据的检索标识被记录在该资源的描述数据

之中。这种方式形式直观,可以在检索时直接阅读资源的记录,是传统文献单位常用的方式。这种方式的不足是,一个对象在检索系统中只能有一个检索的入口。在手工主题检索工具中,当一个资源涉及多个主题因素时,一般必须组配成一个明确的次序,例:将“河流水污染对鱼类的影响”的主题标记确定为:河流水污染—影响—鱼类。这一标记在按字顺组织的检索工具中,只能从第一个词出发才能查到。由于这类系统的检索标识,在进行检索以前就已经在检索系统中组配好了,因此,也称为先组式检索系统。资源单元方式的检索系统通常均为先组式检索系统。

标识单元方式,则以标识为中心记录资源,一般将一资源的号码记录在相应标识之下,所有的标识按一定顺序排列成为一个系统。检索时,先查出相应的标识,再根据标识下的号码查找对应资源。在进行复杂主题资源查找时,标识单元方式的检索系统,可以按照先组方式揭示信息资源,也可以采用与先组式不同的形式,在单元主题标识的基础上,通过相应主题词的组配进行查找。后一种方式形式简练,使用灵活,在进行组配检索时,也不存在检索入口的问题,便于在计算机环境下进行多途径检索。不足是,这类系统一般不能进行浏览,比较抽象,由于这类系统中检索标识的组配,是根据用户提出检索要求后再进行的,因此,这类系统称为后组式检索系统。一般情况下,计算机检索系统通常采用标识单元结合后组的方式进行标引和检索,手检系统在采用标识单元的形式时多数仍采用先组式。

由于资源单元方式和标识单元方式记录资源的特点不同,在实际使用中,两者往往结合使用。例如,在印刷型书目工具中,一般在采用资源单元方式组织检索系统的同时,往往以标识单元的方式提供多种查找途径,用户可以根据标识下的文献号,查找对应资源的有关数据;在机检系统中,通常以资源单元方式完整输入数据,同时提供从标识角度组配查找资源的机制,检出后即调用描述记录,供用户了解资源的描述信息,通过两者的结合,提供检索的便利。

此外,检索工具还可以按照采用的形式,分为卡片式、书本式、缩微式、电子方式等类型;按照信息资源对象的类型,分为图书、期刊、报纸、专利、标准、档案、图像、会议文献等类型;按照检索系统的功用,分为提供文献线索、提供事实、数据、原文、图像等多种类型以及按照专业范围、地区等进行区分。表 1-1 为一些常见的信息组织类型。

表 1-1 信息组织的常见类型

信息组织划分标准	信息组织类型
按组织对象分	信息资源
	检索工具
按检索工具的类型分	目录
	索引
	数据库
	网络搜索引擎

续表

信息组织划分标准	信息组织类型
按检索对象分	图书检索系统
	论文检索系统
	新闻检索系统
	会议文献检索工具
	专利检索工具
	标准检索系统
	档案检索系统
	博物馆藏品检索系统
	网络资源检索系统
	图像检索系统

按照检索途径分	题名检索工具
	责任者检索工具
	分类检索工具
	主题检索工具
	按照引文关系的检索工具
.....	
按照组配的方式	先组式检索工具
	后组式检索工具
按照专业领域分	经济
	法律
	医学
	教育
.....	
按照地区分	北京
	天津

按照设备	卡片式检索工具
	书本式检索工具
	缩微式检索工具
	电子方式检索工具

第三节 信息组织的处理方式和研究内容

信息组织从总体而言,是根据检索的需要和信息资源的特点,将信息资源组织成有序的系统的过程;就个体而言,则是根据信息组织的整体要求和规范,对信息资源进行描述和处理,并将处理结果纳入系统的过程。信息组织的总体目标是通过具体的信息资源

