自动化情报和联机系
统培训班教材之四

数据库

南京航空学院
藏书

科学技术文献出版社

数　据　库

《自动化情报和联机系统培训班》

教　材　之　四

（限国内发行）

编 辑 者：中国科学技术情报研究所

出 版 者：科 学 技 术 文 献 出 版 社

印 刷 者：中国科学技术情报研究所印刷厂

新华书店北京发行所发行　　各地新华书店经售

☆

开本：850×1168 1/32　印张：2.5　字数：63千字

1980年11月北京第一版第一次印刷

印数：1—7,320册

科技新书目：174—41

统一书号：17176·260　　定价：0.30 元

出 版 说 明

中国科学技术情报研究所、联合国教科文组织和法国科技情报局自1979年9月3日至28日在北京联合举办“自动化情报和联机检索系统培训班”，时间共四周。第四周（9月24日至28日）讲授的是数据库部份，由法国巴黎地区中央实验室主任H．维拉(HENRI．VIELLARD）教授和法国化学情报中心工程师M．J．C．波內（M．J．C．BONNET）先生主讲。这部份內容主要包括数据库的结构，功能和应用方法，并结合美国化学文摘服务社（CAS）和法国DARC系统的具体情况重点介绍了化合物结构式的描述，存储，检索的原理和方法。其中对CAS系统和DARC系统化学结构的编码方法作了比较详细的介绍。

本教材主要是根据讲课笔记整理，个别词句作了删节和修改。担任课堂口译工作的是中国科学技术情报研究所的李亚民和江秋明同志，参加教材整理工作的有杨俊林、陈治金、夏爱芳、陈伟娴、崔亨朱、贾同兴、武兆园、周鼎恒、张凤楼等同志，由武兆园、贾同兴、吴荣荣、葛葆森、周鼎恒、张凤楼负责编辑和审定。

编者1979.11


30303436

399116

目　　次

第一章　综　　述

一、情报系统中的数据库……（3）
1．情报的功能……（3）
2．情报的表示方法……（4）
二、数据的内容和分类……（9）
1．数据的内容……（9）
2．数据库的定义……（11）
3．数据的分类……（12）
三、数据库的功能……（13）
1．收集……（13）
2．标引……（16）
3．记录……（19）
4．检索……（20）
5．发送……（22）
四、应用数据库的方法……（23）
1．组织结构……（24）
2．办公设备……（25）
3．软件系统……（26）
4．网络问题……（26）
五、数据库的运行……（27）
1．建立数据库的组织和工作步骤……（27）
2．数据库的提供和使用……（29）
六、国际合作……（30）

第二章 化学数据库

一、化学情报……（33）
1．化学文献库和数据库……（33）
2．化学情报的特点……（35）
二、化学数据库的计算机执行过程……（39）
1．结构数据库的建立……（39）
2．结构数据库的检索……（54）
三、法国DARC系统的成就……（66）
1．DARC系统……（66）
2．DARC PLURIDATA系统……（68）

第一章 综 述

建立情报系统的目的在于能处理多种多样的信息，使能及时的准确的满足用户在科研、生产和教学等方面对情报的要求。数据库技术的产生和发展，为设计和建立新型的现代化的情报系统提供了一种崭新的手段。

科技一次出版物是人们从事生产斗争和科学实验的记录，它汇集着科技工作者和劳动人民对无数有用的事实、数据、方法、理论的运用和发展情况，代表了一定条件下科技的发展水平、动态和方向，它是科学技术情报基本的、主要的来源之一。一次出版物和二次出版物的出版时间，在一定程度上反映出一个情报系统的服务水平和服务质量。美国、英国和法国出版的九种出版物从编辑到出版的时间如表 1 所示。

