

第一章 化学信息与计算机

1.1 化学信息的产生及计算机处理的必要性

在中世纪，炼丹术士的秘方是口授的，即师傅将配方和技艺传给他的徒弟。后来，开始以插图或图形方法代替口传。到15世纪出现了第一本教科书，于是，诞生了化学。化学信息的传输就是以教科书这种初级形式进行的，并且由古一直延续至今。所不同的是，在古代的化学书籍中没有理论，也无化学名称，而仅以简略的符号来表示化合物。在化学科学的发展进程中，随着 Avogadro 和 Boyle 气体定律的发现以及 Dalton、Berzelius 和 Cannizzaro 的近代结构理论的确立而首次产生了“化学结构”一词，从而，可以区分于非结构的化学信息。

从19世纪开始，科学技术的飞速发展引起了信息的大爆炸，Price^[1] 曾对广阔的学科领域中信息增长的速率进行了研究，观察了不同的数据增长模式。当一项新兴领域出现时，信息增长率就加大，如液晶的出现即为一例。再如干扰素，在最初的10个月内，其文献量成倍地增长（图1.1）^[2]。

在过去40多年中，化学文献每年都约以5倍的速度增加；同时，新实验技术所产生的信息量也极其浩繁。而且，近代科学的显著特点是学科之间相互渗透。如生物学家，不仅需要熟知生物化学，而且需要了解和掌握化学、医学及物理学等；物理学家，不仅需要熟知物理学，而且需要了解和掌握数学、化学及工程科学等。同样，化学家不仅需要熟知化学，而且需要

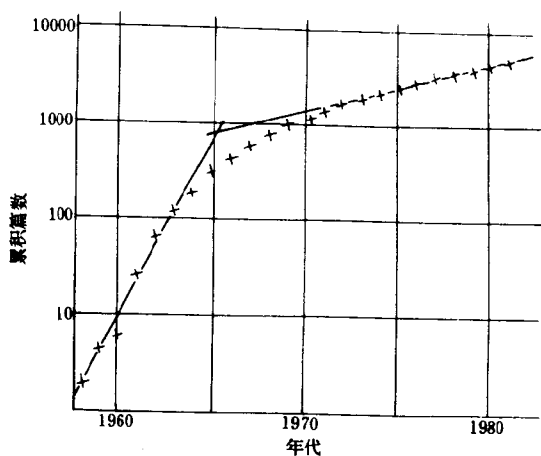


图 1.1 干扰素文献增长情况

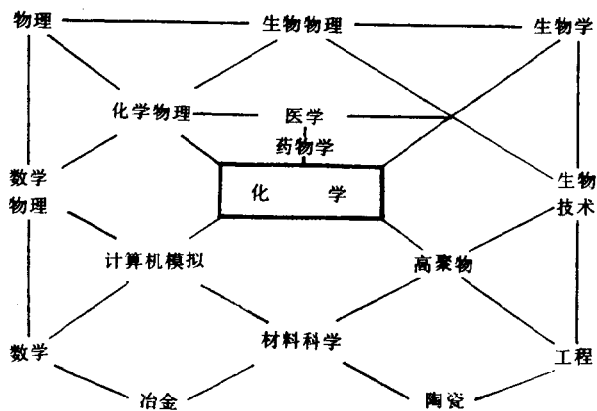


图 1.2 近代化学与其他学科的相关图

了解和掌握生物学、医学及物理学等等。在近代科学中，化学与其他学科领域的相关网络示于图1.2。由图可见，化学家为了进行科学研究，所需要的信息范围甚广，所涉猎的学科范围甚宽。因而，在信息时代中，对化学若仅仅以传统方式进行数据的处理，则会与实际需要极不适应，由此，客观上提出了应用计算机技术的要求。于是，化学必然地与计算机结下不解之缘运用计算机进行化学信息的管理是势在必行。

