Introduction à l'information chimique informatisée

par H. J.-M. Dou et P. Hassanaly

(Laboratoire de chimie organique A, Centre Saint-Jérôme, 13397 Marseille Cedex 4)

H. J.-M. Dou

P. Hassanaly

Depuis environ sept ans, nous voyons autour de nous une multitude de nouveaux outils informatiques et, parmi eux, les terminaux, imprimantes et modem acoustiques ou magnétiques, vont nous permettre d'accéder rapidement, par couplage entre un téléphone et un ordinateur, à de nombreuses sources documentaires, d'une manière instantanée.
Mais, pour bénéficier pleinement des avantages de ces systèmes, il est d'abord nécessaire de bien comprendre comment ils sont conçus, afin de déterminer leurs possibilités, mais aussi leurs limites.

Dans cette introduction, nous présenterons d'abord :
1. comment sont nés les systèmes informatisés,
2. comment sont créées les bases de données,
3. comment elles sont interrogées,
4. comment les ordinateurs sont reliés aux utilisateurs (les réseaux).

Dans la suite de l'exposé, nous présenterons plus particulièrement les systèmes utilisables en chimie, et qui sont à la disposition du chercheur. Parmi toutes les bases disponibles, il est bien évident que *Chemical Abstracts,* et les produits informatisés qui en découlent, occuperont une place de choix. Les brevets, et les interrogations par structures et sous-structures, tels les dictionnaires chimiques dérivés de *Chemical Abstracts* (ou leur traitement pour les rendre interrogeables selon le logiciel DARC), ne seront pas oubliés. Enfin, le seul fichier de réaction disponible sur une large base sera présenté : le CRDS de Derwent. Un certain nombre de fichiers annexes existent ; sans les passer en revue dans le détail, nous les présenterons brièvement, une place particulière étant dévolue au fichier Scisearch issu de *Current Contents* (fichier bibliographique et de citations).

I. Introduction

1. La naissance des systèmes informatisés

Vers les années 1960, les progrès de l'impression des produits papiers ont été accélérés par l'introduction de l'informatique et de la photocomposition. Un texte, saisi sur bande magnétique (ou disque), peut être ensuite reproduit, après avoir été corrigé ou recomposé (les traitements de textes). Par exemple, l'impression actuelle de *Chemical Abstracts* est réalisée par ordinateur. Les numéros des résumés suivis d'un chiffre destiné à la reconnaissance informatique, les numéros de registre (registry number, RN) des composés chimiques, les « parents compounds identifiers » tels que KMTSB [pour identifier le cycle cinnolino (2,1-*a*) ($C_{16}H_{12}N_2$)]... sont directement dérivés du traitement informatique des données saisies. Ainsi, tous les grands systèmes mondiaux de « produits papiers » présentant des références par sujets, auteurs, avec ou sans résumés (abstracts), dérivent *tous* d'une saisie informatique de données : les publications thèses, brevets, conférences et autres documents sont pris en compte suivant une politique préalablement définie. Ainsi, *Chemical Abstracts,* les produits papiers de l'Institute for Scientific Information, *Bioabstracts, Pollution Abstracts, Oceanic Abstracts, Excerpta Medica,*

Physical Abstracts, Derwent, Pascal... représentent une somme considérable de données saisies sur support magnétique. Parallèlement, et à cause des contraintes issues de la « Guerre froide », les Américains développaient, sur leur territoire, des réseaux de télécommunications fiables, permettant d'acheminer et de recevoir des informations vitales depuis les centrales de mesures jusqu'aux ordinateurs de traitement et aux centres de décision.

Ainsi, deux outils différents au départ étaient, vers la même période, disponibles. Le phénomène de fertilisation croisée avait bien lieu, et mariait ainsi les premiers services de distribution de l'information scientifique en ligne (online), tout d'abord à des fins militaires, puis ensuite à des fins civiles. C'est donc dans le berceau des applications militaires que ces systèmes se sont développés. Ils ne sont plus aujourd'hui situés dans un tel environnement et sont gérés par des entreprises commerciales, mais ils sont, de par *leur nature même,* des outils de décision qui, malheureusement, sont encore peu utilisés en France (et même en Europe, dans ce sens).

