解説

荒木啓介*

JOIS-Fデータベースから JOIS文献ファイルへの渡り検索 —JICST化合物辞書の効率的な使い方—

Crossfile search of the databases in the systems of JOIS and JOIS-F —An effective use of JICST Chemical Dictionary Database—

— ARAKI Keisuke —

Commentary article

[著者抄録] 化学情報の高度化の必要性、JICSTの化合物辞書システムの目的、開発の経緯、特徴について述べた。情報源別の登録件数、1985年度の登録開始以来の統計、1989年度から試行的に始めた文献の物質索引の方針と概要、辞書検索上の注意点、ファクト・データベースへの渡り検索および文献データベースへの渡り検索と物質索引の有効利用の方法について解説した。オンライン端末のエミュレーターに化合物辞書検索の結果をキープして、JOISの文献ファイルに接続し、一般のキーワードと組み合わせる文献検索が可能となった。物質索引以前のフリータームで索引された分も、化合物辞書により薬物等の多様な名称を知ることを通して、より網羅的な検索が可能となる。

[著者付与キーワード] 化学情報、化合物辞書システム、化合物検索、物質索引、JOIS-F、JOIS、渡り検索、JICST

[Author Abstract] Necessity to enhance chemical information system, the objects and the historical courses of JICST chemical substance dictionary system are reviewed. The features of the system as well as the information sources, statistical data, the policies and outlines of chemical substance indexing of literatures experimentally began in 1989 are introduced. Some technical points for searching and effective use of the chemical substance dictionary in relation to the correlated factual databases and indexed literatures are explained. Though cross retrieval between JOIS and JOIS-F within a dialogue session is at present impossible because they are loaded on the different host computers, some terminal emulators support to keep retrieved substance numbers of the dictionary and to switch them to the literature database and thus enable cross search. The chemical substance dictionary also indicates many synonymous expressions of such substances as drugs and pesticides, which helps to retrieve the old literatures indexed by free terms of such expressions.

[Keywords by Authors] chemical information, chemical substance dictionary system, chemical compounds retrieval, chemical substance indexing, JOIS, JOIS-F, cross search, JICST

* 日本科学技術情報センター (〒102 千代田区四番町5-3) Tel. 03 (5214) 8401
JICST (5-3 Yonban-cho, Chiyoda-ku, 102)

1. はじめに

物質に関する学問ないし技術であり、知識や経験の積み重ねに依存する率の高い化学においては、その研究対象の中心をなす化合物を正確に記述・蓄積し、これを様々な名称（体系名、一般名、慣用名、商標名など）はもちろん、任意の構造上の特徴からも自由に検索でき、その合成等を記載した文献情報、さらにその物性値や生物活性データの取得につなげるサポート機能が必須である。ドイツ化学会が行った調査¹⁾によると、化合物に関わる情報要求の内容は以下のとおりで、この傾向は今でも変わっていないものと思われる。

① 個別化合物	25%
② 部分構造	23%
③ 反応	22%
④ 物性	7%
⑤ 用途	5%
⑥ ①～⑤の組み合わせ	17%

このために、計算機が出現するはるか前から、印刷物の抄録・索引誌、様々な形のデータ集、ハンドブック類が刊行されて来た。多様な分子構造を持つ化合物に共通の類似性について検索するためにも、体系名の標準化や倒置法による配列、分子式や示性式の利用、特殊な分類やコード法の開発等、多くの工夫がなされて来た。しかしこれらの方法には限度があり、分子構造を任意の特徴や部分構造から自由に検索するためには、化学物質を一旦その最も情報量の多い分子構造そのものの形で蓄積しておくことが必要である。

1960年代に世界主要国の情報サービスや大化学企業においてはこのような課題で研究が行われ、米国 CAS の Registry System、ドイツ IDC の GREMAS、フランスの DARC、スイスの BASIC、イギリス ICI 社の CROSSBOW と次々に独特な化合物構造の蓄積・検索システムが開発され、実用化された²⁾。日本では、JICST においても基礎的な研究が行われたが、化学構造を計算機に効率良く入力する方法および検索におけるスクリーニングの仕組みに見通しが得られず、それに当時の計

算機のキャパシティー的な限界のため、開発研究は中断された。この間、いずれも化学を主たる対象分野とする上記のシステムは70年代を通して着実に発展し、同時期に著しい進歩を示した計算機と通信の技術に支えられ、80年代後半には先に述べた化学研究における必須のサポート機能がほぼ達成されたかの観がある。

JICST において開発が遅れた別の原因として、科学技術のほぼ全域に及ぶ対象分野を、統一分類と統一シソーラスによって統合したデータベースにするという基本方針に、高度な例外処理を伴う化合物のインデクシングが馴染まなかったことがある。東北大学で1926（昭和2）年に創刊され、明治10年に遡って国内の化学文献の抄録を網羅し、英語名 Complete Chemical Abstracts of Japan を持つ日本化学総覧は³⁾、1964（昭和39）年に JICST に移管後もほぼその形式を踏襲する努力が続けられたが⁴⁾、上記の大前提に従ってその優れた化合物名索引と共に第47巻を以って1973（昭和48）年に活版印刷を止め、シソーラス方式のキーワード索引に移行した。化学物質検索については化学の分野、しかも国内化学の分野でのみ高度化したのでは中途半端であり、他方、オンラインによる機械検索の利点を総合的に生かすためにはシソーラス方式への移行が必要な選択であった。むろんこの時点でシソーラス方式によっても化学物質検索に特徴を持たせるべく工夫を施したが⁵⁾、これは先ごろ行われたユーザ訪問による調査でも一定の肯定的な評価が得られていることが分かった。