1.2 数值型数据库

60年代初，世界上一些发达国家开始应用计算机进行数据的存贮和检索，实现了现代化的信息管理。随着计算机硬件及软件系统的发展，从70年代以来，数据库有了突飞猛进的发展。据统计，世界上商用数据库的增长情况为：1975年约390个，1978年为1100多个，1982年达到1800多个。

数据库可分为目录型、数值型和管理型三类。目录型，如文献检索；管理型，如人事档案材料；数值型主要是科学数据库。化学中数值型数据库又可粗略地分成二类：（1）结构数据的数据库；（2）非结构数据的数据库。前者，是与物质化学结构有关的信息，如质谱、核磁和红外光谱等各种谱图数据，以及化合物结构本身的相关数据等。后者包括各种物质的物化参数、热力学性质、相平衡以及传递过程方面的数据等。有关谱图的数据库将在第四章予以介绍；结构的编码、存贮及检索将在第三章予以讨论，而本节仅拟对数值型化学数据库做一概括性描述。

1.2.1 现代化学数据库的特征

数值型化学数据库有如下几个主要特征：

1. 拥有大量的原始数据。我们称之为一级数据，它即是一

般意义上所理解的数据源。例如，目前的谱图库中碳-13已达50 000多张谱图，质谱达到约70 000张，红外光谱达100 000多张；美国NIH/EPA的毒性库所包含的化合物有20多万个。

2. 在一级数据的基础上，可以衍生二级、三级数据。如由碳-13谱图系数，即化学位移和多重性，可派生出化学位移-子结构相关表（表1.1）用于化合物结构的人工智能解析。

表1.1 化学位移-子结构相关表

编 码	结 构	$^1\text{H NMR}$	$^{13}\text{C NMR}$
1	$\begin{array}{c} \text{CH}_3 \\ \\ \text{CH}_3-\text{C}- \\ \\ \text{CH}_3 \end{array}$	84.6 - 60.0	26.02 - 31.13
2		96.0 - 72.0	24.47 - 33.57
3		90.0 - 60.0	25.48 - 34.04
10	$\begin{array}{c} \text{CH}_3 \\ \\ -\text{C}- \\ \\ \text{CH}_3 \end{array}$	90.0 - 48.0	14.72 - 36.27
11		90.0 - 48.0	10.12 - 36.27
12		89.4 - 36.0	6.80 - 32.61
106	-CH ₂ -	220.0 - 108.0	20.46 - 56.42
107		162.0 - 102.0	18.88 - 53.09
108		180.0 - 114.0	13.27 - 54.14
109		144.0 - 0.0	14.66 - 57.16
153	-O-	- -	- -

3. 向智能化方向扩展。数值型数据库总是要在一级或二级信息检索的基础上，进一步扩展功能，使其具备一定智能，即过渡到知识库和专家系统的范畴。如利用萃取分配比，萃取离子的浓度和溶液的酸度等基本数据进行三维空间中响应曲面