2. Comment sont créées les bases de données

Il ne faudrait pas croire que le fait d'avoir saisi, sur un support magnétique, une série de références constitue une base de donnée. En effet, sous cette forme, le produit n'est pas interrogeable en ligne. Il faut, pour cela, le modifier. Cette transformation qui sera effectuée par un ordinateur, à partir de programmes pré-établis, consiste à partir d'un ensemble de données (la référence) et à rendre celui-ci interrogeable par de multiples entrées : titre, auteurs, adresse, source, date, langue... Tous ces éléments (termes), communément appelés des champs *, vont alors constituer autant de listes (par ordre alphabétique) de termes, qui sont interrogeables séparément par ordinateur. Certains auteurs appellent cette transformation l'inversion du fichier de base.

Prenons un exemple; supposons une référence issue de *Chemical Abstracts* :

```
" ref 16
" AN - CA91 - 145818 (18)
" TI - DETERMINATION OF FREE AND BOUND FATTY ACIDS IN
"      ROVER WATER BY HIGH PERFORMANCE LIQUID
"      CHROMATOGRAPHY
" AU - HULLET, D.A.; EISENRICH, S. J.
" OS - UNIV. MINNESOTA, DEP. CIV. MINER. ENG. MINNEAPOLIS
" SO - ANAL. CHEM. (ANCHAM), V 51 (12), p. 1953-60, 1979
"      ISSN 00032700
" LA - ENG
" CC - SEC61-2
```

• Cette référence est décomposée en divers champs : titre (TI), auteurs (AU), organisation source (OS), source (SO), langage (LA), section *Chemical Abstracts* (CC).
• Le champ « titre » est constitué par les mots :

```
" TI - ACIDS, BOUND, CHROMATOGRAPHY, DETERMINATION, FREE,
       FATTY, HIGH, LIQUID, PERFORMANCE, ROVER, WATER
```

• le champ « auteur » par :

```
" AU - EISENRICH S. J., HULLET D. A.
```

et ainsi de suite.
Il est à noter que tous les termes sont placés par ordre alphabétique et que tous les champs ne sont pas forcément interrogeables en ligne. Par exemple : un journal n'est pas caractérisé par son nom, mais par son CODEN.

Ainsi, toutes les références contenues dans le fichier général seront décomposées en champs interrogeables, et les contenus de ces champs seront placés, tous ensemble, dans les sous-fichiers correspondants. De plus, une relation existe entre les termes contenus dans les sous-fichiers, le nombre de références bibliographiques qui les contiennent, ainsi que le fichier central qui renferme toutes les références et qui servira à l'impression des résultats.

* *Un champ peut comporter plusieurs termes.*

Ainsi, au rythme de 500 000 références par an, de plus de 350 000 composés nouveaux, on peut facilement considérer la taille gigantesque des bases disponibles.

3. Comment sont interrogées les bases de données

En étant très succincts, nous dirons seulement que les bases de données sont interrogeables en réalisant des ensembles discrets combinables entre eux.
Revenons à l'exemple précédent,
• à la question 1 : SS1 = SEARCH STATEMENT 1 (équation de recherche n° 1)
« SS1 — CHROMATOGRAPHY
• l'ordinateur répondra par exemple :
« PSTG — 6551
PSTG (posting), soit 6551 références présentes dans le fichier et contenant le mot CHROMATOGRAPHY.

On peut alors constituer d'autres ensembles :

```
SS2    HIGH
PSTG   23582
SS3    PERFORMANCE
PSTG   20612
```

...etc, ceux pour tous les mots du titre par exemple.

On peut alors combiner les ensembles ainsi formés de trois façons différentes :

```
- opérateur utilisé AND      (ET)      A   et   B
-      "       "   OR       (OU)      A   ou   B
-      "       "   AND NOT (ET PAS) A et pas B
```

seules, les parties hachurées seront prises en compte (figure 1).