しかし、JICST が総合的な文献データベースを作成しオンラインによる提供を続けるのであれば、その中で化学は依然として潜在的利用度の高い分野であるから、何らかの高度化が必要であることは自明である。そのために、別途化学物質索引の高度化の研究は続けられた⁶⁾。

2. 化学構造入力的基本的な考え方

化学物質索引を高度化するためには、構造情報

をいかに効率よく入力するか、という課題を解決する必要がある。欧米のシステムでは、化合物構造の計算機内での表現形式である原子結合表を、順次、人手による直接コーディング、構造タイプライターの利用、CRTからの構造直接入力などの方法によって作成してきた。しかしこのプロセスには多大の手間と時間がかかる。何故なら、化学物質のデータベースであるからには、構造情報と共にその名称（体系名、慣用名など）を入力する必要があるため、専門の化学者が命名し、入力された構造図と照合・チェックする作業を伴うからである。構造入力専用のシステムや機器を整え、人員を教育・配置し、複雑な作業工程を管理・運用することも並大抵のことではない。全化学構造を入力しなくても、かなりの部分構造検索を可能とする方法として、WLN (Wiswesser Line Notation) や DERWENT 社の Ringcode, IDC の GREMAS コードなどがあり、場合によってはこれらからプログラムによって化学構造を発生させる可能性もある。しかし名称を別途入力する必要性は変わらないし、何よりもこれらのコード法は訓練・習熟が必要で、人事異動などを考慮すると JICST ではとても採用できない。そこで着目したのが、化学を専攻した者であれば程度の差こそあれ基本的な素養があり、一般に馴染みのある有機化合物の体系名である。これは一定の語彙と文法を持った人工言語であるから、対応する部分構造を持った語彙を解析辞書とし、命名法を計算機プログラムとしたシステムを作成できれば、体系名という文字入力から名称ファイルと分子構造ファイルの両方を同時に作ることができる。この方法は既に CAS にて開発されていたが⁷⁾、分子構造に変換できる体系名の範囲が必ずしも広くないようである上、詳細は不明であった。我々は独自の体系名の解析組み立てシステムの検討を始め、ほぼ全ての種類の体系名を解析し分子構造に組み立てることができるシステムを目指して体系名の分析とアルゴリズムの考案を進めた。

折からの緊縮財政のため、1980 (昭和55) 年度の JICST の本予算では認められなかったが、翌

1981 (昭和56) 年度の振興調整費プロジェクト「ネットワーク共用による化合物情報等の利用高度化の研究」では、オーソリティー辞書として CAS の Registry File に代わるものの重要性が認識され、JICST で化合物辞書を組織的に開発することとなった。欧米から20年近くも遅れたスタートで全く同じようなものを作る不利は承知しており、それらには無い特長、すなわち立体化学的な分子構造の自動作成と自動登録、その図形表示および日本語による入出力機能の実現を目指した。システム分析から設計・プログラム作成、解析辞書のコーディング等に5年を要し、1985 (昭和60) 年に世界で初めての上記の機能を備えたシステムが完成した。振興調整費プロジェクトに参加した専門データベースの対象化合物、約1万を登録して公開デモを行うと共に、引き続き登録化合物数を増やし、1987 (昭和62) 年1月から JOIS-F としての有料サービスを開始した。さらに1989 (平成元) 年からは、一部の文献の物質索引を開始した。20年前とは状況が相当変化しているとは言え、JICST が提供する文献データベースにアクセスする手段としての物質索引の必要性は、基本的には変わらないのである。システムの詳細については情報科学技術研究集会等で発表しているので⁸⁾⁹⁾、その方を参照されたい。

3. 化合物辞書登録の推移、辞書データベースの概要と特長

3.1 辞書登録の推移

JICST の化合物辞書は専門データベースのネットワークにおけるオーソリティー辞書としてスタートし、後から文献の物質索引の典拠辞書と位置付けられたため、蓄積されたデータもそれを反映したものとなっている。即ち、1985年からファクト系データベースが対象としている一般的なあるいは重要な化合物をまず登録し、次いで1989年から部分的に開始した文献の物質索引により最新の文献に記載されている化合物が順次追加された。辞書登録の実績は、表1の通りである。

3. 2 ファクト系データの情報源

(1) 振興調整費プロジェクトおよびその後参入の
専門データベースの対象化合物

現在我が国で作成・提供されている主要な専門データベースの概要は表2の通りである。これらのファクトデータ（様々な物性データや専門的な事項に関するデータ）は、4, 7, 10, 12番のデータベースについては、JOIS-Fによって、3はケミカル・データ社のオンラインでサービスされている。その他はそれぞれ独自の方法で提供されている。これらの情報源は、化合物辞書の出力にロケータとして表示され、これらのファクトデータベースと論理的に結合していることを示している。

(2) JICSTでは実用規模の化合物に関する典拠リストとしての品質を確保し、より一般性を持たせるために、上記の他に日本、米国、欧州の以下の法令によって規制の対象となっている化合物を収録している。

日本の「化学物質の審査及び製造等の規制に関する法律」（化審法）の関連 約12万

日本の「労働安全衛生法」（労衛法）の関連

約2万

米国のToxic Substance Control Actの関連

約3万2千

米国、保健福祉省のRegistry of Toxic Effect of Chemical Substances

約4万

欧州のEuropean Inventory of Existing Commercial Chemical Substances

約7万

化審法、労衛法については絶えず官報をウォッチし、可能な限り早急に追加している。これらの法令によって規制される化合物は、各法令独特のコード番号が入力される。

(3) 便覧類、ハンドブック類

上記(2)と同じ理由で重要な医薬品、農薬、染料、食品添加物、天然物等を網羅する意味から、Merck Index, USAN（共に医薬）、NANOGEN（農薬）、Colour Index（染料）などに収録されている有機化合物は全て調査し必要なものは収録している。したがって登録化合物数は約60万と少ないが、実用されている低分子有機化合物について

