21 CENTURY
世纪计算机化学丛书

化学信息系统开发技术

Chemical Information System Development

乔园园 著

化学工业出版社

21 CENTURY
世纪计算机化学丛书

化学信息系统开发技术

Chemical Information System Development

乔园园 著

化学工业出版社
·北京·

化学信息系统即“以化学数据库和化学在线计算为核心的网络服务系统”。化学信息的数据类型、表现方式多种多样，既有文献、数据，又有图形、图像。在用户界面方面，则需处理图谱、分子结构式和化学反应式等。因此，相应的设计、开发工作需要综合运用多方面的知识和技术。本书简要介绍了信息的基本知识，阐述了信息的重要性。围绕化学信息对象，介绍了数据结构与算法，分析了系统架构模式；针对化学信息的特点，选择不同类型的化学信息系统实例讲述了其系统设计思路和开发方法，对实际研发有重要的指导作用。

本书适合化学、化工、生物化学、药物化学以及相关专业的科技人员、高校师生参考。

图书在版编目（CIP）数据

化学信息系统开发技术/乔园园著. —北京：化学工业出版社，2011.6
（21世纪计算机化学丛书）
ISBN 978-7-122-11287-3

Ⅰ.化… Ⅱ.乔… Ⅲ.化学-信息系统-系统开发
Ⅳ.O6-39

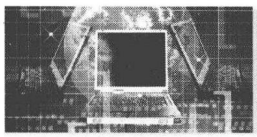
中国版本图书馆CIP数据核字（2011）第089647号

责任编辑：成荣霞　　文字编辑：孙凤英
责任校对：宋　玮　　装帧设计：王晓宇

出版发行：化学工业出版社（北京市东城区青年湖南街13号　邮政编码100011）
印　　刷：北京永鑫印刷有限责任公司
装　　订：三河市万龙印装有限公司
710mm×1000mm　1/16　印张15¼　字数261千字　　2011年8月北京第1版第1次印刷

购书咨询：010-64518888（传真：010-64519686）　售后服务：010-64518899
网　　址：http://www.cip.com.cn
凡购买本书，如有缺损质量问题，本社销售中心负责调换。

定　　价：68.00元



21世纪
计算机化学丛书

序

计算机化学的兴起与发展是与化学知识创新的迫切需要紧密联系的。十年前化学家使用计算机的还不多，现在却已十分普及；十年前对化学计算的要求主要是化学信息的采集、加工、储存和利用，而如今除了以上的基本要求之外，更强调了由化学信息发现新知识和化合物物性的定量预测。计算机网络技术的飞速发展与普及，对计算机化学来说是一个发展的机遇，而愈来愈高的计算要求是计算机化学发展面临的新挑战。今天，以计算机及其网络深入到社会的各个层面为标志的数字化新世纪的到来，将使传统化学发生深刻的变化：以计算机及其网络系统为工具，建立由化学化工信息发现新知识和实现知识传播的理论和方法；认识物质、改造物质、创造新物质，认识反应、控制反应过程，创造新反应、新过程，将成为计算机化学研究的主体。化学数据挖掘、知识发现、计算机辅助结构解析、分子设计和合成路线设计等是当前计算机化学的主要研究方向。可以深信，在21世纪，数字化新世纪的化学不仅要靠“湿”实验室来发展，同时也要依赖于“干”实验室。所谓“干”化学实验室就是指数字化虚拟化学实验室。“干”、“湿”相结合才能更高效地孕育出新的化学实体，才能促进化学由实验科学向严密科学转化，才能大大提高化学非凡的创造力。

为了推广计算机化学的新理论、新技术和新方法，促进科技进步，我们策划了这套《21世纪计算机化学丛书》，主要介绍计算机化学近5年间的新理论、新技术和新方法。希望这套丛书不仅能够大大推动我国科技水平的进步，更能对我国生产力水平的提高产生巨大的影响。

