



郑 守 瑾

主 编

计算机检索 与 国际联机

西南交通大学出版社

计算机检索与国际联机

主 编 郑守瑾

副主编 陈敏英

西南交通大学出版社

1990年 成都

内 容 提 要

本书为教材系列参考书,供高等理、工、医专业本科生、研究生上文献检索与利用课实习时作为指导,亦可作为自学者的参考。对于从事科技工作和科技情报工作的各专业人员 and 教学人员,都有阅读和参考价值。

本书详尽介绍了电子计算机在科技文献检索中的应用。把国际上具有权威性的联机检索系统的特征、收录范围、文档的数量及名称、各种常用指令都一一作了概括的说明,尤其对 DIALOG、ESA-IRS、BRS-ORBIT、STN 等世界大型检索系统以及一些常用文档,更作了全面叙述。为使读者便于应用,在许多节段里,补充了必要的实例。

计算机检索与国际联机

JISUANJI JIANSUO YU GUOJI LIANJI

郑守瑾 主编

*

西南交通大学出版社出版发行

(四川 峨眉山市)

四川省新华书店经销

西南交通大学出版社印刷厂印刷

*

开本: 787×1092 1/16 印张: 13.375

字数: 334 千字 印数: 1—5000 册

1990年7月第1版 1990年7月第1次印刷

ISBN 7-81022-127-2/T 044

定价: 3.05 元

主 编	南京铁道医学院	郑守瑾
副主编	西南交通大学	陈敏英
编 委	北方交通大学	姜淑媛
	华东交通大学	宋如億
	上海铁道学院	汪碧雯
	太原运输管理干部学院	单宝昆
	兰州铁道学院	贾桂荣
主 审	刘基唐	

前 言

自从1986年底中国铁道出版社的《科技情报学》出版发行以来,电子计算机检索的应用和推广发生了很大的变化。不少读者指出:“《科技情报学》所介绍的第四章在实践中显得很不够用”。尤其是研究生所做的课题,不少需要利用计算机进行检索,而他们对这些内容又不太熟悉,因此深感迫切需要一本实用的教材参考书。为此,我们在1988年的铁路高校科技情报网乌鲁木齐会议上作了决定,争取在一年内完成《计算机检索与国际联机》一书的编写工作,作为系列教材参考书,并成为铁路高校科技情报网共同合作的第二本著作——《科技情报检索指南》的姊妹篇。前者以手工检索为主题,而后者则以电子计算机检索为中心来进行探讨。

国际联机检索发展日新月异,不仅文档千变万化,而且系统增添也时有所闻,原有的资料显得已不能适应当前形势的需要。尤其考虑到本书的读者是高等学校的师生,从事科技研究工作的专家、学者以及其他对计算机检索有浓厚兴趣的科学家们,他们对知识的需求具有层次高、内容新、深度广度兼要等特点。因此,给我们的撰写工作增添了一定的难度。为了尽可能达到上述要求,我们在编写过程中,参阅了国内外有关联机检索的近期杂志和参考书,请教了从事联机检索工作多年的科技情报工作者。所以说本书是集体智慧的合作结晶,也是使用电子计算机检索的科技工作者的良师益友。

计算机检索所涉及的范围极广,由于篇幅有限,本书只重点介绍理、工、医等学科有关的专业部分,即使如此,也只是其中有限的一部分。当今世界科技专业分工之细,文献拥有量之多,磁盘存贮还感困难,区区数十万言之书,更无法包含于其间,只有请同行及读者们见谅。

我们在编写定稿过程中,中国情报学会学术委员会副主任委员、中国铁道学会情报委员会主任委员、铁道部铁道科学研究院科技情报研究所刘基唐研究员亲自给予指导和帮助并主审全稿,铁道部科技情报研究所张连友所长为本书的编写和出版、发行都给予多方面的帮助,铁道部教育司、科技司、铁科院以及作者们所在单位的各级领导,都对本书的撰写给予了极大的支持,使之早日与读者见面,我们编委会全体同志,在此向他们一并致谢。

全书共分十章,铁路与非铁路高校均可适用。参加本书撰写的有:北方交通大学姜淑媛,华东交通大学宋如憶、张连英,上海铁道学院江璧雯,太原运输管理干部学院单宝昆,兰州铁道学院贾桂荣,南京铁道医学院郑守瑾和西南交通大学陈敏英、朱文倩、王改云、卢荣等同志。由于编者才疏学浅,误谬之处还请前辈同道们批评指正,以便再版时修改。