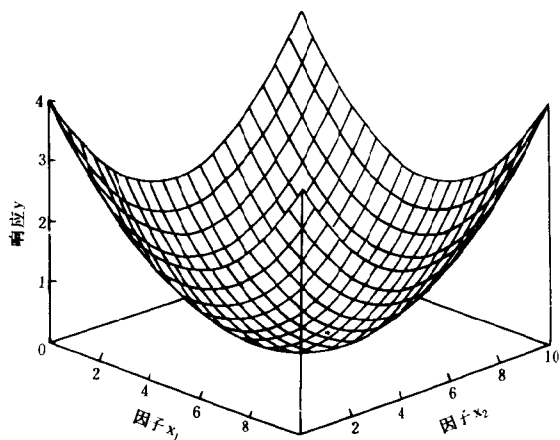


图 1.3 两因子响应曲面

$$Y = 4 - 0.8x_1 - 0.8x_2 + 0.08x_1^2 + 0.08x_2^2$$

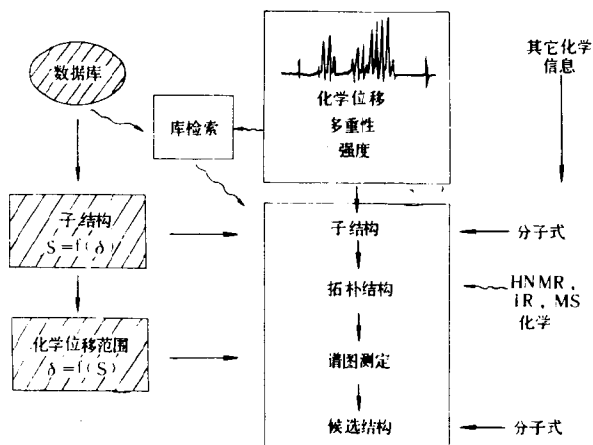


图 1.4 BASF公司的 ACCESS系统

图中 S 表示子结构；δ 表示化学位移

(response surface) 的研究。图 1.3 所示为因变量 Y 对于自变量 x_1 和 x_2 的响应曲面及其数学模型。这类响应曲面可为某种流程的调优提供理论依据。

再如联邦德国 BASF 公司的综合谱图库信息系统中, ACCESS (Automatic Computer Correlation and Evaluation of Spectra based an Substructure) 程序包可以自动解析化合物结构, 其解析过程示于图 1.4。

1.2.2 数据库的作用

数据库不仅是提供原始数据的, 也是信息再加工的必须手段, 而且能为进一步研究提供条件, 开拓通路。因此数据库是基础研究工作之一。

1.2.2.1 化学数据库与人工智能

人工智能的研究 (如: 有机化合物解析) 与数据库的关系尤为密切。

DENDRAL 系统是大家所熟知的最早 (60 年代中叶开始) 的专家系统之一 (见第六章), 此系统是美国斯坦福大学的质谱学家、遗传工程学家、化学家和计算机专家通力协作的一项卓越成果。在该系统中, 运用低分辨质谱中的断裂规则及碳-13 和氢核磁共振波谱等进行化合物结构的演绎推理。随后就在许多企业、研究机构 and 大学获得了应用。这一先驱工作的成功振奋了整个人工智能领域。

在 DENDRAL 系统演绎所得候选化合物的验证中采用了谱数据库^[3]。如图 1.5 所示, 程序功能分为两个部分。第一部分根据子结构环境和 H 谱化学位移的相关性建造 H 谱数据库:

(1) 程序 HCODE, 完成子结构编码的自动生成和数据库的构造;(2) 程序 HDCODE, 完成数据库中子结构-化学位移归属。第二部分, 未知化合物 H 谱的分析:(1) 程序 HNMRP,

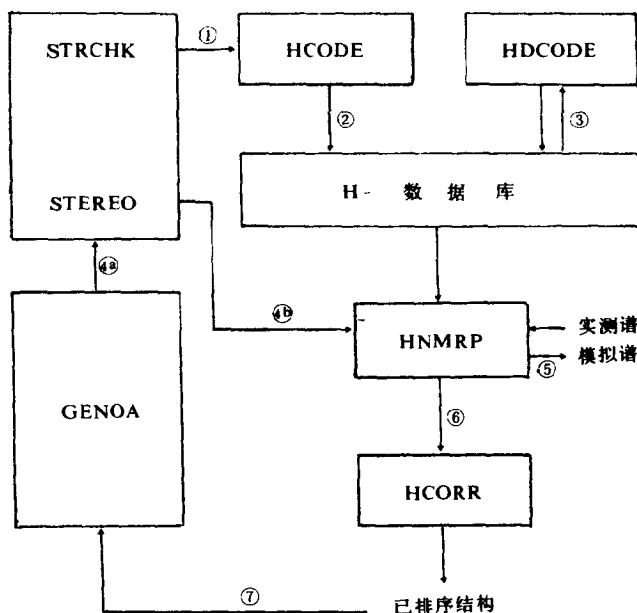


图 1.5 H 谱分析程序

未知结构说明；② 与数据库联接；③ H 谱数据库的维护及检验；④a 同分异构体生成；④b 空间异构体生成；⑤ 波谱预测；⑥ 预测谱与实验谱的比较，并将候选化合物排序；⑦ 将此结果用于约束化合物的最终确证