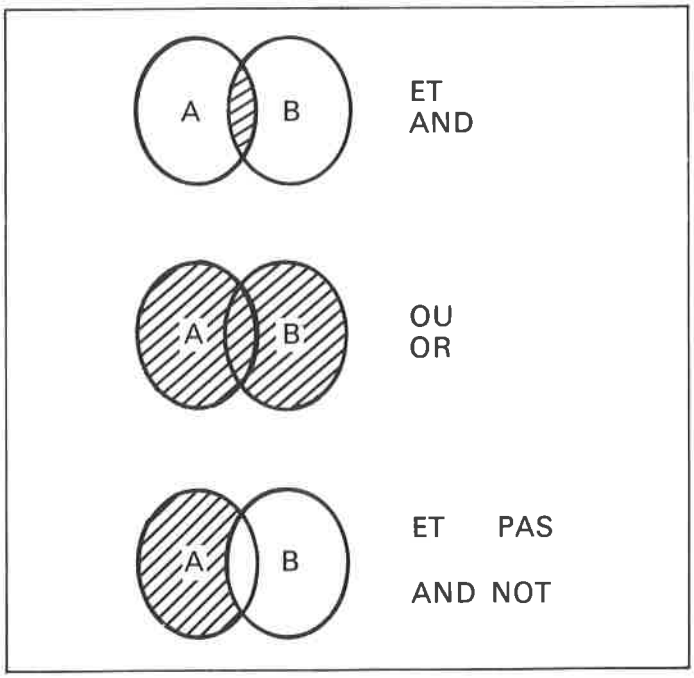


Figure 1.

Ainsi, dans notre exemple précédent, la combinaison de tous les termes :

```
DETERMINATION et FREE et BOUND et FATTY et ACIDS, etc.
```

amènera l'ordinateur à retrouver la référence choisie, pourvu que celle-ci reste dans le fichier. Le temps de recherche sur plusieurs millions de références, ne demandera que quelques secondes.

Il est bien évident que la présentation de la méthode d'interrogation est ici très sommaire, car les ensembles obtenus à

partir de divers champs peuvent être combinés entre eux, par exemple la (ou les) section du CA, le journal (sous forme de Coden), les auteurs,...

On peut ainsi, en choisissant les termes convenables, réaliser un *profil* qui conduira à un ensemble de références qui sera ensuite imprimé, soit au terminal, s'il y a peu de références en ligne (online), soit sur les sites de l'ordinateur en utilisant une impression en différé (dite offline). Le même profil pourra être ensuite stocké dans l'ordinateur central, rappelé si besoin est, etc. De même, s'il ne comprend pas de termes appartenant au thésaurus [donc en langage libre : généralement le titre mots clés (keywords) de CAS], ce profil pourra être sauvegardé et exécuté sur d'autres fichiers.

4. Comment les ordinateurs sont reliés aux utilisateurs

Les ordinateurs, ou serveurs, ou hosts (ce sont les U.S.A. qui sont très largement en tête dans le monde), sont reliés aux utilisateurs par une série de réseaux de télécommunications qui travaillent en temps partagé (comme c'est le cas aussi pour les serveurs). Ainsi, on partira d'un réseau national : le réseau Transpac, qui reliera par un simple numéro les abonnés Transpac à un concentrateur situé à Paris. A partir de ce dernier, on aura accès au réseau européen Euronet (serveurs européens : Agence spatiale européenne, Télésystèmes-Questel, Dimdi, Data Star,...) ou au réseau international Tymnet (grands serveurs internationaux : Lockeed Information System, System Development Corporation, Bibliographic Retrieval System,...).

Ainsi les opérations à réaliser, pour être connectées à un serveur, sont extrêmement simples : à partir d'un téléphone et d'un terminal portable avec imprimante d'environ 3 kg, il suffit de composer un numéro de téléphone pour être relié au concentrateur de Paris. On communique alors à celui-ci, par l'intermédiaire du clavier de la console, des numéros personnels permettant l'identification du correspondant et la facturation des transmissions téléphoniques (à un prix beaucoup plus bas que celui des communications ordinaires) et comportant aussi l'indication du serveur auquel on veut être relié. Une fois relié à ce dernier, il faudra à nouveau se faire connaître de celui-ci par une série de mots de passe, avant d'être relié à la base de son choix et d'effectuer la recherche. Cette opération de connection (dite de Login), demande généralement 1/100e d'heure.

II. Les systèmes utilisables en chimie (CAS)

Nous ne pouvons pas, dans un exposé aussi bref, faire une analyse complète de ce problème, nous ne ferons simplement qu'aborder, avec des exemples issus de fichiers différents, les divers aspects qui peuvent être traités (1).

Quelles sont les références publiées depuis 1977 sur l'analyse des acides gras en chromatographie liquide sous pression ?

La stratégie très simple consiste à demander :

```
" SSI   FATTY AND ACIDS  AND LIQUID AND HIGH AND (PERFORMANCE
        OR PRESSURE)
" PSTG  54 références
```

La bibliographie a duré 3 minutes, les résultats imprimés en différé ont été reçus cinq jours après.