陈凯先
2010年3月

前言

FOREWORD

化学作为中心科学，研究内涵越来越丰富，应用也越来越深入。同时，信息技术的快速发展，更促使化学信息系统的开发技术不断更新。伴随这个进程，化学信息也被赋予了太多的含义，使得化学信息系统不得不覆盖众多的领域，数据类型、表现方式也多种多样，粗略看来，就有以下特征。

（1）类型广泛　既有文献、数据，又有图形、图像（如图谱、分子结构、电子云、反应式等）。

（2）处理多样　既有计量学处理（投影、降维、傅里叶变换），又有信息学处理（分子结构的匹配、结构描述参数及相似性计算），还有能量计算、结构优化等。

（3）界面特殊　除了一般的文本、超文本，还需处理图谱、分子结构式、化学反应式的输入输出等。

（4）功能复杂　文献检索、数据查询、图谱或结构匹配、在线计算等。

（5）应用广泛　搜索引擎、信息管理、检索服务。

化学信息系统可以概括为“以化学数据库和化学在线计算为核心的网络服务系统”，因此，相应的设计、开发工作需要综合运用多方面的知识和技术。一个化工厂的信息管理系统，通常只需按常规的网站来设置，即使需要处理若干种分子结构或者图谱，也完全可以用图片来处理。再比如，谷歌或百度这样的搜索引擎，并不能接受用户输入的二维或三维分子结构来进行检索。因此，化学信息系统所需要的很多功能，是传统的信息系统无法实现的，而必须开发底层的、有针对性的算法或模块。

书名含有“化学信息”的著作，大致有两类：一类是化学文献、情报检索，而近年来则转向介绍化学相关网站、数据库的服务内容、使用方法等，如李晓霞的《Internet上的化学化工资源》；另一类是化学信息学，内容主要是研究如何通过分子的图谱数据、拓扑特征和量化参数等研究其物理化学性质或者生物活性，

如德国 J Gasteiger 主编的《Chemoinformatics：A Textbook》。本书或许可以算是第三类，即以化学信息的共享为目标，介绍化学信息系统的开发技术。

本书概要介绍了信息的基本知识，阐述了信息开放共享的重要性。针对化学信息的特点，选择不同类型的化学信息系统实例，讲述其系统设计思路和开发方法。对实际研发有重要的指导作用，适合计算机与化学领域的科研、教学参考。不过，书中既没有什么秘诀能让读者快速掌握某种信息开发技术，也不可能面面俱到地评述所有的信息共享模式；所能做的，是尽量帮助读者追溯化学对象的设计思想，分析信息系统的构成框架，从而避开花样繁多的开发语言、设计模式的困扰，练习从比较底层的角度来观察和理解一个个看似纷繁、实则有序的化学信息系统。希望读者今后面对自己的实际需求时，在设计理念和开发技术方面，能做出与时俱进的合理选择，创造出更多的专门适应化学研究需要的模式和技巧。

本书所讲解的化学信息的数据结构、算法和开发实例，大都源自本实验室的研究工作，部分参考了国内外的报道。南开大学刘冲、张树众老师为本书部分章节提供了重要的参考资料，并对部分文字进行了核校。在此，笔者对从事化学信息系统开发的各位同行表示感谢。