个体的处理实现的。因此,根据信息组织的需要,明确对信息资源处理的要求十分关键。传统文献单位对文献处理的基本目标主要有以下五个方面,包括:*

- ① 识别、确认存在的所有物理媒介的文献资源;
- ② 确认文献中或其部分中的著作;
- ③ 按照标准引文规则生成这些文献和著作的系统;
- ④ 提供系统有价值的检索点,至少包括题名,责任者和主题等;
- ⑤ 在文献单位或其他检索集合中提供某种评论资源的方式。

上述方面同样也适用于各种类型的信息资源的处理。

在目前的多数信息检索系统中,对信息资源的组织一般是在信息处理的基础上,以信息记录作为代替物实施的。信息记录是根据组织检索系统的需要对信息资源进行处理的结果,是对各种媒体信息进行多层次、多角度组织和揭示的基础。对信息资源处理的基本做法是:

① 对信息资源进行描述。包括:著录或编写题录,对信息资源外部特征进行描述,用于识别或确认该资源。

② 对信息资源的内容进行浓缩。包括编写内容提要或文摘,使用户能迅速了解其内容和特点,供检索查找时选择、判断。

③ 提供检索点。主要是确定检索特征及可检信息,以便据此对信息资源进行有序化组织。信息资源著录中描述的各种特征,如:责任者名、题名、出版机构名等,均可用作检索点。同时还可以在对主题内容进行分析的基础上确定相应检索标识,包括分类号和主题词。此外,在计算机检索系统中,文摘和提要中的语词亦可以作为检索依据。

这一信息处理形成的描述结果,记录了信息资源的基本数据,是组织检索系统的基本单元。将众多的描述记录纳入系统,并依据各种检索点提供检索途径,就构成了相应资源的检索系统。

从上面的分析可以看出,信息组织活动包括两个基本的方面:

① 信息组织的操作。指根据信息资源的内容和特征,对信息资源进行描述、标引,并将其组织进相应系统的过程。通常是依据一定的信息资源描述处理的规范,对信息资源的特征、内容等进行分析、归纳、选择、记录,并在此基础上将其以一定形式纳入信息检索系统。其中,信息描述通常应与系统处理的要求一致,一般应遵循一定的描述或著录的条例进行操作;对主题内容的组织和揭示,则需要在对信息资源的主题内容进行分析的基础上,以一定的类表或词表为工具,将分析出来的结果转换为相应标识,以便可以在需要时供用户匹配查找。

② 信息资源组织规范。这是根据信息组织的需要,按照信息资源的特点预先确定的

* Hagler, R., The bibliographic record and information technology, 2nd edn, Chicago, USA, American Library Association; Ottawa, Canada, Canadian Library Association, c1991.

规则和方法系统,是进行信息组织操作的依据。包括:

信息描述规范。描述规范是对信息资源特征进行描述记录的依据。通常根据特定信息资源的特点和识别、组织的需要加以规定。一般确定需要记录的项目、记录的文字、次序、详略程度等。以便可以依据它,以统一、有效的方式对信息资源加以记录。

检索语言。检索语言是一种由表达信息资源主题概念及其相互关系的词汇及规则组成的人工语言系统,国内亦称为情报检索语言。它是根据检索系统的特点,按照使用需要和对知识关系的理解预先组织好的一种信息资源组织方法的体系,是信息组织方法的具体体现。检索语言由词汇和语法两部分组成。其中,词汇是检索语言的主体,主要用于对所使用的词汇及其相关性进行规定,是信息资源存储和检索的语言保障。其表现形式是各种词表,包括分类表或标题表、叙词表等。它们是根据一定的知识结构,预先建立的标引系统。语法,则是分类标引和主题标引的规则系统,它们规定词汇集的使用方法,是文献标引时必须遵守的规范。自由标引与检索属于自然语言应用,但在一定程度上仍需要遵守一定的语法规范。

此外,信息组织规范还应包括对信息资源及检索工具的各种基本组织、检索的方法等。

显然,上述以信息资源为对象,通过对其内容特征等的分析、选择、标引、处理,使其成为有序化的过程的活动,从本质上讲,是一种知识组织活动,这一过程中使用的依据,如检索语言等,实质上是将知识结构引入信息资源组织的工具或规范。