郑 守 瑾

目 录

第一章 电子计算机检索概述.....	1
第一节 电子计算机检索的历史及其沿革.....	1
第二节 计算机检索的用途及优缺点.....	2
第三节 计算机检索的必备条件.....	3
第二章 电子计算机检索的数据库.....	5
第一节 数据库与数据库管理系统.....	5
第二节 数据库的建立、选择与发展.....	7
第三节 联机联网.....	10
第四节 国内数据库.....	12
第三章 国际联机情报检索.....	15
第一节 基本原理与途径.....	15
第二节 检索提问标识切题与匹配.....	20
第三节 联机服务方式.....	27
第四节 联机检索应知事项.....	30
第四章 计算机检索技术与手段.....	33
第一节 情报需求与课题分析.....	33
第二节 检索单元的确定及逻辑组配.....	38
第三节 选择数据库.....	44
第四节 联机检索的信息反馈.....	48
第五章 美国 DIALOG 国际联机情报检索系统.....	53
第一节 DIALOG 系统简介.....	53
第二节 DIALOG 系统数据库文献记录的标引方式.....	54
第三节 DIALOG 系统常用指令.....	56
第四节 DIALOG 系统常用运算符.....	63
第五节 DIALOG 系统联机检索的基本过程.....	65
第六章 美国 DIALOG 系统第二版本简介.....	69
第一节 检索单元的限定查找.....	69

第二节	常用运算符简介	75
第三节	主要辅助检索功能	79
第四节	检索结果的输出	82
第七章	美国 ORBIT 与 BRS 国际联机情报检索系统	84
第一节	美国 ORBIT 系统	84
第二节	美国 BRS 系统	92
第八章	欧洲空间组织 ESA—IRS 系统	103
第一节	ESA—IRS 系统的数据库	104
第二节	文献记录结构	104
第三节	ESA—IRS 系统的基本指令及功能	105
第四节	ESA—IRS 与 DIALOG 系统检索功能的比较	110
第五节	ESA—QUEST 系统标准输出格式	116
第六节	检索实例	117
第九章	STN 国际联机检索系统	120
第一节	基本检索方法与途径	121
第二节	基本指令与打印格式	122
第十章	有关数据库联机检索	129
第一节	CA Search 数据库	129
第二节	世界专利索引 (WPI) 数据库	137
第三节	PTS 数据库	148
附录 1	DIALOG 系统数据库主题分类目录	155
附录 2	ORBIT 数据库目录	175
附录 3	BRS 系统文档按主题分类表	179
附录 4	按专业归类的 ESA—IRS 数据库	189
附录 5	STN 系统数据库	193
附录 6	《WPI》的有关代码表	203
附录 7	PTS 数据库在几个国际联机检索系统中的文档	205
参考文献		206

第一章 电子计算机检索概述

第一节 电子计算机检索的历史及其沿革

电子计算机技术应用到文献检索上来, 仅仅是近四十年的历史。由于科学技术的高速发展, 文献量也伴随剧增。以物理学为例, 据有关资料统计, 该领域内的新知识 90% 都是 1950 年后为科学家们所发展起来的。从 20 世纪 50 年代初期开始, 美国就将电子计算机技术应用到科技文献检索上来。早在 1954 年, 美国海军军械中心 (NOTS) 利用 IBM—701 型电子计算机建立了世界上第一个科技文献检索系列。随着技术的迅速发展, 又经过了十年的努力, 美国国立医学图书馆 (National Medical Library) 建立了以医学为主的文献检索系统, 即 MEDLARS (Medical Literature Analysis and Retrieval System)。直到今天该系统仍为世界上最权威性的医学生物学文献数据库。此后, 1963 年起美国洛克希德导弹和宇航公司的情报单位研制成功人机对话而命名为 DIALOG 系统。1981 年 6 月起独立经营。随着电子计算机处理技术的进步, 磁盘和磁盘机技术的出现, 70 年代初期, 世界各地的电子计算机检索系统相互联机联网, 从而使单台电子计算机通过通信线路连接着众多的检索终端; 利用分时技术使用户可直接与数据库对话。现代通信技术已经应用到情报数据的输送传递上来, 使网络检索系统的建立成为可能。每个节点 (即大型情报检索系统的电子计算机) 又可连接许多检索终端, 各个节点之间也可利用各种通信线路彼此连接, 彼此纵横交叉, 在世界范围内提供了能相互利用情报资源的检索网络。依靠这种技术和设备, 才能为情报资源共享建立了物质基础。

电子计算机检索的发展与空间技术的发展相结合而形成的国际联机情报检索系统, 是情报——卫星通信——电子计算机三位一体的新阶段。这个系统大大缩短了空间与时间的距离, 冲越了洲界与国界, 用人工检索方法需要几个月、几年、几十年、甚至几代人方能完成的工作, 如果利用电子计算机国际联机去进行, 则往往可以减少到几分钟、几十分钟, 就可以浏览当今世界数以千万计的有关期刊资料、学术论文、会议记录等名目繁多的情报信息。电子计算机检索的成功运用, 使快速获取全球性科技情报资料成为可能。这对于人们能及时继承并发展前人的科学技术、发明创造, 继往开来地使社会进步, 确已起到了难以估量的积极推动作用, 是情报科学历史上的又一里程碑。但是, 到目前为止, 即使电子计算机检索的重要性已达到了如此的高度, 仍然不能完全替代手工方式的传统检索法。两者可彼此取长补短, 为人类提供优质服务。