表1 化合物辞書登録の推移

年度	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	累計
ファクト	9,115	24,997	57,539	69,622	76,625	60,337	14,739	9,176	12,273	5,974	340,397
文献索引											
雑誌数					23	29	491	665	586	964	
文献数					560	8,153	29,320	41,374	33,809	25,563	138,779
物質数					36	37,704	53,746	75,234	54,140	34,589	255,449
登録物質	9,115	24,997	57,539	69,622	76,661	98,041	68,485	84,410	66,413	40,563	595,846
合計											
蓄積数	9,115	34,112	91,651	161,273	237,934	335,975	404,460	488,870	555,283	595,846	

注1) 89年度は2か月間のみ。

注2) 94年度は9か月分

はその全てを網羅している。

3.3 対象化学物質

入力対象として可能な化学物質は、当面以下の範囲のものである。

(1) 平面構造以上で分子構造が明確な低分子有機化合物。相対的にのみ立体化学が明確なものも含む。

(2) 分子を構成する水素以外の原子数が247以内のもの。

この範囲内の、有機化合物の無機酸エステル、有機金属化合物、オリゴペプチド、オリゴ糖を含む。

(3) 以上の有機化合物を主成分とする混合物、ラセミ混合物、付加塩、オニウム塩、水和物を含む。

現在のところ、一般の無機化合物、錯体、高分子化合物、合金、蛋白質、多糖類、ポリヌクレオチド等は対象外である。これは後述する文献の物質索引の場合も同じである。

3.4 化合物辞書のエントリー

化合物辞書のデータ項目および検索可能項目の概要は表3の通りである。

1 化合物エントリーの出力例を、図1に示す。

内容についてのこの他のコメントは、検索上の注意と共に、5項に記す。

3.5 化合物辞書の特長

(1) 立体化学的な構造の入力、検索と表示

R/S方式および α/β 方式のいずれによって記述された体系名も解析し、立体化学的な検索と図形表示を可能とする。立体を考慮しないと極めてノイズが多くなる糖類の検索も有利である。

(2) 日本語による化合物名検索機能

エチホキシシなる薬物名に対し、この英語名を推定しようとする、

チ: ti, ty, thi, thy, chi, chy

ホ: ho, fo, pho

キシ: xi, xy, ksi, csi

ン: ne, n

合計 $6 \times 3 \times 4 \times 2 = 144$ とおりの可能性があり正解を得るまで時間を要する。

(正解は etifoxine)。カタカナではほぼ間違いなく記述できる。

また、酢酸、クエン酸のような身近な物質名はわざわざ英語でアクセスするのでは不便である。

体系名についても、

表2 ファクトデータベースの概要

コード	データベース名	作成機関	対象化合物の特徴
1 BC	バイオケミカル	農林水産省食品総合研究所	栄養素, 食品添加物等
2 BL	バイオリジカル	国立衛生試験所	変異原性, 発癌性, 催奇形性物質
3 CD	化学商品	化学工業日報社	化学商品
4 CR	結晶構造	JICST	重要化合物
5 EN	環境物質	国立環境研究所	環境に放出された物質
6 IS	医薬品産業	医薬品産業研究所	医薬品特に治療薬
7 MS	質量スペクトル	日本質量スペクトル学会, NIST	重要物質
8 PE	農薬	農薬工業会	農薬
9 PH	医薬品	日本医薬品情報センター	医薬品
10 SF	化学物質法規	JETOC	法規制化学物質
11 SP	スペクトル	工業技術院物質工学研究所	基礎的な物質
12 TH	熱物性	JICST	基礎的な物質
13 TK	試薬カタログ	東京化成(株)	試薬
14 TX	急性中毒	日本医薬情報センター	医薬品

benzenamine naphthalen-1(4H)-one などの場合は骨格名の末尾の e は脱落するが

benzeneacetic acid の場合は e は脱落しない。その理由を含めて知っている人は化合物命名法の相当なベテランであるが、一般のユーザには難しい。

以上の状況を見ても、英語名と並んで日本語による化合物名の検索機能も用意しておくことは必要であり、JICST 化合物辞書の第 2 の特長である。

(3) 国内で発生する化学情報のオーソリティー辞書

後述する文献の物質索引では、重要な外国雑誌と共に国内誌を網羅的に対象とし、1996(平成 8)

年度から全面的に実施することを計画している。CA に比べても国内誌に関しては網羅性で勝ることは確実であり、本辞書利用上の最大の特長となる。

4. 物 質 索 引

JICST では科学技術文献データベースとして蓄積する文献において主題をなす化学物質を、化合物辞書によって索引する方向を目指している。これは化合物辞書の側から見ると今や最大の情報源となっている。

表 3 化合物辞書のデータ項目と検索可能項目

項番	表示項目名	記号	内 容	検 索 法 概 要
1	化合物辞書番号	SN	登録した物質に自動付番した背番号 (日化辞番号) 形式: JZZZ.ZZZ.ZZ9X J: 固定, Z: 数字で 0 は省略, 9: 数字 X: チェックディジット文字	SN :
2	物質クラス	—	SN の後カッコ内に表示。 C: 構造あり, X: 混合物, T: 名称のみ	TY :
3	ロケータ	LC	関係付けられた専門データベースのコード, 文献に対し索引された物質にはロケータは無い。	LC :
4	CAS Reg. No.	RN	CAS の Registry Number 形式: ZZZZZ-99-9	LC :
5	分子式	MF	Hi11方式 元素記号 C H ABC の順	MF: 全体または部分の分子式
6	分子量	MW	小数 3 桁まで	MW= 2 次サーチ
7	構造図		グラフィック端末で標準形式を表示	質問構造の作図または構造の特徴を示すキーによるサーチ
8	体系名	NM	IUPAC 命名法規則に基づく日本語表記	全体または名称フラグメントの直接入力。完全, 前方後方および中間任意一致
9	慣用名		正規の体系名以外の, 日本語, 英語の各種名称 一般名, 通俗名, 商標名, 治験番号, 官報公示整理番号, 各種コード番号。 英語体系名もこの下に表示	直接入力 一致条件は体系名と同じ。英語体系名はフラグメントも検索可
10	旋光方向符号		+ または - で表示	OR= 2 次サーチ