信息组织作为以信息资源为对象的组织和揭示的方法,长期以来一直受到人们的重视,并且是与信息资源的状况和现代信息技术的发展密切联系的。在目前情况下,信息组织的研究领域包括下述内容:

① 信息组织基本理论方法的研究,包括信息组织的基本任务、要求、方法的研究;描述控制、词汇控制、句法控制研究;相关知识门类如知识分类、概念逻辑、语言学等与信息组织关系的研究;

② 信息组织的发展及其类型的研究;

③ 信息描述规范的基本理论方法研究,包括信息描述规范的原理、类型、特点、发展等的研究,数据编码与通用标识语言建立规律等的研究;

④ 信息描述理论和技术的研究,包括各种信息资源描述方法、数据编码、通用标识语言使用方法等的研究;

⑤ 类表、词表的基本理论及方法研究,包括:结构原理、性能、特点,编制及管理的技术方法等的研究;

⑥ 各种类型类表、词表研究,包括各种综合性类表、词表研究,专业性类表、词表研究等;

⑦ 标引理论与技术的研究,包括各种标引方式、类型研究以及各种文献、主题类型标引方法的研究等;

⑧ 信息检索工具研究,包括各种手工检索系统、计算机检索系统的构成、技术方法等的研究;

⑨ 计算机技术在信息整序领域应用的研究,包括词表的自动编制、自动赋词标引等研究;

⑩ 自然语言在信息检索中应用的研究,包括各种自然语言标引和检索的方法,汉语自动标引和检索的方法等的研究;

⑪ 各种专门类型资源的组织和处理研究,包括图像资源处理与组织,音频资源的处理与组织,多媒体资源的处理与组织等的研究;

⑫ 信息组织在联机系统和网络环境下的应用研究。

从上面信息组织领域包括的基本内容可以看出,信息组织的研究领域总体上包括三个层次,依次为,理论层面、规范层面、方法层面。理论层面主要研究信息组织的基本规律和理论基础,为规范层面和方法层面提供理论依据,指导信息组织的操作和实施。规范层面主要通过描述规范、检索语言的建立,为信息组织的操作提供实施的规范和依据。使用它,可以按照规定的方式,一致、有效地进行描述和标引,是进行信息组织必须共同遵守的约定和规则系统。方法层面包括操作的技术方法、信息组织的管理、标引规则、描述规则等,是实现信息资源组织的具体技术方法,其作用是按照信息资源组织的要求进行操作、实现信息资源系统的组织。三个层面之间是相互联系、互相作用的。应在对信息组织了解的过程中有意识地加强对三个方面的把握和结合。

本章思考题

1. 什么是信息组织? 信息组织与信息检索有什么关系?
2. 信息资源组织有哪些类型? 什么是文献目录、索引?
3. 按照标识特征和提供的检索途径,信息组织可以分为哪两种类型? 从内容角度标引和检索的方法有哪些?
4. 信息组织过程中对信息资源的描述处理有哪些要求? 基本做法是怎样的?
5. 信息资源组织一般应包括哪些方面的规范? 什么是检索语言?

第二章 信息组织原理

第一节 对信息组织的要求

信息组织的目的是为了有效地检索和利用。一般认为下述因素是一个信息组织系统必须考虑的重要因素。

1. 检全率

亦称查全率、命中率,指系统实施检索时检出的与某一检索提问相关的信息资源数与检索系统中与该提问相关的实有信息资源总数之比,可以表示为:

$$\text{检全率} = \frac{\text{检出相关信息资源量}}{\text{系统相关信息资源总量}} \times 100\%$$

上述表达式中,检出相关信息资源量为检出的信息资源总量排除其中不相关信息资源数量后得到的值;系统相关信息资源总量则为已检出相关信息资源数与未检出相关信息资源数量之和。如一数据库中共有 40 篇相关文献,所实施的检索检出其中的 30 篇,则该检索的检全率为 30/40,即 75%。检全率表示检索系统检出相关信息资源的能力,是评价检索系统效率的一个重要参数。较高的查全率是检索系统的基本目标之一。