表 1 亦可用图 1 表示，图中实线代表一次出版物从编辑到出版

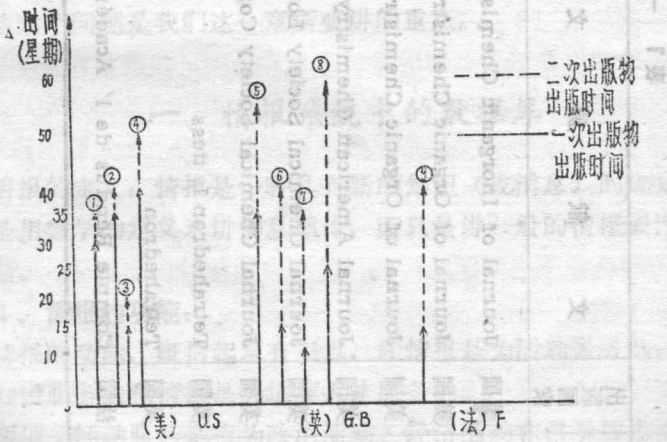


图 1 文献出版时间

表 1　一次出版物和做二次文摘时间

序号	出版国家	文献原文名称	中文译名	一次出版物编辑出版时间	做二次文摘时间
1	美国	Journal of Inorganic Chemistry	无机化学期刊	30（星期）	5（星期）
2	美国	Journal of Organic Chemistry	有机化学期刊	35（星期）	5（星期）
3	美国	Journal of Organic Chemistry Communication	有机化学通讯	15（星期）	5（星期）
4	美国	Journal American Chemistry Society	美国化学社期刊	40（星期）	10（星期）
5	英国	Journal Chemical Society London	伦敦化学社期刊	35（星期）	20（星期）
6	英国	Journal Chemical Society Communication	化学社通讯期刊	15（星期）	25（星期）
7	英国	Tetrahedron L ttess	四面体通讯	9（星期）	25（星期）
8	英国	Tetrahedron	四面体	25（星期）	30（星期）
9	法国	Compte Rendus de l′ Académie des Science	法国科学院报告	15（星期）	25（星期）

时间，虚线表示从一次出版物出版后到二次出版物（文摘）出版的时间。

从上面的图表可以看出，不管是一次文献的编辑出版所需时间还是二次文献的加工时间均是美国最短。编辑、出版和加工时间长短是非常重要的问题。例如，一九五五年，美国化学文摘社沒有做过一篇法国有关方面的文献，好像法国化学家沒做什么工作。实际上，这是由于法国文献编辑出版时间长，未能及时提供足够多的文献所造成的。

数据库技术的进步，为设计和建立高水平、高质量的情报检索系统创造了条件。我们将用信息分析的观点来介绍数据库的建立和利用方面的有关问题，主要内容包括：

情报系统中的数据库；

数据的定义和分类；

数据库的功能；

应用数据库的方法；

数据库的运行；

世界合作。

这六个问题是我们这一章所要讲的重点。

一、情报系统中的数据库

情报的定义：情报是一种具有新的知识（或消息）的成分。这里不是用数学的定义来讲信息理论，而只是讲一般的情报理论和实际问题。

1．情报的功能：

情报的功能，概括起来有三点，即情报是为活动服务的；情报是为知识服务的和情报是为培养人才服务的。

所谓为活动服务是指为政治活动、经济活动和科学技术活动服务的情报。

我们课程的重点是讲科学技术活动。虽然活动的范围属于技术方面的，然而所讲的原理对于其他方面的活动也是有效的。

掌握情报有助于做出决定。例如，通过了解情报分析全面情况就能决定安全和经济等方面的问题。教育方面的一个主要任务可作为从情报到活动之间的过渡，很遗憾，这一方面总是不易被人理解；很多人的脑子里有很多情报，但是不知道如何利用这些情报发挥作用，所以我的意见，从小学到大学的整个教育过程中应当告诉给学生情报和活动的关系，使得情报能在科技活动和生产活动中发挥应有的作用。

情报的第二个功能是为知识服务。人们希望知道自己所生活的世界里的情况，这个愿望促使人们去研究新的情况新的设想。一个人所需具备的知识可能是对整个情报领域的了解。科学技术的基本理论问题，也就是一个科技情报人员所了解的基本问题，这种知识的获得是通过多种方式（直接的或间接的）得到的。对于有些学科，主要是通过间接方式来获取。例如，天文的知识，对一个人来讲，不可能都是直接的知识，而是需要人们进行多方面了解情况，深入的进行科学研究才能得到的。

情报的第三个功能是为培养人材服务。人员培养的目的是建立有系统的情报存储并合理的进行组织和协调，以使情报能为未来的活动所利用。情报在培养人员问题上的作用应当是循序渐进的，也就是说重点在于训练学生形成科学的思想结构，在人脑中科学的排列和组织具有的情报。得到的情报不能从这方面和那方面任意的抛出去，情报本身有其一定的科学体系的，当头脑中形成这种科学体系后，那么在考虑问题的时候，就会有规律的顺序想象，就像化合物的原子排列有顺序的有序集合一样。