进入 80 年代以来, 世界上国际联机已经达到了相当大的规模。联机网络和检索终端遍及美、澳、欧、亚、非五大洲的各国主要城镇。数据库有几千个 (仅 DIALOG 检索系统就有数据库近 400 个), 检索终端在千万台左右。现在正在研究利用普通家用电视机来作为声象数据情报的检索终端, 用以实现计算机检索普及化的设想。每个家庭用户都可以利用自己

的电话机、电视机和转换器来完成此项情报服务。根据现在已有的设备和软件,科学家们预测:到21世纪的早期,人们将通过通信卫星去连接全世界各地为数众多的数据库群,使人类的绝大部分知识财富都贮存到电子计算机网络相连的数据库群中去。用户只要坐在家用电视机显像管前,拨用电话号码便可摄取所希望得到的各种信息。中国有句话叫做“秀才不出门,能知天下事”,那时候方可真正成为现实。

我国计算机检索的发展也是十分迅速的,除了我们自己正在逐步建立的各种类型数据库外,1980年中国建筑科学研究院情报研究所等机构联合在香港中国海外建筑公司内,装置一台DTC—382型检索终端,通过香港大东电报局直接连接国际通信卫星TYMNET或TELENET网络,与美国CALIFORNIA州的DIALOG系统或ORBIT系统或欧洲空间组织的ESA—IRS系统进行联机,迈开了国际联机检索服务的第一步。1983年,北京中国科学技术情报研究所与意大利的有关部门达成协议,开展中意远程数据通信业务,建立了数台传输速度较高的数据终端(原为300波特,近几年来已调增到1200波特)。同年9月,正式为ESA—IRS系统联机并开展检索服务。除以上三大系统可供检索服务外,近年来美国BRS系统及联邦德国STN系统也为我国提供服务。1989年起BRS系统与ORBIT系统实行合并。STN系统对我国用户实行优惠,该系统自1983年9月建立以来,发展十分迅速,已被誉为世界第二的联机检索系统,也是世界上第一个可以显示图象的数据库,并可实现全文检索,尤以物理学文档更具有特色。原来专业范围较窄,现已开放到遍及各学科的文档。该系统有三个服务中心(美国CAS IN COLUMBUS;联邦德国FIZ KARLSRUHE;日本JAICI),彼此通过卫星互相连接。德美之间有56 k BIT/秒和9.6 k BIT/秒的两条通信线路,德日之间有一条9.6 k BIT/秒的通信线路。任何一个终端用户通过任一个服务中心网络接口同样可使用任一个数据库,大大方便用户,既节省通信费用,又能较快提供脱机打印。STN系统的检索功能已达到世界领先地位。我国情报部门提出应尽量多使用该系统,争取二三年内能与DIALOG基本持平,以期消除当前主要只靠一家检索的局面。

第二节 计算机检索的用途及优缺点

计算机检索的用途十分广泛,除可作为文献检索外,还可用来作定题服务(SDI),订购原文资料,专利权的肯定与否定,情报监护,科研课题追溯与咨询等。计算机检索与手工检索都有各自的优缺点。一般说来,计算机检索的优点,往往是手工检索的缺点;反之亦然。如前者检索速度快,耗时少,查阅范围广,甚至可查到国外刚刚出版的期刊以及非订阅读物(如论文集、会议录、汇编),检索内容的专指性强(如可专查疾病的治愈率或死亡率,各种金属的熔点,各种放射性元素的半衰期等),但追溯时间受一定限制(目前正在逐步向前推移收录年代),检索价格昂贵(不同文档收费标准不一,联机按占用机时计算,脱机则按文章篇数计算),检索时间也有一定限制,国外每周工作五天,星期六星期日两天停机。此外,国际联机要经过众多的环节,如人造卫星地面发射接收,中美间通信要经过欧洲罗马地面中转站,任何一个环节的故障,都可影响检索效果。再加上电子计算机工作必须具备的保证条件,都可左右结果。