化合物辞書検索システム

DATE: 95. 06. 29 PAGE: 1
JICST COPYRIGHT

#00001

SN=J9. 266E (C)

ロケータ=BL (バイオロジカルDB); MS (質量DB); PE (農薬DB);
SF (化学物質法規DB); TH (熱物性DB); TK (東京化成DB)

RN=54-11-5

分子式=C₁₀H₁₄N₂

分子量= 162. 236

旋光性= -

体系名= 3 - [(2 S) - 1 - メチル - 2 - ピロリジニル] ピリジン

体系名= (-) - 3 - [(2 S) - 1 - メチル - 2 - ピロリジニル] ピリジン

体系名= 3 - [(2 S) - 1 - メチルピロリジン - 2 α - イル] ピリジン

体系名= 3 - [(2 S) - 1 - メチル - 2 α - ピロリジニル] ピリジン

体系名= (-) - 3 - [(2 S) - 1 - メチル - 2 α - ピロリジニル] ピリジン

体系名= (2 S) - 1 - メチル - 2 α - (3 - ピリジニル) ピロリジン

慣用名= β - ピリジル - α - N - メチルピロリジン

慣用名= β - Pyridyl - α - N - methylpyrrolidine

慣用名= [S , (-)] - ニコチン

慣用名= [S , (-)] - Nicotine

慣用名= (-) - ニコチン

慣用名= (-) - Nicotine

慣用名= (-) - 3 - [(2 S) - 1 - Methyl - 2 - pyrrolidinyl] pyridine

慣用名= (-) - 3 - [(2 S) - 1 - Methyl - 2 - pyrrolidinyl] pyridine

慣用名= (S) - ニコチン

慣用名= (S) - Nicotine

慣用名= ニコチン

慣用名= ブラックリーフ40

慣用名= Black Leaf 40

慣用名= 1 - ニコチン

慣用名= Nicotine

慣用名= 3 - [(2 S) - 1 - Methyl - 2 - pyrrolidinyl] pyr
idine

法規制番号= (9) - 990 ← 化審法番号

法規制番号= ENCS2001933 ← EINECS番号

法規制番号= QS5250000 ← RTECS番号

法規制番号= TSCA (54 - 11 - 5) ← TSCAで規制の対象 () 内はCAS Reg No.

法規制番号= 8 - (1) - 675 ← 労安法番号

図1 化合物辞書の出力例

4. 1 物質索引の対象雑誌

JICST では約9,400種の雑誌類を収集・処理し、年間約70万件の文献データベースを作成しているが、主題が低分子有機化合物に関連する文献は、約2,000誌、73,000文献程度と見積もられている。さらにこれらは、複雑な新規化合物が多数出現する基礎化学分野の雑誌から、平均2～3個程度の、しかも医薬・農薬のように通常は一般名・慣用名で表わされることの多い雑誌まで幅広い。後者は体系名と共に一旦化合物辞書に登録されてしまえばその後は索引作業は楽になり、既登録リストやJOIS-F検索で確認し、その慣用名のみを記入すれば済む。新規の場合のみ体系名を組み合わせる。ファクト系データの登録で既に主要な慣用名は相当登録されているので、1991年度の統計調査によっても、この種の平均的な雑誌では3個のうち新規化合物は0.3個以下である。物質索引が困難なのは基礎化学の雑誌で、平均13個もある対象化合物で、慣用名を持つものは0.5個と少なく、しかも既登録は0.1個であって残りのほとんどは、複雑な構造を体系名で記述する必要のあるものである。この状況を図2に示す。

現在のシソーラスによる索引方式を継続したま

図2 物質索引の実態
—ランク別、出現化合物パターン

ま物質索引担当者の教育・訓練を on the job 式に行うにも、新たな要員を確保するためにも、関係者にかかりの負担がかかるので、どうしても比較的簡単なものから始め、時間をかけて徐々に量を増し難易度を上げていく方式を採らざるを得ない。以上を勘案し、物質索引は以下の方針および年次計画で推進しつつある。

- (1) 医学・薬学関連の雑誌の早急な全面物質索引の実施。これは1992年度から既に実現している。
- (2) 農薬を含むその他の応用分野への可能な限りの物質索引の拡張。
- (3) 有機化学分野での重要な雑誌からの実施。
(1989年度からの試行を通し、中断された雑誌がいくつかある)

4. 2 文献の主題パターンにおける役割

文献の主題を構成する概念の中で、化学物質が重要な役割を果たしている場合は、その対象化学物質の種類、主題パターンごとのその役割などから物質索引の基準を決めている。主題パターンとしては大略以下を想定し、各々について化学物質の役割を考慮する。実際にはこれらの組み合わせとして出現する。これらが原著論文の場合に、必ず索引するものに○印を付けて示す。それら以外には重要な場合のみ索引する。