2．情报的表示方法：

然而学生除了应具有对情报有规律地排列并能在必要时加以发挥外，还要具有科学表示情报的能力。表示情报的物理方法有以下几种：

口头的方法（授课）；

报告、通讯的方法；

出版刊物的方法。

以下主要介绍出版的方法，重点讲科技文献的出版。

科技文献是科技情报的重要来源之一，因此，科技文献编辑出版工作是情报工作的重要组成部份。科技文献有它的基本结构，编辑人员通过阅读的方法选择质量好的材料，这些是文献本身的源材料。欧洲、美国、苏联的出版物，一般有一非常简单的、共同的计划。我对中国出版物的情况不太了解，我想，中国出版物的计划可能比较复杂、多样化。虽然你们的工作与出版物也有关系，但不是专门搞出版工作的，可能对出版物的计划的了解不完全清楚。不管是何种出版物，它的发表总有三步过程：

提出数据——→论据——→推导结论

出版物结构一般包括四个部份：

前言；

加以发展；

结论；

参考文献。

从以上四个方面来看那一方面能够发现情报呢？为了说明方便再回忆一下情报的定义“情报是一种具有新的知识（或消息）的成分，……。”按照这个定义去衡量各个部份是否具有情报。

前言：前言的目的是确定问题的范围。通过一个历史情况的说明（可能是一个作者的，也可能是几个作者的情况），但是一个作者的本意总是要表现自己，而不是为了表现其他作者。讲了作者的历史以后，前言中一般又告诉文献计划和方法。从整个前言的内容来讲，情报几乎等于零。因为作者的历史，文献的计划和方法可以在其他出版物上找到而没有新的特点。

加以发展：在文献的加以发展部份中，一般可以找到实验性的方法、器材、设备、处理过程，有的还可以根据不同的情况找到计

算方法。如果在社会科学文献中，可以找到调查的结果；在生物学科文献中，可以找到观察的结果。总之，在这一部份中，可以或多或少找到一些情报。因为，在这方面所做的工作与以前所做的工作不一样，它是在原来工作的基础上有一些发展。如一篇文章中采用的不是新的计算方法(即以前有人采用过)，不能说有新的知识成份，那么，情报等于零。如果在文章中谈到新的方法、新的技术，那它就是一种情报。在这一部份中又可以找到实验的结果，观察的结果，这些结果可能是数字，也可能是文字，还可以找到图表、曲线、照片和论据等。论据是一种非常重要的情报。论据就是当我们对一个问题在讨论时为了解它的深度应用的一些逻辑形式和比较形式。一般过程为：

讨论——→逻辑的形式——→比较的形式也就是找出比前的结果和文献中所得结果之间进行比较，提供论据的目的就是检验前言中所讲的假设的东西。

在大多数情况下，往往发展的方法不是新的东西，包含的情报不是很多。当然，在认识这个问题时，应当非常小心，一个已经认识情报的结合可以促使产生另外新的情报，如果不注意这一点，那我们就可能这样认为，自从创造了A．B．C．这种书写方法之后(不管是中文的，阿拉伯文的书写方法)，再沒有产生新的情报了。我们必须小心的注意到，情报本质和他们之间的联系对所产生新的情报的作用。

结论：回顾前言中所总结和讨论的问题，提醒人们注意工程的作用，进一步告诉关于文章的下一步内容，出版物的结论方面，情报亦等于零。

参考文献：参考文献不包括情报的内容，所讲的同样是人们知道的内容。

几点意见：

（1）情报的相对性：由于每个人接受的情报不同，有的人了解的多，有的人了解的少，源情报的情况不同，对于不同的人情报

具有相对性；

（2）同一的概念可以用无数的形式表示，在一个句子里，一种表示法将一个词的定义改变，那么就是另外一种意思了，如果一个词在很多句子里都使用，得出的概念是无数的。这就产生了对文献工作者来说的主要问题之一，即文献中的词在句子中的确切定义事先是不知道的，所以使用自然语言的系统当然是很困难的；

（3）同样的事实可以有不同的方法来表示，即用同样的观察方法也可能会得出不同的结果，也就是取决于先前的假设了；

（4）有一种永久的表示方法就是数据，这是从实践中得出的结果，是永久的。当一篇文献发表很久，对于其他的讨论是无意义的，因为讨论的内容已经过时，但是对其经过实践得到的数据对科研和生产仍有作用。