笔者学浅，难免疏漏，不妥之处，还望海涵。

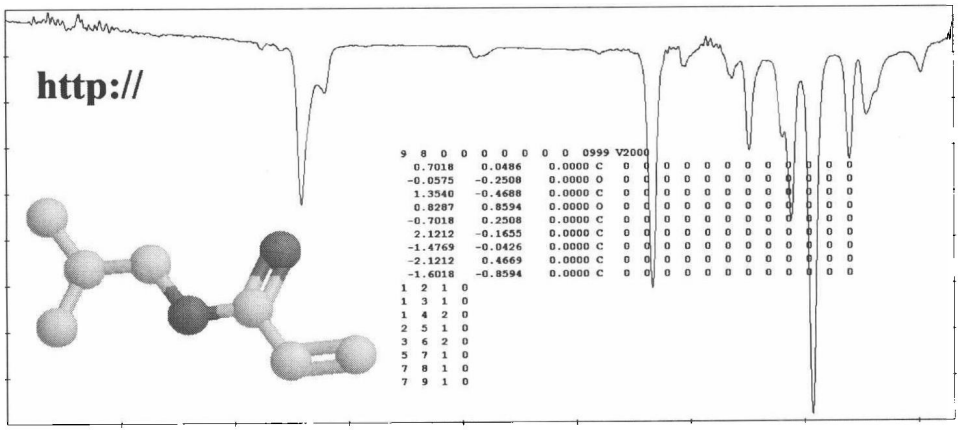
乔园园

2011 年 5 月于天津南开大学

目 录

CONTENTS

1 概　述


http://
9 8 0 0 0 0 0 0 0 0 0999 V2000
0.7018 0.0486 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
-0.0575 -0.2508 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
1.3540 -0.4688 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
0.8287 0.8594 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
-0.7018 0.2508 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
2.1212 -0.1655 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
-1.4769 -0.0426 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
-2.1212 0.4669 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
-1.6018 -0.8594 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
1 2 1 0
1 3 1 0
1 4 2 0
2 5 1 0
3 6 2 0
5 7 1 0
7 8 1 0
7 9 1 0

概要介绍信息的发展历史，对信息的传播、管理及其重要性进行了阐述。

1.1 信息系统概述

自古以来，人类生活就离不开信息，“日出而作，日落而息”，农耕借助太阳运动的信息，“知己知彼，百战不殆”，战争依赖实力对比的信息。存储信息的介质多种多样，从远古的“结绳记事”到龟背、竹简、毛皮、纸张，再到磁带、磁盘、光盘。传递信息的方法层出不穷，殷商时就有“烽火”，古波斯人有高塔“喊话站”，古罗马人有“悬灯”，还有风筝、信鸽等。后来邮政、电报、电话、广播和电视占据了主导地位。现在，因特网又成了最有竞争力的新方法。介质的进步和传播的高效极大地促进了信息数量的增长和质量的提高。

那么信息究竟是什么？答案可以有很多种，就科学研究而言，信息是对科学体系进行探索而获得的有序数据。科学理论是建立在不断积累的经验、知识之上的，科学研究离不开信息。随着信息数量不断增长，对信息的管理就成了一个重要问题。以纸质档案为主体的传统形式图书馆早已不能适应现代化科学研究的需要，必须采用数字化的信息存储介质与网络化的信息传递方法，简单地说，就是用计算机数据库和因特网，建立有效的、现代意义上的信息系统。

从数据库方面来说，信息系统要求存储对象必须是数字化和检索化。数字化似乎比较容易理解，而检索化的含义可能需要解释一下。例如，一幅扫描成图片格式的书页并不能满足检索化这个要求，因为图片的构成单位虽是数字化的像素，但图片所代表书页的检索单位是字、词，而不是像素。类似地，一幅扫描成图片格式的图谱也不能满足这个要求，因为图谱检索的是谱线的位置或强度，而不是像素。那么，如果存储对象本身就是一幅图像怎么办？显然，从技术上来说，完全可以按照像素来检索，但在实际应用中并不现实，更多的是为图像附加文字说明，或者从图像提取特征数据作为检索依据。同理，音频、视频对象也有同样的问题，也需要类似的处理步骤。杂乱无章的信息是无法有效检索的，为了实现检索功能，必须对存储对象进行管理，也就是建立有效的索引。

这样的信息系统还要求信息传递方法是网络化的，对存储对象的存储、检索等操作通过网络进行。网络化信息系统面临的一个主要问题是信息对象的类型标识。虽然通过因特网能够传递丰富多样的信息对象，但是其基础却是相当简单甚至脆弱的若干协议。最早的超文本传输协议（Hyper Text Transfer Protocol，HTTP）只支持超文本（Hyper Text Mark-up Language，HTML）文