主題区分 化学物質の役割

合成、製造：出発物質，中間体，○生成物，溶媒
最終目標化合物

反応研究：反応物質，○キーとなる物質，
○生成物，溶媒，触媒

構造研究：○対象化合物（新規および既知）

分析、物性研究：○対象化合物，試薬

生物活性研究：○対象化合物，比較化合物

薬物等の代謝：○対象化合物，代謝・排泄物質

用途等：○対象化合物

これらの数が多い場合は、やむをえず最も重要なものの30個程度に制限する。

また、レビューないし解説記事で、その数が極めて多数である場合は、個々の化合物としては索引せず、別途シソーラス語の類名で索引する。

4. 3 化合物索引の方法

文献中では化合物は、1)構造図、2)体系名、3)慣用名およびこれらの任意の組み合わせとして出現する。このうち、3)が化合物辞書に既登録であれば、一般には短いその名称のみを記入すれば足りる。未登録の場合はこれに2)を組み合わせで索引する。1)のみの場合は命名規則に従って2)を作成する必要がある。いずれの場合にも、文献中のデータは全てが正確であるとは限らないので、厳密なチェックを行っている。

なお現在は、シソーラスによる物質の索引を併用しており、構造検索や体系名のフラグメント検索には馴染まない観点、例えばテルペン、アルカロイド、アミノ酸などの類名から検索できるようになっているが、この長所は将来、全文献を化合物辞書方式で索引するようになってシソーラス中に残して、索引することになっている。この場合、現在のシソーラス中の化合物タームには、部分構造を示すのか類名を示すのかがあいまいな、脂肪酸やアルコールのような語があるので、そのスコープを類名に限定した上で索引されることになるが、対象化合物を限定できるものについては、個別の化合物との対応付けによる自動付加の可能性も検討しつつある。

5. 化合物検索の概要と注意点

有機化合物はその分子構造に従う古典的な分類の仕方があり、これを全く無視した検索要求も無いわけではないが、基本的にはこれに沿ってこれらに適した検索機能を用いるようにした方が効率が良い。詳細はJOIS-F活用の手引きの「JICST化合物辞書データベース」を参照いただきたい。ここでは主として検索戦略的な問題や注意点に絞って話を進める。

5. 1 名称検索

化合物の何らかの名称、一般名、慣用名、商標名、治験番号、コード番号等あるいは体系名そのものまたはその一部から、該当物質の構造や付随

する情報を知りたい場合に名称検索を行う。これらの名称については、英語名と日本語名の両方で検索可能である。日本語名は日本化学会の化合物日本語表記規則に準じて和訳しているが、慣用的な日本語表記があるものはそれらも入力し、アクセスの多様化を考慮している。部分一致条件も前方、後方、中間任意の3条件が適用できる。部分一致指定の場合は時間を要するが、例えばビタミンD₂のような前方一致指定でD₁からD₄や関連誘導体をまとめることが出来る。HOE- $\&$ とするとヘキスト社の治験番号を持つ化合物の集合を作れる(治験番号は文献中にバラついた形で出現するが、治験番号の特定英文字と後続の数字の間にハイフンを入れた形を採用している)。後方一致は $\&$ マイシンのように指定すると多くの抗生物質をまとめることができるが、語尾がこの形でないものや、 --- マイシンA₁のようにさらに後がある場合があるので注意が必要である。 --- グルコピラノシドや --- カルボアミドのように、化合物名の末尾であることが確実な文字列に適用することが肝要である。

体系名の検索では、短いものはそのまま入力するが、一般的にはそのフラグメントを検索しその組み合わせで対象を絞る。これらの場合以下の点に注意が必要である。

- (1) 体系名の要素でもある名称で、その単独のものを求める場合

ベンゼン、酢酸などは体系名の要素でもあるので、そのまま入力するとヒット件数は大きな数字が表示される。この場合2通りの対処法がある。

- (a) 分子式による検索結果を掛け合せる

例えば ベンゼン の名称による検索結果を

MF: C 6 H 6

による検索結果と掛け合せる。

- (b) $\&$ E コマンドを適用し、後方一致指定で文字列を入力する

酢酸 の場合は

$\&$ E $\&$ 酢酸

と入力する。列挙される検索語のうち単独

の酢酸そのものが常に先頭に表示されるので、/01で該当するその1件を得ることができる。

- (2) 体系名はまずカンマ、カッコ、ハイフン等で、次いで得られた名称のみの文字列を化学的に意味のある単位で以下のように切断してそのフラグメントを得ている。

2, 3-ジニトロ-3-(2-クロロエチル)

ベンゼン酢酸メチル:

ジニトロ	ニトロ
クロロエチル	エチル
ベンゼン酢酸メチル	酢酸メチル
メチル	

これらのフラグメントが全て検索キーとなるが、切断の際、前方の要素名は捨てられるので、クロロ、ベンゼン、酢酸の単独形は検索キーにならない。したがって、クロロとエチルのあるいはニトロとベンゼンのAND検索ではこの化合物は検索されない。クロロエチルまでの文字列を指定する必要がある、あるいはニトロとベンゼン酢酸& (ベンゼン&のみでは時間がかかり過ぎる) のAND検索をするなどの配慮を要する。これらは名称検索の欠点であり、改善が急務である。

- (3) 体系名のフラグメントによる検索は、体系名の性格上、カッコレベルの異なった直接には関係のない部分がAND検索でヒットし、ノイズを生む可能性がある (ニトロとエチルの組み合わせで、これらが直接には結合していない上例の化合物がノイズとして得られる)。したがって名称フラグメントの組み合わせによる検索は、そのような名称で表記される要素が構造中に共存していることの要求と見るべきである。
- (4) 名称自体に多様性があり、構造的には同じであっても他の構造要素との関係で主官能基名であったり置換基名であったりするので、想定される可能性を網羅する必要がある。

(例1)

アニリン、ベンゼンアミン、アミノベンゼン、アミノフェノール、アミノフェニル、フェニ

ルアミノ

(例2)