预测候选化合物的 H 谱；(2) 程序 H CORR，比较预测谱与实验谱，并根据二者的符合程度进行候选化合物的排序。

再如，新近已由南斯拉夫化学计量学教授 Zupan 研制成了 CARBON 这一结构解析专家系统^[4]。此系统的显著特点是它具有浓厚的数据库色彩，Zupan 等在 70 年代中已经建立起他们

的 C-13、IR 等谱图库信息系统，同时创造了三距离聚类分析方法。CARBON 系统主要运用的是 C-13 谱图数据，目前，此数据库中 共有 2536 张 C-13 NMR 谱图，30000 个已作归属的化学位移。以此数据库为基础，提取不同取代基的化学位移信息，用于谱图的模拟和归属。

专家系统 CARBON 的知识规则由四部分组成：(i) 相似谱图的系统聚类树（三距离法）；(ii) 不同取代基、不同空间位置及不同构象的化学位移表；(iii) 功能团类别的统计和数学表达；(iv) 子结构检索、归属等的图论算法。由此可见，CARBON 系统是在坚实的数据库基础和技术的根基之上研制而成的。

1.2.2.2 化学数据库与分子设计

研究活性物质，如药物、杀虫剂及农药等的合成方法，多年来一直是沿用纯经验的方法，这要先合成一系列化合物，然后进行生物实验进行筛选。对于有效物质的改进，通常采用结构的修饰方法，如以穷举法进行各种功能团的取代。这种用经验、以母体衍生物为线索的药物合成方式效率极低。以农药为例，在国外市场上，一种化学药品的需要合成量随着年代变化示于表 1.2。若以表 1.2 增长速估测，到 1997 年，要创制一种新的农药，须合成的化合物将变成天文数字。也就是说，时至今日，

表 1.2 创制一种农药的合成量

年 代	1956	1964	1970	1972	1979
合成量 (篇)	1800	3600	8000	10000	12000

单靠纯经验探索药物的方法已经陷入困境。因此，人类在寻求新的方法来摆脱这种状况。其中的一种途径是运用计算机来辅助

分子合成。所谓辅助分子合成主要包括如下内容：定量结构物性相关分析（QSAR）（见第八章）；模式识别分析（第五章分子模型化技术和定量结构-活性相关的信息基础为X-射线晶体结构数据库和Hansch疏水性等参数的物性数据库。

1.2.2.3 化学数据库与合成路线方案设计

有机化合物计算机辅助合成路线方案的设计就是一类专家系统（见第七章）。1969年美国哈佛大学的知名化学家Corey和Wipke报道了世界上第一个合成路线方案设计系统，之后其他系统相继问世。迄今，最知名的为Wipke研究的LHASA系统，此系统目前已被美国和欧洲的瑞士、德国等国的跨国公司所广为利用。在Merck公司的实验室中，每天有60多位科学家在使用着的LHASA的派生系统SECS对Merck公司的新药创制起了重要的作用。

LHASA系统在演绎分析中采用逆向推理，即由目标产物前推，得到前体，再继续前推，……形成一棵倒挂的树。支撑这一推理过程的后盾是庞大的知识库，而知识库中的规则，一是来源于有机合成化学家的经验，即非书本的；二是来源于书本的，如反应信息的数据库等。