档，为传递多种类型的信息对象，借用了多用途网际邮件扩充（Multipurpose Internet E-mail Extension，MIME）协议，才能传递图像、音频、视频等信息对象。简单地说，信息对象的 MIME 类型类似于文件的扩展名，操作系统可以通过扩展名来寻找对应的处理程序，而因特网则通过 MIME 类型来标记不同的信息对象。更重要的，MIME 类型也如同文件的扩展名一样，是可以根据需要增删的，因此，一方面，越来越多的化学信息对象有了自己的 MIME 类型，另一方面，很有可能这些类型并未包含在通用的 MIME 类型范围之内，从而导致传递失败。网络化信息系统面临的另一个主要问题是超时。一般的存储、检索等操作基本上以秒或者分钟为单位，问题不大，但化学研究往往需要进行几小时甚至几天的计算，这种情况下就要采取特别的措施了。

因此，从通用信息系统到化学信息系统，特别需要在化学信息对象的数字化与检索化、MIME 类型定义与超时操作处理等几个方面，开展更多的工作。此外，在用户界面方面，化学信息对象也需要有针对性的开发技术，从而使化学信息系统有用、能用、好用。

1.2 化学信息系统概况

化学信息对象类型广泛，既有文献，又有图谱、分子结构、反应式等，所以相应的处理方式也是多种多样。例如，计量学方面的数据处理，有投影、降维、傅里叶变换、小波变换、建模等，信息学方面有分子结构匹配、指纹比对和结构参数生成等，计算化学方面有轨道、电荷密度计算、能量分析与结构优化以及动力学模拟等。因此，化学信息系统即是以化学信息对象的数据库和在线计算为核心的网络服务系统，并带有特殊的用户界面，功能需求复杂。

1.2.1 化学信息系统的主要类型

化学信息系统有很多种类型，在此并非要像参考手册那样列多少种化学信息资源，而是按照功能来介绍不同类型的系统。

1.2.1.1 化学文献

化学文献信息系统有很多，既有专业学会旗下的，如美国化学会（American Chemical Society，ACS）、英国皇家化学会（Royal Society of Chemistry，RSC）等，也有大型科研机构主办的，如 Pub Med/Entrez，当然更多的来自文献出版商，如 Elsevier B. V. 的 Science Direct、John Wiley & Sons，Inc. 的 Wiley Inter Science、Springer 的 Kluwer Online、Ingenta 的 Ingenta Connect 等。此

外，信息技术公司与出版商合作，也推出了 Amazon、Scirus 和 Google Scholar 等。目前，网上化学文献正在逐步取代印刷文献，这些信息系统也都具备在线检索功能。

除了搜集的文献范围有所不同之外，绝大多数文献信息系统的功能大同小异，都只是给出作者、文章或书籍的标题与出处、内容摘要以及关键字等。有些还能提示相关文献和展示一些样张。只有购买了相关文献全文授权的用户，才能查看和下载更多内容。绝大多数文献的全文文档，采用的是兼容性很强的超文本或 Adobe 公司的可移文档格式（Portable Document Format，PDF）。

此外，很多系统还推出了“学术趋势”这类服务，用户可以查看与特定关键词相关的文献分布、增长情况，迅速了解相关学术领域的发展形势。例如，我国的国家知识基础设施（China National Knowledge Infrastructure，CNKI）网站也可以给出“学术趋势”图（图 1-1）。

图 1-1　CNKI 的学术趋势图

鉴于读者对此类信息系统都相当熟悉，就不再多做介绍了。但是，通过分析读者对一篇文献有哪些要求，就可以看出化学文献信息系统目前还存在什么问题。

（1）从大量文献中找到特定的文献

文献信息系统能够很好地采用通用技术来管理大量的文档资料，也都具备多种文字检索功能，如作者、标题、关键词以及全文词频检索等。但是，没有考虑图谱、数据表、分子结构、反应式等化学文献中经常出现的信息对象，也