カルボン酸、カルボキシ、カルボキシラート
以上を踏まえた上で工夫すれば名称検索で一定の良い結果が得られる。

5. 2 分子式検索

MF: の後に Hill 方式 (炭素数、水素数次いでその他の元素を ABC 順に並べる方式) の分子式を入力して該当化合物を検索する。分子式による検索の場合は一般に、原子数が多く複雑な構造のものほど該当する化合物が少ないため効率よい結果が得られるが、簡単なものほど該当する化合物が多くなるので、名称検索あるいは後述する構造上の特徴を表わすキーによる検索を併用する必要がある。

分子式は完全形だけでなく、部分分子式およびその組み合わせでも検索でき、また特定の元素を含まないとする NOT 条件を安心して適用できるが (文献の主題の検索の場合は安易に NOT 条件を適用するとモレが生ずることが多い)、水素を指定しないとどうしてもノイズを多く含む結果となる。有機酸の金属塩の分子式は原則として酸としてのプロトン水素が減りその金属を含めた形を分子式としているので、Chemical Abstracts とは異なるので注意されたい。

(例) 酢酸ナトリウム

JICST	$C_2H_3NaO_2$
CA	$C_2H_4O_2 \cdot Na$

5. 3 構造上の特徴キーによる検索

任意の構造の検索は最終的には質問構造の作図による構造検索に頼ることになる。しかし以下のような、構造上の特徴を表わすキーによる検索を組み合わせると、対象化合物の集合を事前に少数に限定して効率よく構造検索を行うことができる。多くの場合、構造検索をするまでもなく、該当化合物の小集合を得ることができる。

項目名	検索コードと検索例
-----	-----------

環 数	NR: 5 縮合環系の構成環数と単環の環数の和
環の大きさ	RS: 7/5 縮合環系では構成環の大きな順に/で区切る 一つのみ指定すると単環のみを検索する
環構成元素	RE: C3NS 水素を除いた Hill 方式で指定
環元素配列	CL: NCSCC アルファベット 順の若いヘテロ原子からつぎのヘテロ原子の方向へ途中の炭素 C を含め、元素記号を配列
鎖系数	CS: 3 枝別れしているものや官能基を含め一続きの鎖を 1 本と数えた総数
鎖系の形	FC: 7/3 鎖の全ノード数/鎖の末端数を指定

なお環を含まない鎖のみの化合物は鎖系数 1 の指定では不十分である。環に 1 本の鎖しかないものも含まれるからである。この場合は以下の方法によってほぼ鎖式化合物のみの集合を作れる。

[1] TY: C

によって、構造情報を持っているものの集合を作る。

[2] NR: 1 + NR: 2 + NR: 3 + NR: 4 + + NS: 12... によって、環を持つ化合物の大部分の集合を作る。

[3] 1#2

によって、大部分の環を除いた鎖系である集合を得る。

5.4 構造検索

¥GRA コマンドにより作図画面に変え、マウス、キーボードまたはメニューとして用意してあるテンプレート図形を利用して質問構造を作図する。詳細はマニュアルに譲るが、基本的な戦略として、上記 (5.1~5.3) の検索手段により求める化合物がその中に含まれる必要かつ十分なできるだけ小さな集合を作り、この直前の集合に対してのみ作用する、

¥EXEC *

のコマンドで検索すべき点を指摘したい。求める構造が比較的単純なほど、このような措置が必要であるが、逆に複雑な構造の場合は直接

¥EXEC

で全ファイルを対象としても良い。その他構造検索にはマーカッシュ的な検索や混合物検索、立体化学的な検索等特色ある検索法があるが省略する。

6. 渡 り 検 索

6.1 ファクト系データベースとの渡り検索

ファクト系データベースとの渡り検索は JOIS-F のホスト計算機の中でどの方向に対しても自由に行うことができる。化合物辞書からの渡りの場合は、ある集合を作ったら結果をセーブし、ファクトデータベースに渡ってその集合を呼び出して、ファクト系データとの相関検索を行うことができる。

[例]

[1] U: ビタミン C& (前方一致検索)

S: 14件

[2] U: ¥XSAVE VITC

(VITC は仮りの集合名)

S: [1], VITC を登録しました

[3] U: ¥CSTR TH

(熱物性データベースに渡る)

熱物性データベースがスタートしたら

[1] U: ¥XCALL VITC

によって化合物の集合を呼出し、熱物性データベースにおけるその物質の融点、沸点、溶解度、電離平衡定数等のファクトデータを取得する検索を開始する。

ファクトデータベースにおいては逆に、特定の物性値からこれを持つ物質の候補を探し辞書ファイルに渡ってその構造や別の名称を知ったり、さらに別のファクトデータベースに移ってスペクトルチャートを見たり、法的規制状況を知る手掛かりを得ることができる。

6. 2 文献データベースへの渡り検索

JICSTの化合物辞書システムは、前述したようにファクト系データベースシステムの中心的辞書として開発されたこともあって、その検索システムはJOISとは別のJOIS-Fとしてスタートした。その後JICSTファイルの物質索引を強化する観点から両者を統合することを検討したが、その改造には多大の費用を要するということで見送られた。幸いファイルを作成する部分は両者に橋渡ししのできたので、物質索引を開始しファイルを蓄積することが可能となった。1989年度末から開始し、1994年12月末で138,779件の化学文献を索引し蓄積した。JOIS-IIIの文献ファイル自体にも索引された化合物名がそのSN番号と共に物質索引セグメント(SW欄)に収納され表示される(図3)。

前述のように1992(平成4)年度からは医学・薬学分野の文献については全面的に化合物辞書方式による索引を実施しており、この分野の化合物検索には威力を発揮するはずである。未だ蓄積量が少なく、RSには十分ではないが、SDI検索としての利用には有効である。そのための、また物質索引の結果を利用して一般の検索で一定の効果を