1.2.2.4 化学数据库与分析实验室

分析化学是最早引入计算机的学科之一。在分析实验室里，分析测试仪器尤其是大型仪器与计算机的联机运转已广为采用。在一些发达国家，不少实验室已有以计算机为中枢的，与多类分析实验室、众多用户终端、大型数据库相联形成的网络系统，其中，殊为典型的是联邦德国Max Planck学会煤炭研究所的计算机的网络系统。该所目前的主要研究方向为金属有机络合物（用于均相催化剂）。围绕这一主攻方向的是双机（VAX11/760）并行运转的计算机系统，它把“触角”伸到研究所主要的大

型仪器分析化学实验室，诸如质谱、气相色谱、X-射线衍射、核磁共振、红外、紫外及偏共振等等，控制着分析仪器的运转和联机采集数据。同时，分析化学家运用设置在分析实验室中的用户终端，实施联机对如质谱和核磁等的数据库检索，来进行化合物的结构解析（图1.6）。目前，数据库检索成为质谱信息处理中最关键、最重要的一环。

该所的质谱数据库系统为SISCOM。此系统由许多模块所组成，功能甚强。在晚间机器可运行自动解析程序，次日，可

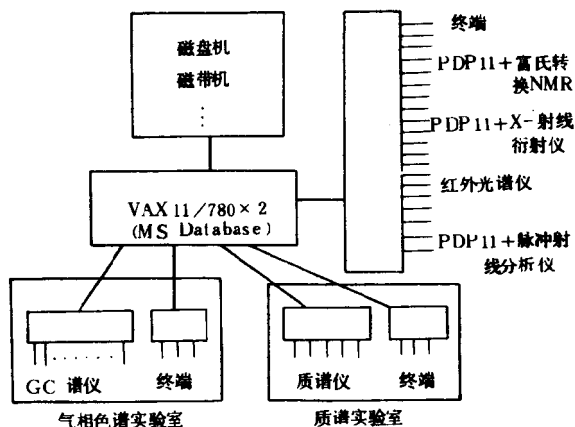


图 1.6 联邦德国Max Planck学会煤炭所计算机系统示意图

将结果自动输出，然后，分析人员通过人机交互方式对之进行优化。此系统不仅可以进行相似检索，而且还能进行智能性的功能阐明，因而它是一种解释型谱图库系统。

煤炭所的计算机系统已经成为该所的“生命线”。由于这一系统得力的配合，使得该所每年都有众多高水平的成果。

1.2.2.5 化学数据库与分析仪器工业

在国际上，化学数据库的建造始于60年代，其高峰时期在70年代的前半期。70年代中至80年代初，主要手段的谱图数据库均已建成，随着数据库的发展与完善，这一新兴的技术开始介入仪器分析工业。70年代已有分析仪器的厂家把科学研究的成果——数据库“装饰”到他们的产品上。进入80年代以来，在诸如质谱仪、核磁共振波谱仪、红外光谱仪等，许多大型分析仪器上，数据库已成为一个重要的组成部分。现在，这类仪器系统的数据库部分已具有相当的规模。如英国VG公司的质谱仪，质谱库中的谱图多至30000张以上。由于计算机和谱图库系统的应用，极大地增强和延伸了大型分析测试仪器的功能。

1.2.2.6 化学数据库与化学工业

早在60年代，化学物质物性数据库已经被引入到化学和化工工业中。在世界上著名的美国杜邦化学公司、道化学公司、柯达公司、联邦德国BASF公司、英国帝国化学有限公司以及联邦德国的某些染料工业公司等大型工业企业中，计算机联机的化学信息起到了重要的作用。

在60年代末，杜邦公司已经建造起它与美国化学文摘服务社（CAS）共同协力研制成的结构信息系统CS⁺。CS⁺系统中，其信息源尚包括该公司的技术报告和世界上的有关专利。此时，库中已有聚合物14000个，非聚合物5500个，库中数据每年增加18000个。系统具备子结构检索和结构筛等功能。

道化学公司亦是早在60年代就引入了结构信息数据库。到60年代末期，库中已有结构10000个，化合物的结构为线性编码（WLN）。另外，库中存贮了子结构碎片表，与生物相关的结构筛等，可进行子结构的检索。