就是无法查询某幅图谱或者某个分子结构来自哪篇文献。后面要提到的SciFinder Scholar、Crossfire Commander和Reaxys等部分地解决了这个问题。

再有，读者如何管理自己搜集到的文档。检索文献时往往需要从不同的信息系统下载若干PDF（或HTML）文档，其文件名本应由信息系统提供，可实际上却总是得到密码般的结果，如ar700192d.pdf、b602241n.pdf，且下载时用户也很难赋予适当的文件名。结果，面对一堆文档却找不到所需的，更不用说查找图谱或分子结构了。专门的文献管理软件，如EndNote，算是一种补救措施，但仍不能解决图谱或分子结构这类问题。

另外一个可能的漏洞，是读者根本没有访问某个文献信息系统，从而导致检索不完全。对于查新工作来说是不小的挑战。

（2）阅读其摘要或者全文

这是最核心的要求。目前，读者利用多种软件，可以做笔记、标注或者与作者沟通。然而，如果一位读者发现了文献中的问题，除了告知出版商和作者以外，能否及时地在文献信息系统中留下自己的意见？出版商和作者在对文献做出修正之后，除了刊登声明之外，能否对原来的文献也进行标注？因为读者看到一篇文献之后，并不再查看是否后来又有什么修正。

（3）找到该文的参考文献

文献的HTML文档大都为参考文献提供了超级链接，而PDF文档虽有类似的功能，却没有提供。

此外，很多文献信息系统都提供了按照关联度列出相关文献的功能，非常有用。但是，这个功能只能在线使用，读者不能从下载的文档中得到相关文献的功能。其实，文献信息系统完全可以动态地修改文档的关联信息。

（4）在写作时引用该文

论文写作过程中，参考文献的格式问题往往耗费很多时间。大多数文献信息系统都提供了多种引用格式。此外，前面提到的文献管理软件，如EndNote以及ReferenceManager等在这方面也是非常有用的工具。

（5）使用该文中的数据

化学文献的数据主要体现为图谱、数据表、分子结构、反应式等化学信息对象。很多文献信息系统都支持和鼓励作者，在正文之外，把上述内容存储为支持文档（Support Information），供读者自由下载。但目前很多支持文档也用PDF格式，导致这些化学信息对象达不到检索化的要求。

1.2.1.2 化学物性标准数据

物性标准数据是建立在大量科学实验基础上的可靠资源，对化学研究具有

重要的价值。因此，能够快速准确地提供有效的数据，是这类信息系统的主要目标。

● Chemistry WebBook (http://webbook.nist.gov/chemistry)

Chemistry WebBook是美国国家标准与技术研究院（National Instituteof Standards and Technology，NIST）的标准参考数据库（Standard Reference Data）中与化学有关的物性数据库网络版（图 1-2），包括气相热化学数据、凝聚相热化学数据、相变数据、反应热化学数据等，还有红外光谱、质谱、紫外可见光谱、电子能谱等，支持多种检索方法。