以上讲的这些出版物，情报的概念包括一些什么东西呢？主要有：

题目，

作者，

关键词，

文摘。

这是一般的情况，并不是所有的出版物都包括这四方面内容，有的期刊中，只包括题目、作者，可能沒有任何的情报。文章的题目，不总是包括很多情报，特别是有些坏的做法，因为有些人认为，题目是吸引人的，这样题目成了广告，和内容联系很不紧密，甚至于面目皆非，这对于报道一篇文献来讲是很不够的。

一般情况下，一个文献系统，必须能回答下列问题，即：

出处——在什么地方能找到这些情报；

作者——谁写的情报；

内容——具体情报。

这样的文献系统，对所有的出版物这些问题都能得到回答。但是用戶到图书馆找他所需要的情报，往往是局部的，不全面的。不

同的用户对情报要求的侧重点不同，如研究人员和实践人员对情报的要求就有很大的差别。

研究人员一般对文献中所有的问题都很注重，对所有的事实、讨论问题有兴趣，对美观等一类问题也有兴趣，他们希望文献的结构好，论据全面，他们可以用别人研究的经验来检验情报。

实践人员，包括实验者、操作人员，他们可以是工程师、技术员，出于自己的工作性质，对事实很感兴趣，为了找到情报尽快地用到工作中去，希望得到的这些情报是经过检验有效的情报。他们的目的主要是想尽快地利用这些情报。

研究人员可通过其研究经验来检验情报，但是一般情况下实践人员是没有时间也没有方法做到这一点。我想你们中间也是有这样情况的，把全篇文章是否从头到尾都读呢？读了是否能全部记住？如果全部记住了是否全部有用？我可以说明的是，我所谈的很多东西在很多时候是无用的。研究人员没有足够的时间来阅读所有的文献，有些特别的文章，一般只有十个左右的人去读它。可以看出文献发表原则与目的往往不太适应，对于这个问题，不能花费许多时间去讨论。值得提醒一点的，也就是在读一篇文献和写一篇文献时要注意上述问题。

目前，实践人员所需要的是快速的、经过检验的、准确的情报，他们对目前所能提供的情报不太满意，经调查得到：

15%的用户认为文献服务不满足要求，因为技术术语没有标准化；

55%的用户认为，原始标引不够好，也就是关键词选的不好，文摘不全。

30%的用户认为问题的表达的形式不好。

综上所述，事实上一个文献中心只能告诉你的是到什么地方找到出版物，不能对它有更高的要求，目前我们从中看到了用户的不满意，需要很好重视并加以改进。

二、数据的内容和分类

数据的定义：

数据是一个事实的情报，它是得到很好表达的事实情报，它本身所具有的内容，不需要解释，不需要上下文联系，就有其独立的内容，换句话说，数据是一种独立的情报。

1．数据的内容：

数据可以表示数量，也可以表示质量，并非是一讲到数据便是数量问题。数据包括的主要内容是：主题、属性和状态。

主题是数据的基础，包括一个自身存在的东西。在一张企业卡片中主题可以是一个人，也可以是一个化合物，也可能是一个地理的位置。一个主题是由二个因素来确定的，即性质和价值。性质可以认为在数据库里的是人，这个人在企业里工作。价值是一些名字。价值在文档中的表现可以是名字，也可以是代码数字。如在医疗证上的代码数1360451043001的就是这样的一个值，其中“13”代表男性，“6”出生年，“04”为出生月，“51”代表省，“043”代表地区，“001”代表4月份此人在本地区是第一个出生。

上面的例子可以说明，数据的性质可能是一样的，然而价值是可以区别的。

属性：可以叫谓语。

例：主题：x　　性质　　值

　　属性　　职业　　电工

这个例子就是说，x 先生的职业是电工。如果性质指的是家庭状况，那么值就表现为已婚和两个孩子。如果性质为年龄，那么值就是三十四岁了。这些所有成份，只有跟主语联系起来，才有它的意思。

状态：有些数据库里它是主要的成份，但在有些情况下，如一

个企业里的数据库，状态不是非常主要的东西，因为上例中的职业和年龄中的状态彼此是独立的。用情报性质和值来对“状态”下定义。如在一个人的卡片中，“性值”是顾员在大学里的学习状况，“值”是名字，“状态”可以是顾员加上他的名字，也可以是顾员在大学里的状况。