柯达公司的化学信息系统已运行多年。目前，该公司在着力研究光谱数据管理 (SDM) 软件包。SDM 的一个重要功能是提供专家系统的骨架，可使波谱学家由经验的结构-波谱相关规则来建造有机化合物解析的专家系统。SDM 中具有众多的数据库检索指令：(i) 整体结构检索；(ii) 子结构检索；(iii) 全谱图匹配；(iv) 子谱图检索；(v) 子结构与子谱图检索的逻辑组合，也就是求得结构与谱图的相关表，用于专家系统规则的自动生成。柯达公司的局部计算机网络示于图 1.7。

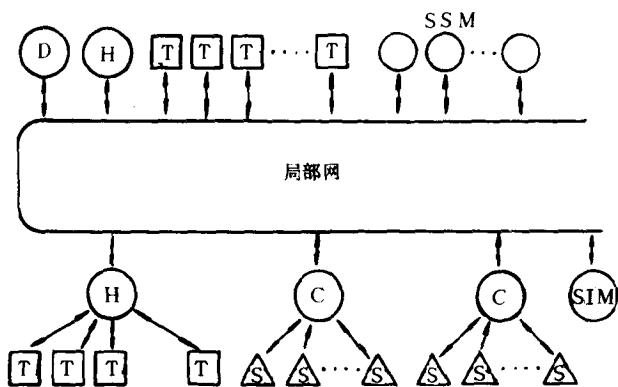


图 1.7 柯达计算机局部网

①：程序调度；②：结构阐明（主机）；③：数据库；
 ④：终端；⑤：波谱仪；⑥：子结构检索机；⑦：波谱解析机

BASF公司的化学信息系统早为大家所熟知。以Bremser为首的研究小组，提出了HOSE编码方案，即对化合物的子结构按照层次进一步完善而提出的HORD编码（见第三章）他们把大量以HOSE或HORD编码的子结构作成缩微胶卷，对外出租

或出售，并以手册形式加以出版。目前，BASF公司的C-13 NMR谱图库已建过50000张，可以购买，但售价较高。该公司还建有红外光谱的谱图库系统。在数据库的基础上，该公司已建立了有机化合物结构的自动解析系统。这些成绩在计算机化学领域中产生着广泛的影响。

1.2.3 化学数据库系统举例

1.2.3.1 化学信息系统 (CIS)

美国化学信息系统 (CIS)^[5,6]是包括范围最为广泛的一个系统，多个子库系统由化合物的结构及系统命名检索系统SANSS所衔接，即SANSS是CIS之枢纽。

CIS中的子系统包括：质谱、C-13 NMR波谱、红外光谱、X-射线衍射晶体数据、哺乳动物剧毒数据。其中，X-射线晶体数据及NMR数据文件，包括目录型信息。

分析程序库包括：统计分析和多种数学模型。如分子的模型化程序、溶解度和分配系数的计算及化学药物的生物分解等等。

CIS所具有的检索功能为：名称检索、分子式检索、CAS登录号检索和子结构检索。子结构检索包括化合物名称片断及子结构碎片的检索等。

1.2.3.2 剑桥结构数据库 (CSD)^[7]