得るための方法と注意点を述べたい。

6. 2. 1 JOIS 文献ファイル内での物質索引の利用

JOISの検索においては、シソーラシステムないしフリータムとして索引されている物質名、特に医薬品等の慣用名、商標名、治験番号等は、知っているままの名称を入力して大体は検索できる。化合物辞書を通して構造まで登録された物質名も、SW:で指定すれば直接検索できるし、SNが知られていれば、SN:の後に入力してこの番号の物質が索引された文献を検索できる。しかしJOISの文献ファイルで索引されている化合物は、その特定の物質に慣用名や治験番号が多数ある場合、たまたまその文献で表記されていた名称で索引されるのが原則なので、名称のみのサーチではかなりのモレが生ずる可能性がある。

これを救済するために物質索引の結果が利用できる。まず分かっている名称で検索し、ヒットした回答書からそのSNを知り、再度SN:の後にそのSN番号を入力して検索する。これにより、見掛けの名称は異なっても実際には同一の物質に関する文献を、より幅広く得ることができる。さらに次の段階として、物質索引は未だ4年分し

```
#000001* JICST COPYRIGHT
CN 94A0911243
TI 気管求心性神経からのCGRPに対するカプサイシン、ブラジキニン及びニコチン
  の作用の多様な機序 プロスタグランジン、交感神経及び肥満細胞の役割
ET Multiple Mechanisms for the Effects
  of Capsaicin, Bradykinin and Nicotine
  on CGRP Release from Tracheal
  Afferent Nerves: Role of
  Prostaglandins, Sympathetic Nerves
  and Mast Cells.
AU HUA X-Y, JINNO S, BACK S M, TAM E K, YAKS
  H T L (Univ. California, San Diego, CA
  , USA)
JN H0537A (NEPHB) (0028-3908) Neuropharmacology
VN VOL. 33, NO. 10 PAGE. 1147-1154 1994
CI (A) (a1) (EN) (GBR) (写図4, 表1, 参50)
AB ラット気管求心性神経からのCGRPの遊離に対するカプサイシン(C)、ブラジ
  キニン(B)及びニコチン(N)の作用を検討した。CのCGRP遊離作用は知覚
  神経への直接作用であり、B及びNの刺激作用はプロスタグランジン合成を介すると
  考えられた。Nの作用には交感神経活性化も関与していることが示唆された
CC EB08100U, GW05040V, GN01020W (577.112.6.02, 615.217.4, 616.8-09)
KW 神経ペプチド; カプサイシン; ブラジキニン; ニコチン; 知覚神経; プロスタグラ
  ンジン; 交感神経; 気管; ラット; 灌流; in vitro実験; 動物; 肥満細胞;
  Compound 48-80; インドメタシン
FT [CGRP]
SW カプサイシン(J1.529F); ブラジキニン(J10.684D); ニコチン(J9.266E); B-ヒド
  ロキシドーパミン(J126K); 化合物48/80(J34.548B); インドメタシン(J1.368D
  ); (-)-ノルエピネフリン(J9.223A); ATP(J10.680A); プロスタグランジン
  E1(J3.319G); プロスタグランジンE2(J9.243F); プロスタグランジンF2α(J9
  .246K); プロスタサイクリン(J17.550A); ホスホラミドン(J18.388A)
```

図3 物質索引された文献データベースの例

か蓄積していないので、その以前の文献はフリータームとして索引されていることを期待し、こうして得た同一物質についての多様な名称を、今度はSW:を使わず、直接入力して検索する。

[例] 知っている薬物名: テブフェノジド

U: テブフェノジド

[1] S: 1 テブフェノジド
U: ¥P A (抄録付きの出力を指定)
回答出力 (中略)
SW テブフェノジド (J540.289A)
(SN が分かる)

U: SN: J540.289A

[2] S: 2 J540.289A
(物質索引は2件)
U: 2#1 (No 1の重複を除去)

[3] S: 1 2#1
U: ¥P A
回答出力 (中略)

SW RH-5992 (J540.289A)

(同一物質の治験番号が分かり再度これで検索する)

U: RH-5992

[4] S: 5 RH-5992
U: 4#3

[5] S: 4 4#3
(従来からのフリーターム索引の分が追加される)

U: ¥P A

回答出力 (中略)

FT [RH-5992]

6. 2. 2 JOIS-F から JOIS への渡り検索

化合物からの本格的な文献検索は JOIS-F によって名称、分子式、構造図およびこれらの組み合わせによって目的化合物の集合を取得し、JOIS の文献データベースに接続してその SN 番号で今度は文献の集合を得ることによって行う。渡り検索というこの方式は既に述べたように、JOIS と JOIS-F のホスト計算機が別であるため、目下のところ同一の会話の中で化合物検索と文献検索を自由に結び付けて検索することはできない。化

合物の集合が小さい場合は、JOIS-F で検索した結果をプリントアウトし、回答中の SN を今度は JOIS において順次以下のように+で結んで入力すれば、それらの SN が付与された文献を検索できる。

[例]

U: S N: J253.795H+SN: J345.194A+
SN: J3.467C+J22.2 53D+
J245.477G

[1] S: 48件

しかし化合物の集合が大きい場合はこの方法は適当でない。同一の会話の中で SN を自動的に転送する機能を JOIS 側に付加できないため、次善策として端末側で対応する方式を開発してきた。すなわち JOIS-F の検索結果をセーブし、オートログオンで JOIS に接続してから蓄積した SN 番号で自動的に文献を検索するもので、パーソナル・ビジネス・アシスト社との共同開発で通信エミュレーター、PCOM- α に実装したものである。このエミュレーターの場合は化合物辞書検索での1回の会話で最大5テーマまで、1テーマ最大999化合物までセーブできる。これらの数字は決して十分なものではないが、大部分の化合物検索の結果はこの制限内に納まると思われるので、その限りにおいての活用が期待できる。PCOM- α による渡り検索は、化合物検索の直前の結果に対し、