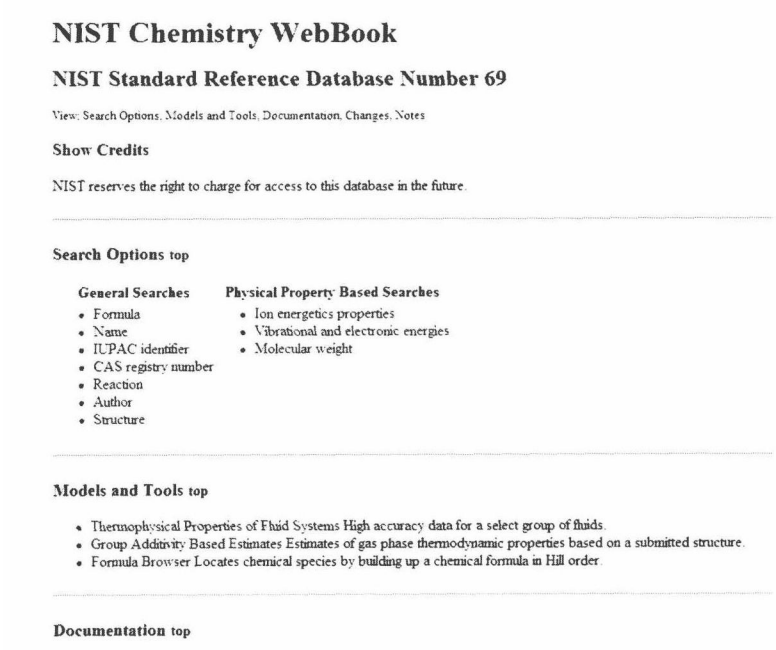


图 1-2

● CODATA (Committee on Data for Science and Technology) (http://dc.codata.cn/)

国际科学与技术数据委员会（Committee on Data for Science and Technology，CODATA）是1966年成立的一个跨学科委员会，其宗旨是提高所有科技领域内重要数据的质量，改进这些数据的管理，并扩大科技数据的可获取性，所有数据均在网上公开（图 1-3）。我国于1984年加入该委员会。

物性标准数据绝大多数是数字，因此对于界面操作没有特殊的要求。但是

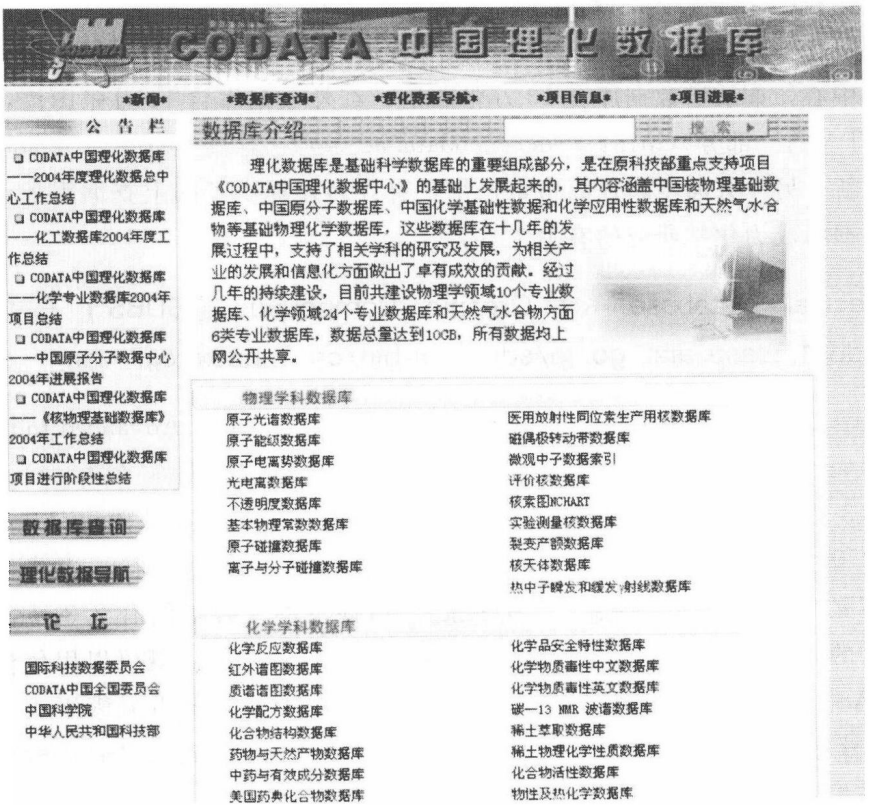


图 1-3

有两个方面的问题需要注意。一是这些数据信息可否按照一定的格式（如 Excel 文件）下载。二是如果要提供化合物二维结构式的检索功能，仅靠浏览器是不行的，还必须利用一些插件。