手工检索必须具备丰富的馆藏(包括各种工具书及期刊),但往往由于图书资料不足,即使在检索工具书中找到了有关的题录及期刊名称,也无法获得原文献而不得不再委托机检

去订购原文资料或文摘。完满的馆藏可收藏到上个世纪的有关文献,但在数据库里目前还无法做到。

数据库的发展方向当前已初显前景。全文数据库已经开始建立,图象检索也已成为现实。从现在各国数据库发展趋势来看,我国的联机联网也不会是遥远的未来。中国科技情报研究所已在 IBM—4381 机上利用 CDS/ISIS 软件建立了约有 220 万记录的包括 INSPEC (1980—1987 年)、COMPENDEX (1986—1988 年)、Ei—MEETING (1985—1988 年)、中国科技情报研究所西文资料馆藏 (1985—1988 年)及中国科技情报研究所西文期刊馆藏、中国学报论文及中国学位论文等数据库。另外在 VAX—11/750 机上利用汉化软件 TRIP 建立了中国学术会议论文数据库 (1987—1988 年)、中国科技期刊联合目录、适用技术及国家级科技成果等数据库。根据权威人士预测:科技文献数据库将向全知识数据库方向过渡,今后不论是自然科学还是社会科学,甚至其他一些与人们生活有关的知识 and 信息,都将会被情报学家们收录到数据库中来,这将大大方便人们的生活和享受。只要在检索终端旁边拨一下电话号码,想知道的东西一律得到满足。这可能就是真正进入到了全信息的时代。

第三节 计算机检索的必备条件

计算机检索当然要依靠计算机来完成,计算机就成为首要的必备条件。但是单靠计算机也是不可能完成检索的。一般说来计算机检索的必备条件有下列三个方面:

一、物质条件

从计算机检索的全过程来看,物质条件应该由三个部分组成,即数据库(包括电子计算机和软件)、通信系统和检索终端(包括微型电子计算机、调制解调器和民用电话机等)。随着技术进步而来的数据库更新速度也是十分惊人的。近十年来数据库的数量增加很快,平均每年递增 20% 左右。数据库的专业覆盖面几乎已涉及到所有的科技门类。仅以 DIALOG 系统为例,1979 年有数据库 109 个,总文献为 3 500 万篇,到五年后的 1984 年,数据库增加到 227 个,总文献为一亿篇左右。1985 年数据库为 267 个,1986 年为 294 个,1988 年又增加到 378 个,1989 年初达到 385 个,十年间几乎增加了 300%。设备更新也十分显著,1979 年主机使用的是 IBM—3032 及 IBM—3033 各一台,1984 年便使用 NAS—AS 9080 机两台。1985 年主机为三台 NAS—AS/XL 60 (内存 64 MB×3),速度为 28 MIPS。1988 年 8 月主机仍为三台,其中二台 NAS—AS/XL 60 (内存 64 MB×2),一台 NAS—AS/XL 80 (内存 128 MB),速度为 50 MIPS。

通信系统在联机检索中也占有举足轻重的位置。DIALOG 系统除与 TYMNET 及 TELENET 两大公共数据网连接外,还增加了一些其他网络,如日本和英国的专用线,美国 and 加拿大所用的 UNINET 数据网和 DIALNET 数据网。至于与终端间的联系,主要通过电话线路去完成,如果采用专线电话,则既可保证检索速度,又可维护检索质量。

现以国际联机某终端为例,说明联机的物质条件:

机房一间(最好配有套间,作为办公室并可存放检索资料档案):地板为木质或铺以地

毯, 装有空调设备, 使温度保持在 $10\sim 25^{\circ}\text{C}$ 范围内。温度过高或过低, 都可能发生误码而影响检索质量。

直通或专线电话一门。

微型电子计算机一台 (IBM—PC/XT 即可)。

调制解调器一台 (Modem VI—2242 PA—300~2400 波特)。

打印机一台。

显示器一台。

软盘若干。

为防止某些地区突然发生计划外停电事件而致输入的检索策略全部消失, 可配置不停电电源设备一套, 以应急需。

二、人员素质

计算机检索的质量虽然与物质基础密切相关, 但与人员素质的好坏也是不可分割的。单位内部利用计算机检索, 则与文献资料的标引质量休戚相关。标引人员除应具有高度的责任心外, 既要熟悉专业知识及标引准则, 又要具有较高的外语水平 (采用汉语标引例外)。国际联机检索虽不必考虑标引质量, 但仍然需要懂得文章的标引方法, 否则就不能制订出较好的检索策略。素质好的检索人员, 首先必须把检索目的弄得一清二楚, 然后才有可能研究为达到目的而采用的方法与措施。最好在上机前对检索课题进行一次较充分的讨论, 起码编写出 2~3 套检索策略。要考虑到第一套方案输进去, 检不出文章该怎么办? 检出的文章过多时又该怎么办? 检索者要学会扩展指令 (EXPAND) 的使用方法及查阅主题词表。在联机检索的全过程中, 检索者要全神贯注屏幕的各项显示, 防止输入和输出中发生错误。对于上机操作人员来说, 不可能要求他们精通各种专业的词汇, 即使外语水平很好, 打字速度很快, 也熟悉各系统各文档的使用方法和特征, 但输入时词汇错误还是可能发生的。只有靠检索者本人高度专注屏幕和敏捷的思维反应方可避免误谬并得到正确处理。