2. 检准率

亦称查准率、相关率,指系统实施检索时检出的与某一检索提问相关的信息资源数与检出的信息资源总数之比,可以表示为:

$$\text{检准率} = \frac{\text{检出相关信息资源量}}{\text{检出信息资源总数}} \times 100\%$$

仍以前面提到的检索为例,如该检索的实际检出信息资源数量为 45 篇,则其检准率为 30/45,即 66.7%。检准率表示检索系统排除与检索提问无关信息资源的能力。提高检准率可以节省用户分离无关信息资源所花费的时间,对提高检索系统的使用效率具有重要作用。因此,一般均采用各种措施,力求达到较高的检准率。

在实际检索系统中,检全率与检准率的计算当然远比上面的说明复杂。在许多情况下,要明确判断一信息资源是否与检索提问相关,以及在多大程度上符合检索要求并不容易。但一般都同意,检索系统的检全率与检准率之间是一种互逆关系。因为,检全率高,必然会检出一些内容关联程度较低的信息资源,从而影响系统的检准率;反之,提高检准

率,则要求排除与检索提问关联程度较低的信息资源,必然会反过来影响系统的检全率。二者的关系如图 2-1 所示。如何使检索系统保持适当的检全率与检准率,以及允许用户根据检索需要选择和调整系统的检全率和检准率,是保证检索系统效率的重要内容。通常应根据检索系统的特点及用户的需求确定。相对而言,规模有限的中小型检索系统对检全率有较高的要求,对检准率则并不如大型检索系统那样重要;但在信息资源数量较大的数据库或搜索引擎中,随着检出文献量的增加,对检全率的需求往往退居其次,检准率成为主要的追求目标。例如,在网络检索中,数据检出量有时往往在数千条以上,在这种情况下,对普通检索用户而言检全率的高低几乎可以忽略不计;如何保证较高检准率,有效揭示符合用户需要的信息资源,成为检索的关键。此外,不同的用户对象,在不同的阶段对检索需要也是不同的。因此,检索系统对检全率与检准率的确定,应考虑检索系统的任务、服务对象的需求、信息资源集合的规模等多种因素,并根据使用需要提供选择的可能性。这样,通过将检全率与检准率结合使用,就可以较好改进系统的使用效果。

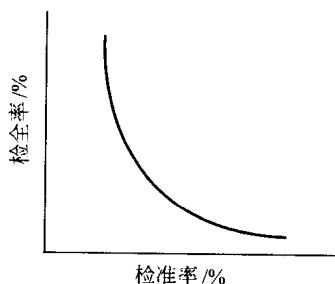


图 2-1 检全率、检准率反比关系示意图

3. 检索速度和系统的及时性

这是涉及信息资源组织时间和检索时间的两个指标。检索速度指用户实施检索时获得检索结果花费的时间。检索速度快,可以及时检出所需要的信息资源,直接影响到检索系统的使用效果。这一指标不仅与系统响应检索提问的速度有关,同时也涉及查全、查准等因素,与信息资源的组织是否合理直接有关。系统的及时性指检索工具能否在较短时间内将信息资源纳入系统,以较快速度提供使用。为了提高信息组织的质量、缩短检索时间,检索系统通常需要花费一定时间进行标引处理。但处理时间过长,也会影响系统的效率,特别是一些时效性较强的系统,例如一些通报性的系统、新闻系统等。对于一些特殊的系统,应根据其特点,对信息资源的组织方式加以改进或调整,包括结合采用各种自动标引方式等,力求在保证信息资源处理和标引质量的同时,缩短处理时间。

4. 系统的易用性

对于信息资源的处理和标引人员而言,指系统是否便于进行描述和标引操作,是否需