Citons, en plus de la référence prise par exemple en début de texte, les références suivantes :

```
" ref 29 : SEPARATION OF LONG AND SHORT CHEM FATTY ACIDS
"          AS NAPHTACYL AND SUBSTITUTED PHENACYL ESTERS
"          BY HIGH PERFORMANCE LIQUID CHROMATOGRAPHY
"          JORDI HOWARD C.
"          WATER ASSOC. INC. LIFE SCI. DIV. MILFORD MASS.

" ref 32 : ANALYSIS OF NONIONIC SURFACTANTS (EMULSIFIERS)
"          USING HIGH PRESSURE COLUMN CHROMATOGRAPHY
"          BRUESCHWEILER H.
"          EIDG. MATERIALPRUEF-VERSUCHSANST ST.GALLEN SWITZ
"          MITT. GEB. LEBENSMITTELUNTERS HYG. (MGLHAE) V.68
"          (1) p.46-63 1977
```

Remarquez que les références les plus récentes sont imprimées en tête, ce qui permet d'effectuer par exemple le même travail un an après avec, par exemple, 72 références, au total, et de n'imprimer que les 18 plus récentes, et ainsi d'actualiser la bibliographie initiale.

Est-il possible d'obtenir un résumé des CA en ligne.

Actuellement, cela n'est pas possible, car la taille du fichier ne le permet pas. Notons cependant que le fichier CBAC (480 000 références) qui est constitué par la partie biologique de *Chemical Abstracts* (accessible par Télésystèmes), comporte un résumé. De même, depuis 1974, un important fichier : Biosis (Bioabstracts) renferme aussi le résumé accessible en ligne. En ce qui concerne *Chemical Abstracts*, l'utilisation de l'impression complète de la référence permettra d'obtenir de précieuses informations, en utilisant les termes d'indexation (aussi interrogeables en ligne) ajoutés à la référence bibliographique, ou, dans certains cas comme CBAC, le résumé * :

```
" -1- 337742 C.CNIC.ACS
" AN : CA87-194698(25)
" CC : S3-2
" TI : - INHIBITION OF CLOSTRIDIUM BOTULINUM BY 5-NITROTHIAZOLES
" AU : DYMICKY,M. HUHTANEN,G. N. WASSERMAN,A. E.
" AF : - ARS
"      - ERRC
" LO : - PHILADELPHIA
"      - PA.
" DT : J
" SO : - ANTIMICROB. AGENTS CHEMOTHER. (AMACCQ) V12(3) P.353-6
" DP : 77
" LA : ENG
" AB : - A NO. OF 5-NITROTHIAZOLES WITH VARIOUS SUBSTITUENTS
"        IN THE 2-POSITION WERE TESTED FOR INHIBITION OF C.
"        BOTULINUM IN A CULTURE MEDIUM.
"      - THIAZOLE ITSELF OR 2-BROMO- OR 2-METHYLTHIAZOLE AT
"        30 .MU.G/ML DID NOT INHIBIT THE ORGANISME.
"      - AN AMINO GROUP IN THE 5-POSITION OF 2-AMINOTHIAZOLE
"        INCREASED THE INHIBITORY LEVEL TO 0.12. .MU.G/ML
"        ACETYL-, PROPIONYL, OR BUTYROYL-2-AMINO-5-NITROTHIAZOLE
"        INHIBITED AT 0.04 .MU.G/ML.
"      - BENZOYL-2-AMINO-5-NITROTHIAZOLE INHIBITED AT 0.16
"        .MU.G/ML THIS INCREASED TO 0.01 .MU.G/ML WHEN THE
"        BENZOYL GROUP CARRIED A NITRO GROUP IN THE M- OR
"        P-POSITION  A NITRO GROUP IN THE O-POSITION, ON THE
"        OTHER HAND, INHIBITED AT 0.04 .MU.G/ML.
"      - UNSATD. ALIPH. ACYLS DECREASED INHIBITION.
"      - THE GREATEST ACTIVITY WAS EXHIBITED BY 2-NONANOYL-
"        AND 2-LAUROYLAMIDES, WITH MIN. INHIBITORY CONCNS.
"        OF 0.005 AND 0.0025 .MU.G/ML, RESP.
" IT : - 96-50-4 288-47-1, DERIVS. 288-47-1 1603-91-4
"        1606-76-4 3034-22-8 3034-48-8 3034-53-5 3581-87-1
"        7305-71-7 16243-71-3 19783-57-4 41663-73-4 64724-76-1
"        64724-90-9 64724-91-0 64724-92-1 64724-93-2
"        (CLOSTRIDIUM BOTULINUM INHIBITION BY)
"      - 121-66-4 (ACYLATION AND BACTERICIDAL ACTIVITY OF)
"      - 140-40-9 14538-16-0 14645-50-2 64398-84-1 64724-77-2
"        64724-78-3 64724-79-4 64724-80-7 64724-81-8 64724-82-9
"        64724-83-0 64724-84-1 64724-85-2 64724-86-3 64724-87-4
"        64724-88-5 64724-89-6 (PREPN. AND BACTERICIDAL
"        ACTIVITY OF)
```