1.2.1.3 化学图谱

通过分析测试仪器获得的图谱，作为数字化的化学信息对象，可以用来建立相应的信息系统，如图谱数据库检索系统。国外的图谱数据库起步较早，以质谱、核磁共振、红外和 X 射线衍射图谱为主，如美国国家标准与技术研究院（National Institute of Standards and Technology，NIST）、环保局（Environmental Protection Agency，EPA）和国家健康研究院（National Institutes of Health，NIH）联合建造的质谱数据库，德国的 BASF 核磁共振谱数据库，美国 Fiveash Data Management（FDM）公司的傅里叶（Fourier）变换红外（FT-IR）数据库，美国国际衍射数据中心（The International Centre for Diffraction Data，ICDD）的粉末衍射标准图谱数据库（Powder Diffraction File，PDF）等。

中国科学院的化学研究所、长春应用化学研究所、大连化学物理研究所、上海有机化学研究所等也开发了一些大规模的图谱数据库，建立了中国化学专业数据中心，如质谱数据库含 13 万张图谱，红外数据库含 14 万张图谱，成为 CODATA 的一部分（http：//dc. cncodata. ac. cn/）。这些数据库一方面不断充实、更新，另一方面不断提高检索技术，有许多还提供了网上查询服务，使得图谱数据库成为化学研究的有力工具。

● Spectral Database for Organic Compounds（SDBS）（http://riodb01. ibase. aist. go. jp/sdbs/cgi-bin/cre _ index. cgi? lang = eng）

这是由日本产业技术综合研究所（National Institute of Advanced Industrial Science and Technology，AIST）支持的一个免费图谱数据库（图 1-4），集成了大约 3 万 2 千多个有机化合物的质谱（2 万 3 千多张）、氢核磁谱（1 万 4 千多张）、碳 13 核磁谱（1 万 2 千多张）、傅里叶红外谱（5 万多张）、激光拉曼谱（3 千多张）以及顺磁共振谱（5 万多张），所涉及的化合物约有 2/3 是含 6 到 16 个碳的有机物，图谱数据主要来自商品试剂。在检索时，既可以用化合物名称、CAS 注册号、SDBS 编号，又可以用分子式、分子量或者元素组成，还可以输入核磁化学位移、质谱峰位置与强度。用名称、分子式或号码检索时，还可以用“%”作为通配符。

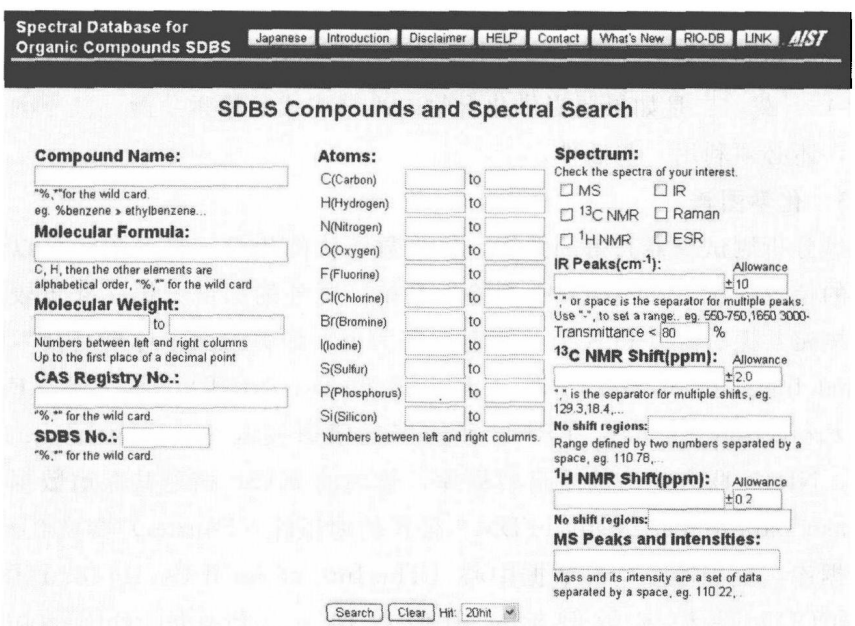


图 1-4