CSD始建于1965年，信息源为从1935年以来有机化合物的X-射线及中子衍射数据。CSD的传播可通过两条途径：(1)已出版的手册；(2)联机检索。

CSD文件和程序可在英国国内的3个中心及世界范围内的20个分中心进行运行。在人机交互检索中可进行结构的显示。我国于1987年引进并运转了该数据库。

中记录所包含的信息示于图1.9。每一记录由参考编码

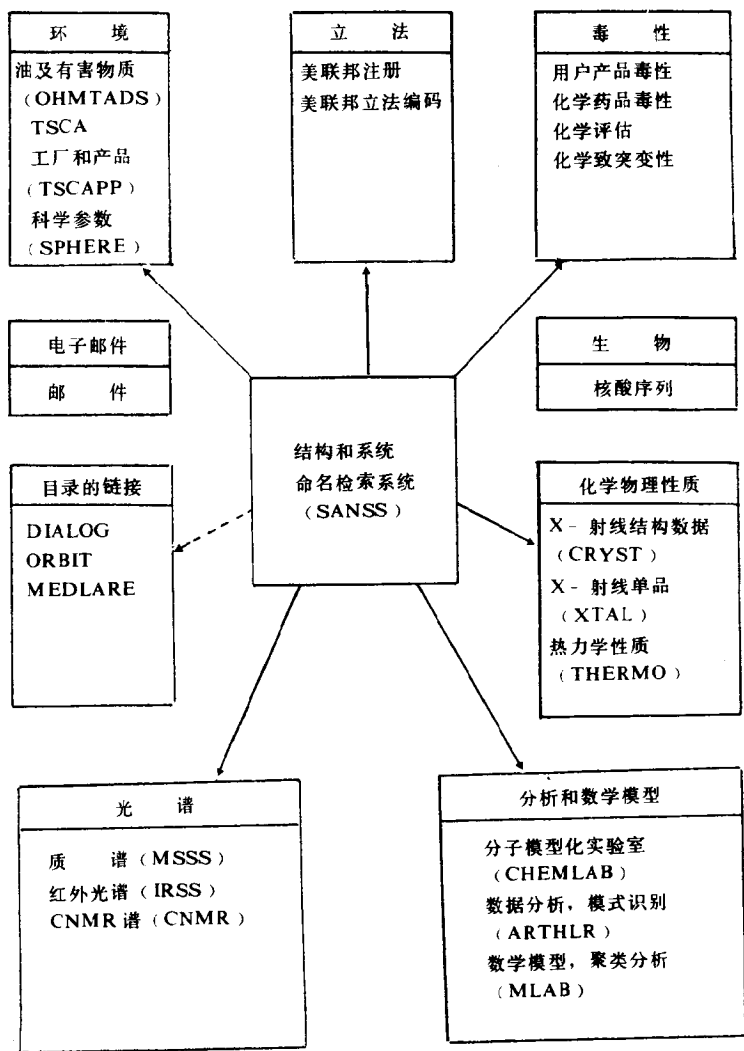


图 1.8 美国NIH/EPA的CIS运行系统

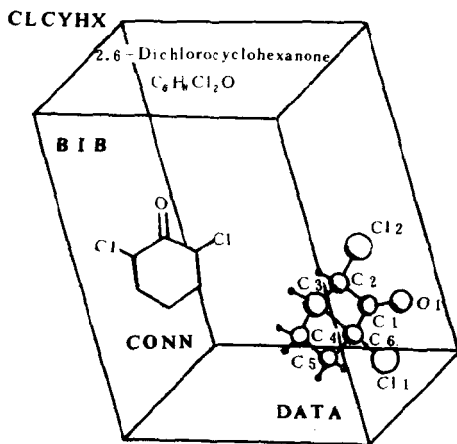


图 1.9 CSD记录的信息内容

BIB : 文献摘引, 化合物名称,
分子式, 化学分类

CONN : 二维联接表的紧凑形式;

DATA : 原子坐标, 晶胞和对称性,
准确度指示器, 文字注释

所识别, 此处为 **CLCYHX**。同时每一记录由 3 个数据段组成:
BIB (文献目录) **CONN** (联接度) 和数值型数据 **DATA**。CSD
允许脱机查询 **BIB** 或 **CONN** 片断, 之后再检索相关的 **DATA**。
CSD 还允许进一步对之分析处理, 以进行结构的显示等。

1.2.3.3 环境化学的数据和信息网——**ECDIN** ^[8]