* *Recherche réalisée par l'intermédiaire du serveur Télésystèmes.*

```
"       - BACTERICIDES, DISINFECTANTS AND ANTISEPTICS
"         (NITROTHIAZOLES AS)
"       - CLOSTRIDIUM BOTULINUM (NITROTHIAZOLE INHIBITION OF)
"       - MOLECULAR STRUCTURE-BIOLOGICAL ACTIVITY RELATIONSHIP,
"         BACTERICIDAL (OF NITROTHIAZOLES)
```

• Issu du fichier Biosis * :

```
" -1-
" AN - BA/BIOI CIT. NO.   BA70-044593
" TITLE                   THE EARLY STAGES AND BIOLOGY OF
"                         ACERBIA-ALPINA LEPIDOPTERA ARCTIIDAE
" AUTHOR                  SOTAVALTA O; KARVONEN E; KARVONEN E;
"                         KORPELA S; KORPELA J.
" ORGANIZATIONAL SOURCE   PUDASRINNE 4B, SF-01600 VANTAA 60,
"                         FINL.
" SOURCE                  NOT ENTOMOL (NOENA), 60(2). 1980.,
"                         P. 89-95.
" LANGUAGE                EN
" WEIGHTED CATEGORY CODES  63584 (INVERTEBRATE TAXONOMY-INSECTA,
"                         LEPIDOPTERA);
"                          07508 (ECOLOGY-ANIMAL); 10614
"                         (EXTERNAL EFFECTS-TEMPERATURE); 16501
"                         (REPRODUCTIVE SYSTEM-GENERAL STUDIES
"                         METHODS); 17020 (ENDOCRINE SYSTEM-
"                         NEUROENDOCRINOLOGY); 25508 (DEVE-
"                         LOPMENTAL BIOLOGY-GENERAL MORPHOGENE-
"                         SIS); 62800 (ANIMAL DISTRIBUTION/
"                         ZOOGEOGRAPHY); 64076
" TAXONOMIC CODES         75330 (LEPIDOPTERA)
" INDEX TERMS             F-2 GENERATION DIAPAUSE POPULATION
"                         SURVIVAL ASIA FINLAND
" ABSTRACT                THIS INTERNATIONAL RARITY OCCURS IN
"                         THE ARCTIC IN THE NEW AND OLD WORLD,
"                         AND IN A FEW DISJUNCT ALPINE AREAS
"                         IN ASIA, INHABITING TREELESS TUNDRA
"                         AND ROCKY AND SCREE BIOTOPES. IN
"                         1977 2 SETS OF CATERPILLARS WERE
"                         REARED EX OVO IN FINLAND AND IN 1978
"                         AN F2 GENERATION WAS OBTAINED. THE
"                         EGGS ARE YELLOW. THE CATERPILLARS
"                         HAVE A TRANSVERSELY STRIPED BLACK/
"                         YELLOW HAIR COVERING AND ARE
"                         POLYPHAGOUS; THEY HAVE A PREPUPAL
"                         DIAPAUSE AT THE 7TH INSTAR, BUT MOST
"                         OF THEM HAVE ANOTHER DIAPAUSE AT THE
"                         6TH OR 5TH INSTAR AND THEY MAY HAVE
"                         ADDITIONAL DIAPAUSES AT EARLIER
"                         STAGES. WHEN REARED AT HIGH TEMPERA-
"                         TURES THEY MAY SKIP 1 INSTAR AND
"                         PUPATE WITHOUT A DIAPAUSE. THIS
"                         PLASTICITY IN THE LIFE CYCLE ENABLES
"                         THEM TO ADAPT TO THE PRONOUNCED
"                         SHORT- AND LONG-TERM VARIATION IN
"                         THE CLIMATIC CONDITIONS OF THE ARCTIC
"                         AND PERMITS SURVIVAL OF THE POPULA-
"                         TIONS THROUGH UNFAVORABLE PERIODS.
```

Quelles sont les publications réalisées en commun par les auteurs de cet article depuis 1977 à nos jours ? (sans publicité et pour éviter tout commentaire).