用另一个例子可以清楚地看出“状态”的重要性。

例：毒性

主题：化合物

属性：药剂的限制量DL50(致命剂量) 即吃了这种剂量，含有50%的动物死亡。美国人把这个属性叫NAC(即一个人能接受最大毒性的限度)这个致命量是从动物身上试验得来的，也适用于人。例如：

CO(一氧化碳)对动物的致命量是$1000 \cdot 10^{-6}$ (1千/百万)

对人的致命量是$40 \cdot 10^{-6}$ (40/百万)

毒性的致命量必须结合动物在实际的状态中加以说明(如说明试验的动物是纯种白鼠或其他品种)。还必须确定吸毒状态，即口服、吸入和肌肉注射，因为在三种不同的状态下，毒性的作用是不同的。时间对吸毒的作用也很重要，对有的药物来讲，早、晚服药它的作用不同，这就是时间效应。有毒金属在某个时候毒性作用可能加倍。从上述情况可知，不同的条件，可以得到不同的致命药剂量。研究这方面情况对毒物数据库有重要的实际意义，如安全部门和社会保险部门都必须重视了解各种各样的具体情况。

例：金属焊接

属性：焊接的类型

值：电焊、氧焊、氩弧焊

状况：机械处理方式，可能是煅造，也可能是压冶。

热处理方式：淬火。

西德已建成焊接数据库，在该库中有十种状态表示其属性。全面了解一个数据库，必须有一定的技术水平，不仅应是一个信息专

家，而且要求懂得专业技术知识。在讲到技术问题时，这类数据库有很多复杂的情况。如果用戶利用影据库的数据，焊接的结果是坏的，那他就会说，情报是错误的，数据库所有者就要找出原因给于纠正。如果提供的数据是错的，提供数据的人有责任，数据库所有人也有责任。情报人员的责任是重要的。对于状况的性质确定后，就确定了数据库的质量，必须认眞对待。

2．数据库的定义：

数据库就是组织和利用、集合、登记和发送数据。

这是一个基本的定义，沒有确定自动化的程度，情报基础的性质，活动的范围，提供服务的內容。根据此定义，一张卡片上的文挡也可以认为是数据库。一个数据库包括主题、属性、状况。我们这里讲的数据库是从有一定数据集合出发的。在欧洲或美国对文献数据库和信息管理系统（DBMS）的定义都有所混淆。在一个特定的时间里，在一个共同体內部，它们有一个共同的术语，在本周的学习中我们要用一个共同的概念来理解数据库，数据库应当包括：

——哪一个科学的部门；

——而且必须确定主题、属性和状况的性质；

——确定数据库的广度：例如毒药的性质有主题、属性、状况；

——了解属性的精确度：例如致命量（DL）：

50毫克（mg）/公斤（kg）±20（动物重量）

即 30毫克（mg）/公斤＜DL＜70mg/kg

为了保证安全，必须用30mg/kg，而不能用70mg/kg。在建立数据库时，精确度相当重要。

——数据库的增长速度，也就是说明数据库的增长是按什么速度进行的，指导用戶能很好的了解数据和使用数据。

对于一个小的共同体建立数据库沒有必要，因为价格贵经济上

很不合算；另外人们总是希望情报范围越广越好，而小的共同体很难做到这一点。

3．数据的分类：

为了表明数据的特点，世界科学技术协会数据委员会(CODATA）提出了如下的数据分类方法，如表2所示。

表2 数 据 分 类 表

数据类型	化学/物理	地质/天文	生物
a_1 可以反复测量的数据	大多数数据	地质结构、岩石重力加速度、恒星	大多数数据
a_2 只能度量一次的数据		灿爆发、太阳的炽斑、新星	稀有品种 化石
b_1 与位置无关	大多数数据	矿石、地球构造	除去地球之外生命的大多数数据
b_2 与位置有关		岩石、化石、天文数据、气象数据	稀有品种、化石
c_1 从观察和试验中获得的基本数据	光谱、结晶学沸点值	地震记录 气象图表	生物数据（呼吸速率、血压）、生物化学数据（细胞组织和器官构造）
c_2 通过基本数据和理论模式结合得到的数据	基本数据 晶体结构	化石区域 太阳的温度分布	
c_3 通过理论计算导出的数据	通过量子力学计算的分子数据	利用天体力学预测的日食	
d_1 可测定数据	大多数客观数据	行星轨道原理	染色体数目