上面已经强调了检索者素质的方方面面, 这里还应当进一步强调对外语词汇的灵活应用。众所周知, 文字是随着科学技术和生产的发展而发展的, 目前数据库里贮存的词汇, 有一部分在正具书里是不容易找到的。换句话说, 也就是这些词汇是近年来人们在科学研究、社会实践中创造出来的。造词的原则也无非在原有词汇的基础上加以充实和发展。所以, 检索者应当学会剖析词汇, 熟悉外文的前缀、后缀和词干的组成与搭配, 以利在检索过程中发生意外时可以自行选词组配。

三、其他因素

室内环境的清洁卫生及附近电磁场、机械、化学因素的干扰都有可能影响检索质量, 因此在干扰严重时有必要采用屏蔽设备。机房地点的选择最好离交通要道的主干道远些, 也不要离铁路、机场、电厂、化工厂、汽车停车场、煤气厂、码头等处太近, 以免影响机房运转。楼层的选择也至关重要, 最好机房不设在顶层或最底层, 更不能设在地下室或半地下室。对于湿度较高的地区和机房, 应配备功率足够的吸湿机 1~2 台, 以备必要时使用。

第二章 电子计算机检索的数据库

第一节 数据库与数据库管理系统

数据库是电子计算机检索系统中基本的重要的物质条件。了解和选择数据库,是电子计算机检索的首要前提,也是影响检索效率的重要因素。因此,每个情报用户必须掌握数据库的类型和记录结构,才能根据需要进行合适的数据库。

一、数据库的概念及分类

数据库 (Database) 是指计算机存储器中合理组织、相互有关的数据集合。这些数据有最小冗余度 (数据重复最小), 可为多个用户共享, 数据的存贮独立于使用它的程序, 数据被统一管理, 可方便灵活地为用户使用。

联机检索的数据库也称为文档 (File), 它是以磁盘或磁带为存贮介质, 并按一定的存取方式编制而成的机读文献记录集合。目前, 全世界国际联机的数据库有 3 300 多个, 它们都是以赢利为目的的商用数据库 (Commercial Database)。

数据库可以根据各种标准进行分类, 基本上可以从以下四个方面划分:

1. 从数据库生产者的角度, 可分为政府数据库、专业协会数据库和商业数据库。
2. 从主题覆盖面的角度, 可分为大学科数据库、多学科数据库、交叉学科数据库、小专业数据库和特种数据库。

3. 从专业角度, 可按不同类型划分, 几乎每个专业可分为一种类型。

4. 从检索角度, 可分为文献参考数据库和源数据库二种。

目前情报界主要是从检索角度, 根据数据库所含信息内容作为基本的分类标准。根据这个标准, 数据库可分为:

(1) 文献参考数据库 (Reference Database), 这主要是二次文献数据库。包括:

① 书目数据库 (Bibliographic Database), 它包括各种机读版的文摘、目录、索引等。

② 指示数据库 (Referral Database), 其内容包括可作为情报取得来源的机构、计划、活动, 乃至有特长的个人介绍。

文献参考数据库本身并不直接提供用户所需情报, 其价值在于指引情报用户找到合适的情报源。

(2) 源数据库 (Source Database), 亦称非书目数据库, 其存贮内容为数值、结构式、原始文献片断或全文等源情报。包括:

① 数值数据库 (Numeric Database), 包含来自原始文献或调查统计的并经过处理的各种数值或数据表格。

② 文本——数值数据库 (Textual Numeric Database), 是指其记录同时包含文本信息字段和数值信息字段的数据库。许多名录型的数据库就属于这种类型。

③ 属性数据库 (Properties Database), 是指词典或手册式的化学、物理数据以及其他数据。

④ 全文数据库 (Full-Text Database), 是指机读化的文献全文。例如法律条文、文学作品作为数据全部加以存贮, 并能用来检索全文中任何一个字、段、节、章等等。

源数据库的特点在于本身含有一次情报。例如社会科学统计数据、化学命名与结构式、公司财政信息、存货市场报价、人口统计数据、文献全文、新闻消息等, 都可以直接满足用户作为检索目的而要求获取的数据、事实或文本的情报需求。与文献参考数据库只提供文献书目线索相比较, 具有更大的直接用途。正因为如此, 源数据库的增长要比文献参考数据库快得多。据统计, 1982 年前全世界联机数据库中, 源数据库占 52%, 到 1987 年, 全世界 200 多个联机情报检索系统中提供公众使用的 3 300 多个商用数据库, 91% 是源数据库。

当前, 由于数据库品种的多样化, 也引起了数据库分类概念的新变化。一种新的划分方法把数据库分为:

(1) 文字型数据库。包括文献型、事实型、百科全书型、词库型和全文型等数据库。

(2) 数据型数据库。

(3) 图象型数据库。例如 STN 系统西德专利数据库 (PATDPA), 是目录、全文、图象三结合的世界第一个图象显示库。

这种新的分类概念, 反映了图象型数据库日趋发展的局面。

二、数据库的记录结构

数据库的记录结构同图书馆中的索引卡片相似, 每一条记录相当于一张卡片, 含有与卡片大致相同的若干著录项目。这些著录项目叫数据项 (或数据元)。每一个数据项中所含真实数据叫数据值。

数据库中一条记录的常用字段及内容包括: 题目 (TI)、登录号 (AN)、作者 (AU)、期刊名称 (JN)、出版年 (PY)、语种 (LA)、文摘 (AB)、叙词 (DE)、识别词 (ID)、分类代码 (CC) 等。在不同的联机检索系统中, 记录的著录项目各不相同, 完全由系统使用的软件所决定。

记录中的题目、作者和叙词是关键数据项。联机检索主要通过上述项目查找有关记录, 每一项为一个查找点。一篇文献, 查找点越多, 命中率越高。记录中的时间、出处、分类号、语种、图别等是副关键数据项, 在联机检索中起到限定控制作用。

三、数据库管理系统 (Data Base Management System 简称 DBMS)

在数据库建立的过程中, 所使用的算法称之为数据库方法。数据库方法主要取决于选择什么样的数据关系结构来构造数据之间的联系, 即选择什么样的数据模型。目前常用的数据关系模型有三种: 层次模型、网状模型和关系模型。对应这三种模型, 可以建立层次数据库、网状数据库及关系数据库。

数据库的信息内容与结构, 关系非常复杂, 要求对这些数据进行操作也是多样的。如数据插入、删除、存贮、查询、安全保护、数据流控制、供多用户同时使用的分时操作、统计

计算等,这就需要管理这些数据的专用软件系统,即数据库管理系统(DBMS)。数据库管理系统是用来管理数据库中数据的一组复杂的软件。目前国际市场中的DBMS软件产品种类很多,有DMS、FRAMIS、INGERS、dBASE、INQUIRE等,并且都具有相应的数据模式和适应机型。

DBMS主要功能是,实现对数据库的描述、管理、维护等。DBMS主要由以下几个部分组成:

1. 语言处理部分。主要包括模式数据描述语言(DDL)翻译程序,子模式数据描述语言(SDDL)翻译程序,数据操作语言(DML)处理程序以及物理数据库描述语言(PDDL)处理程序等。

2. 系统运行控制部分。包括对整个DBMS进行有效控制和协调的总控程序,访问控制程序,并发控制程序,保密控制程序,数据访问及通信控制程序等。

3. 系统建立、维护部分。包括装入程序、监督程序、恢复程序等。

第二节 数据库的建立、选择与发展

数据库建设是一种知识、技术、劳动力密集型的产业,涉及数据、设备、技术、人员各个方面。建设标引质量高、出版速度快、搜集资料全、覆盖年限长的数据库,才能适应科技情报检索日益发展的需要。

一、数据库建设的基础工作

数据的预处理工作是数据库建设的基础,难度大,费时费工,工作庞杂。数据预处理工作包括:

(一) 数据的采集

数据的采集应当做到“全”、“快”、“精”、“准”、“新”。如果数据源不全或支离破碎,特别是专业资料搜集的全度、时间覆盖的长度和及时性差,将会影响数据库的使用和信誉。在当前形势下,缺乏竞争性的数据库将难以生存下去。

(二) 数据的评价

数据的可靠性和准确性是数据库的生命,数值库与事实库尤为如此。要通过数据评价从大量的数据源中精选、优化、鉴定有用数据。没有数据评价工作就不可能真正建立起适用的数据库。

(三) 数据的加工整理

建立规范化、标准化的数据库,必须实行著录与标引工作的标准化,这对数据库的质量至关重要。此外,还有数据本身的标准化问题,这也是实现资源共享的必要条件。美国三个大图书馆的信息网OCLC、WLN和RLIN,从1980年初试图连接起来,但至今未达到目的。原因不在于网络技术,也不在于硬件、软件和记录格式,而在于三个网络内原始数据的标准不一致,规范文件不统一。这种数据标准不统一,涉及到千千万万原始数据,不是由软件进行转换所能解决的,必须在建库之初就进行数据的规范与统一。目前,我国颁布了《文献著录总则》、《普通图书馆著录规则》、《连续出版物著录规则》、《档案文献著录规则》、《汉语主题词表标引手册》以及《铁路机读文献著录规则》等,这是我们建立规范化、标准