ECDIN 是欧洲共同体的一个下属组织所建造的有关环境
化学物质的数据库。**ECDIN** 运行在哥本哈根的数据中心的主中
心 (**Data-Centralen Host Centre**), 可通过信息网进行联机
检索。库中有化合物 60000 个, 其中有 20000 个为剧毒化合物。

ECDIN可视为汇集于一体的一套计算机化的参考书：

- 化学字典；
- 物理化学性质参考书；
- 化学产品生产的字典；
- 一套产品和商业统计表；
- 危险物立法的手册；
- 化学物质毒性影响的注册。

1.2.3.4 中国科学院综合化学信息系统

1979年中国科学院化学部组织院属各有关化学所开始进行综合化学信息系统的研制。到1985年，已先后建成了质谱谱图

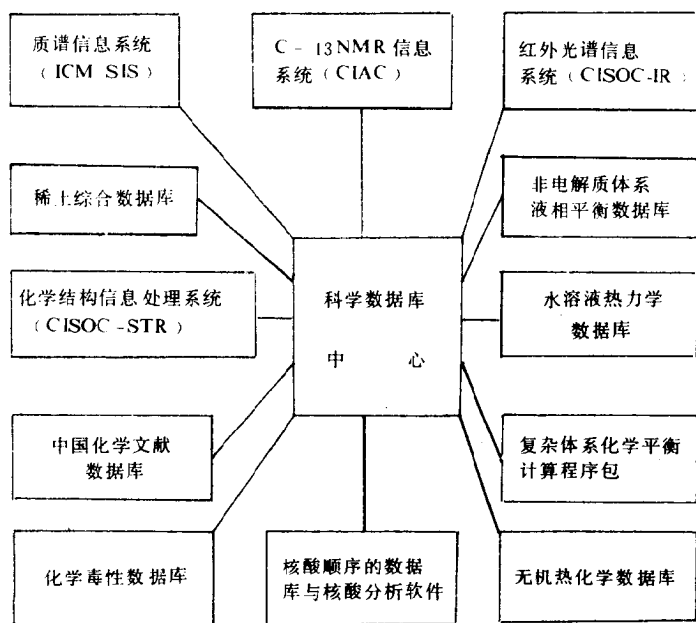


图 1.10 中国科学院化学综合信息系统

数据库、红外光谱数据库、C-13 NMR数据库、化合物毒性数据库及非电解质体系统液相平衡数据库等，目前在中国科学院数据库中心的资金支持和组织下，正在开展一些新库的研制该系统的框图示于图1.10。

其中，质谱数据库，¹³C NMR数据库和红外光谱数据库将在第四章中讨论，本节仅对由中国科学院长春应用化学研究所承担的稀土综合数据库系统作一简单介绍。

稀土综合信息系统包括如下子系统：(1) 稀土萃取；(2) 稀土的物理和化学性质；(3) 稀土材料；(4) 稀土资源；(5) 稀土的市场信息。目前已建和正建的子系统为稀土的萃取和稀土的物理化学性质。

1. 稀土萃取信息子系统

(1) 检索功能：可以进行萃取剂、金属离子、分配系数、萃取平衡条件及参考文献等项的检索。

(2) 数据关联：在数据库基本数据的基础上进行相关性研究。如萃取分配比与金属离子浓度、酸度和温度的关系，这些相关性可以用三维空间中的响应曲面来表示，从而为分离流程的调优提供理论依据。

(3) 串级萃取的数学模型：主要以徐光宪的串级萃取数学模型为基础，给定基本数据系统可以进行串级萃取有关参数的计算。

(4) 萃取剂结构与萃取性能相关性研究：对萃取剂用纯经验的方法以母体衍生物为线索合成化合物的成功率极低，为此客观上提出了建立在科学基础上，有预见性的创制新萃取剂的要求。

本系统对于萃取剂结构与萃取性能的相关性进行了广泛的研究。广义 a_N 指数是其中一种方法（见第八章）。 a_N 指数是HMO