¥WSAVE (渡りセーブ)

のコマンドによってこれをセーブすることから始める。このコマンドの後、1スペース置いて英文字Aを指定し、テーマを区別する。システムからのメッセージに応じてセーブした集合に仮の名を付けると、エミュレーターのプログラムは JOIS-F から次々と SN 番号を受信し、蓄積する。別の化合物検索を行い、その集合を同様に

¥WSAVE B

と今度は別の文字で区別して蓄積する。テーマは A~E と最大5テーマまで区別できる。次に

¥WSAVEEND

により PCOM- α による端末への検索結果のセーブが終了する。次に JOIS へ接続を変え、

¥FILE 010

で JICST ファイルを選んでから、

¥CROSS A (または B~E)

でセーブした各々の SN 集合を呼出す。システムは

U :

のプロンプトの後に自動的に

U : SN :

として番号を列挙し、+で結んで質問式として検索を実行する。1回の質問式が256文字を越える場合は自動的に SN の境目で区切り、いくつかに分割して質問式の入力が続ける。全 SN の入力終了すると各々の集合の和を自動的に作る。この後は得られた化合物の集合は通常のキーワードと任意の順番で組み合わせて AND 演算し、求める文献に絞り込むことができる。

なお、PCOM- α の渡り検索機能はオプションであるが、これを組み込んだバージョンは JICST で預かった分に余裕があるので、お問い合わせ頂きたい。

6. 2. 3 化合物辞書の同義語を表示する機能の活用

化合物辞書の他の重要な役割に、同一化合物の様々な名称を示す機能がある。文献の物質索引はまだ始まって間がなく蓄積量が少ないので、SN 番号のみをキーとしたのでは、その前の古い部分が漏れる。前述のように文献抄録の回答を見て索引された同義語を知ることができるが、医薬品などで同義語が多い場合は一旦辞書を検索し、使われている可能性のある同義語を全て出力してみるのが確実である。文献ファイルで一般のキーワードとして索引されていることを期待してこれらを検索語として列挙すると、過去に遡った網羅的な検索ができる。

[例] ニルバジピン

JOIS-F にて検索した回答

SN J22.253D

S Y ニルバジピン

ニバジピン

ニバジル

ニバジール

FK-235

FR-34235

JOIS にて検索

[1]	SN : J22.253D	143件
[2]	ニルバジピン	235件
[3]	ニバジピン	2件
[4]	ニバジル	3件
[5]	ニバジール	27件
[6]	FK-235	4件
[7]	FR-34235	3件
[8]	¥OR 2-7	241件
[9]	8#1	98件

この最後の98件が物質索引以前のキーワード索引の結果で、この方法により救済された分である。

7. まとめと今後の課題

JICST の化合物辞書システムおよび物質索引は、欧米に20年以上も遅れてスタートし、今なお不完全であるが、遅くとも1996年度からは全面的に実施すべく準備を進めている。また同一ホスト内での渡り検索についても、現在開発を進めている次期システム(JOIS-IV)では可能となる予定であるが、それ以前に JOIS-III の世代においても1996(平成8)年度には化合物辞書のうち名称ファイルを搭載し、分子式や構造の特徴を表わすキーを含む名称系の検索を行った後、¥MAP 機能により一会話内でのマルチデータベース検索ができるよう検討中である。このように JICST ファイルの化合物検索については着実にその質的向上に向かって歩み出したところである。今後は皆様方のご期待に沿った物質索引の実現に向けて努力して行かなくてはならない訳である。前述したように医学・薬学分野では1992年度から全面的に化合物辞書方式による索引を行っているの、是非ご利用いただき、これについての貴重なご意見等をお寄せくださると有難く存じます。

先進諸国の化学情報システムが量的・質的に大きく発展し国際化が進む中で、JICST で独自に物

質索引を行う意義が改めて問われる。その答えは先に述べた三つの特徴のほかにもあると思われる。すなわち、東西の軍事対立が終了した後、世界的な経済競争が国家や企業間の主要な問題になると言われている。このような中で的確な情報をいかに独自に入手するかがもっとも重要となる。1億3千万の人口があり、これだけ科学や工業の

発達した日本が独自のデータベースとその提供手段を持つことは、さらなる発展のためにも必須の要件であろう。水、安全、情報はタダで容易に入手できるという時代は過ぎつつある。ただそれはJICSTの力に余る事業であるから、日本の化学界の長期的な支援を是非ともお願いする次第である。

参 考 文 献

- 1) Kresze, Guenter. *Angewandte Chemie* 1970 82, 15, 563-568
- 2) Ash J.E.; Hyde, E. *Chemical Information System* J. Wiley & Sons 1974
- 3) 岡野 禎二 *情報管理* 1974 17, 3, 176-183
- 4) 斉藤和男, 岡野 禎二 *情報管理* 1965 8, 8, 36-39
- 5) 荒木 啓介 他 *情報管理* 1975 18, 7, 571-577
- 6) 内野 浩, 荒木 啓介 第14回情報科学技術研究集会発表論文集, 1977; pp101-121
- 7) Vander Stouw, G. G.; Naznitsky, I.; Rush, J. E. *J. Chem. Doc.* 1974, 14, 185-193
- 8) 藤原 譲 et al 第21回情報科学技術研究集会発表論文集 1984; pp49-83
- 9) Araki, K.; Kaji, M. *J. Chem. Inf. Comput. Sci.* 1991, 31, 363-374

(事務局注)

今回の解説は平成7年7月末現在で実行可能な渡り検索の方法を解説したものである。JICSTでは、渡り検索について、平成8年度から一部、平成10年度から本格的な運用を行えるようシステムの改造等開発を進めている。