化数据库的重要条件。

建立数据库需要有相当数量的专业人员和专业技术人员。数据的筛选、标引、输入以及国内建库解决汉字处理问题,都需要训练有素的专业人员。因此,抓好人才培养是建库的根本保证。

二、数据库的选择

我国与 DIALOG、ORBIT、ESA—IRS、BRS、STN 等系统进行国际联机检索,涉及数据库 700 多个。选择数据库是国际联机检索的重要环节,是决定整个检索过程成败的关键。选择是否合理正确,取决于三个因素:

1. 对不同的联机检索系统的数据库掌握的程度,包括对数据库的规模、内容、质量、更新周期和覆盖年度的了解。

2. 通过课题分析判断可利用数据库的准确性。

3. 对数据库了解同对用户提问了解进行匹配的能力。

在通常情况下,选择数据库应从以下几个方面考虑:

1. 所需情报类型。不同类型的联机检索数据库收录的文献类型不同。有的只收录论文,有的只收录标准,有的只收录专利,也有收录多种类型的文献。因此,要弄清所需情报类型,因题而异,选择相应的数据库。

2. 数据库的特征。每个数据库都有其本身特定的主题范围,在类型、规格和内容上各有差异,构成各自的特征。例如 DIALOG 系统中涉及化学化工方面的数据库有 20 多个,每个数据库都有其特点和用途,其检索效果各不相同。如果查询有关立德粉颜料的市场情况,涉及商业经济和化工专业两方面的内容。只查化工(CA SEARCH 化学文摘)或只查经济(PTS PROMT PTS 世界工业产品市场与技术概况)都不准确。而 DIALOG 系统中 CHEMICAL INDUSTRY NOTES(化工札记 19 文档)的主题范围为化学化工方面的市场数据、投资、产品价格、政府和社会、组织和研究机构、产品和生产工艺等化学经济和化学新闻。查该数据库的结果符合课题本意。只有将情报需求的课题内容同数据库的主题特征匹配起来,才能达到预期的检索结果。

3. 数据库的质量。数据库的质量主要包括数据库的时间范围、更新频度和标引、词表等内容。

4. 情报检索的目的。情报检索目的可分为普查型和查重型。普查型检索主要用于开题、立项和一般了解工艺技术水平 and 现状,以收集某一技术领域中的文献资料、国外某一学科研究发展动态等。查重型又称为新颖性检索,主要用于申请专利、申报成果,这类课题在选择数据库时除查询相应专业数据库外,还要查询专利数据库,以满足新颖性检索的要求。

5. 数据库费用。联机检索系统都是商用性的,各系统机时收费差别较大。例如中情所自 1986 年 10 月正式开通 STN 系统,该系统 FIZ Karlsruhe 中心的全部数据库从 1989 年 7 月 1 日至 1990 年 6 月 30 日对我国用户实行 20% 优惠。掌握这些情况也是我们选择数据库的参考依据。

选择数据库的最终目的有二个:

1. 提高选择数据库的准确性,确保检索结果的质量。

2. 防止重复或错误使用数据库,降低联机检索费用。

三、数据库的发展趋势

60年代以来,情报供需关系的变化和信息技术的发展,引起世界范围内科技情报体系和技术手段的改变,出现了以加工和提供数据库服务的现代化科技情报检索行业。这个行业的主要任务是对情报进行收集并加工为计算机可读介质,采用计算机和通信技术为全社会提供现代化情报服务。商用数据库投入使用后,数量增长极快。据报道,全世界商用数据库1973年仅有81个,1975年增长到386个,1980年1409个,1986年2505个,1987年竟达到3369个,为14年前的41倍多,其速度极为惊人。

当前数据库发展呈以下趋势:

1. 范围社会化。数据库范围从科学技术向经济、管理、社会生活等各领域扩展,情报需求从科研和管理向技术开发、经济管理、产品开发、市场开发等多领域扩大。情报用户从科研人员扩大到包括科技、管理、办公、企业、教育、家庭等社会的几乎全部阶层。

2. 加工深度化。传统情报工作主要加工和利用的信息对象是书刊和科技报告等成品型情报。今天,数据库加工范围则扩大到随时发生的、流动中的半成品型和一过性情报。加工精度从整册整篇书刊为单位发展到可按章节段落乃至具体数据为单位的细加工;从文字发展到声音、图象等各种介质、载体的加工。因此,书目型数据库已经不再是主要数据库类型,事实型、数值型、多介质数据库所占比重越来越大。

3. 流通商品化。早期由政府部门或学会、协会等非盈利机构生产、所有或资助的科技情报利用形态,全面走向商品化。数据库成为技术市场的重要商品。数据库同矿产需要勘探、开采、精炼、加工、运输才能成为各种产品一样,也需要通过各个信息过程进行加值。数据库的生产和服务,需要采用高成本的信息技术和大量人力,而加值的信息又为用户换来了时间、速度和金钱。数据库的供需双方都接受商品经济的约束,除公用性和福利性机构与系统外,义务和免费服务的时代将告结束。

4. 存贮微型化。数据库的存贮向高密度发展。目前已开始普及的高密度存贮只读光盘(CD-ROM),一片4.75英寸大小的光盘,可以存贮27万页资料,相当于1600片软盘。生产厂家通常把检索软件和资料录制在同一张光盘上,利用微型机可以存贮或检索一个小型图书馆规模的情报资料。据1988年统计,全世界已出版CD-ROM光盘数据库300种以上,许多数据库生产和服务机构已开始常规提供光盘数据库服务。

5. 生产产业化。数据库生产和服务的市场产值越来越大。1986年全世界数据库产值为65亿美元,数据库行业已经具有产业化的规模和地位。发达国家的数据库行业已按照其他行业的模式,实现生产、加工、流通、批发、零售的产业化,自给自足或者手工作坊式的生产方式,已经让位与社会化的大生产方式。

6. 功能战略化。数据库事业的发展和对社会各领域的渗透,提高了情报的社会地位。社会已经进入信息化时代,并把能源、物资和情报称为当代社会的三大资源。人类可以利用知识创造和再生无穷的物质财富。而数据库是控制和处理知识的现代化手段。因此,数据库发展水平越来越被看作与国家经济实力、科技成就、外交地位、文明程度等同样重要。数据库事业的社会功能出现了实质性的变化,数据库作为重要的战略资源受到国家和社会的重视。

第三节 联机联网

计算机技术和通信技术紧密结合,出现了计算机网络和联机情报检索,推动了计算机情报检索系统的深入发展。

一、计算机网络

计算机网络是用通信线路将分散在不同地点并具有独立功能的多个计算机系统互相连接,按照网络协议进行数据通信,实现资源共享(共享网络中的硬件、软件、数据库等)的计算机集合体。

计算机系统相互连接的方式有两种:一种是计算机间通过双绞线、同轴电缆、电话线、光纤等有形通信介质互相连接;另一种是通过激光、微波、地球卫星通信信道等无形介质互连。

计算机网络按结构形式分为集中型、分布型和环状型三种。随着计算机网络的不断完善和发展,又出现了从逻辑功能上把数据处理和数据通信分开的趋势。这种计算机网络是由数据处理网(亦称资源子网)和数据通信网(亦称通信子网)组成的两极网络结构。建立计算机网络有两个目的:一是以传递数据、收集和传递信息为目的;二是以相互共用计算机资源为目的。计算机网络的建立使各用户不受地理条件的限制达到资源共享,充分发挥各地的资源特点,避免重复投资。同时也使计算机的负荷分散,提高了可靠性和可扩展性。

世界上最早出现的计算机网络是分布在很大地理范围内的远程网络(Wide Area Network 简称 WAN),例如美国国防部高级研究计划局等研建的 ARPA 网,早在 1969 年就开始建立,至今地理上不仅跨越美洲大陆,而且通过通信卫星与夏威夷和欧洲等地区的计算机网络连接。由当初试验网的 4 个结点,发展到 60 多个结点。连接起来的计算机有 100 多台。70 年代中期,由于微型计算机和微处理技术的发展及对计算机间进行短距离高速通信的要求,另一种分布在有限地理范围内的计算机网络——局部地区计算机网络(Local Area Network 简称 LAN)应运而生。1975 年美国 Xerox 公司推出的实验性以太网(Ethernet)和 1974 年英国剑桥大学研制的剑桥环网是局部网络的典型代表。

国际联机检索使用的全球性通信网都是以报文分组交换技术为基础的数据通信网。所谓分组交换网(Packet Switching Network 简称 PSN),是指将传送的情报报文(Message)先化整为零,分解为长度较短的信息单位,称为报文分组或信息包(Packet),每个分组都加上传送所需的控制信息(如目的地址、源地址等)和检错信息。分组交换技术有两个主要优点,一是分组信息单位较短,当发送产生错误需要重发时,开销较小,而且出错率减小;二是各个分组可沿网内不同径路传送,根据各径路忙闲情况自动选路,这可以降低总延时。到达目的地后,再把各个分组合零为整,依顺序组装成原来的情报报文。

联机检索使用的分组交换网都是商用网(Commercial Network),要按照使用通信线路和传送的分组数量进行收费。用于联机检索的商用网有两大类:一类是公共分组网。如美国的 TYMNET、TELENET、UNINET,加拿大的 DATAPAC,英国的 IPSS,法国的 TRANSPAC,欧洲共同体的 EURONET 等。另一类是专用分组网。这是联机检索系统专门为自己建设的通信网。例如 DIALOG 系统的 DIALNET,ESA—IRS 系统的 ESANET 等。