Question :
DOU, H:/AU et HASSANALY, P:/AU

Réponse :
13 publications qui peuvent être imprimées.
Temps de réponse :
42 secondes !

Nota : L'utilisation du fichier Scisearch de ISI permettrait de savoir qui cite ces travaux et où.

* *Recherche réalisée par l'intermédiaire du serveur S.D.C.*

Quel est le volume des travaux publiés depuis 1977 sur le charbon dans le monde ? (On utilise le fichier Chemical Abstracts de 1977 à nos jours).

Question : COAL posée en langage libre sur le « basic index ».
Réponse : 13462 références, brevets compris.
Répartition :

Belgique	France	R.F.A.	U.K.	U.S.A.
33	66	575	276	1 455

Quel est le volume des travaux publiés en français dans le Bulletin de la Société Chimique de France, le Nouveau Journal de Chimie, Tetrahedron, depuis 1977 ?

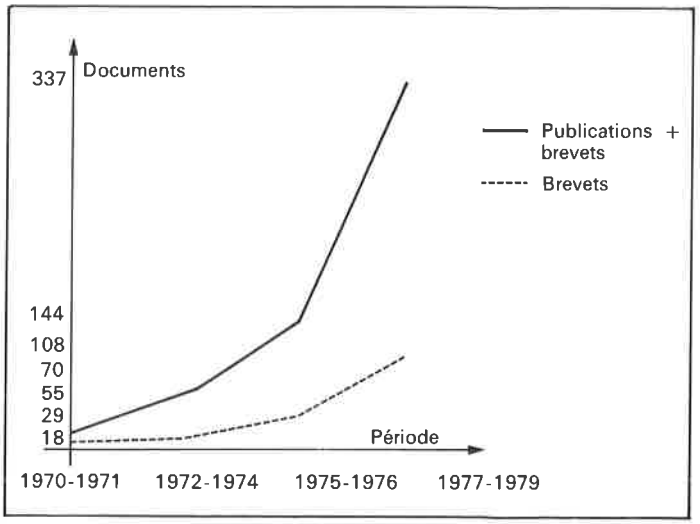
	Bull. Soc. Chim.	*Nouveau J. Chim.*	*Tetrahedron*
Total	855	375	1 781
En français	827	101	294

Quel est le volume des travaux publiés en français, traitant des hétérocycles à un seul noyau, depuis 1977 ?

Question : on limite la section 27 de *Chemical Abstracts* aux travaux publiés en français :
Totalité : 20 370.
En français : 595.

Quelle est l'évolution prévisible de la catalyse par transfert de phase durant les prochaines années ?
(Extrait de la thèse de doctorat ès sciences de Mme P. Hassanaly).

On choisit un certain nombre de descripteurs, et on limite par année. On aboutit à la courbe suivante :



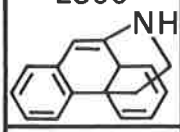
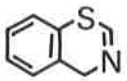
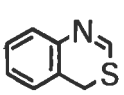
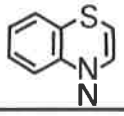
Nota : le ratio brevet-publication met en évidence la situation privilégiée de ce domaine au niveau des sciences de transfert.

*

Nous allons aborder maintenant un autre domaine, celui de la recherche par structure et sous-structure. En effet, pour tout chimiste, la pensée créatrice passe normalement par l'utilisation de structures chimiques. C'est à ce niveau que les analogies joueront pleinement.

Trois grands systèmes sont utilisables :
a. Le codage, utilisé dans le fichier Derwent *, dont nous donnons ici un exemple :

code ↓

E300 	E310 1 — N* 4 (other)	E320 Ergoline	E330 ⩾ 2 — N* 4 (other)	E340 Yohimbane	E350 ⩾ 1 — N ⩾ 5 (other)	E399 Poly	
E630 	E640 	E650 	E660 Benzothiaz-epine * (-ocine)	E670 	E680 	E690 2-(S + N)* 2 (other)	E699 Poly

Il faut noter que l'accès à la recherche par codes (punch codes) nécessite, dans ce cas, un abonnement préalable situé hors des possibilités financières de tout laboratoire universitaire. Nous ne le développerons donc pas en détail.

b. L'utilisation simultanée des codes issus du Ring Index (RSD, RPR, RNO), des formules moléculaires MF et des fragments de formules moléculaires (MFF), ainsi que des noms du « parent » (P), « parent fragment » (PF), nom et fragment de nom (NF),...

Ceci conduit à une recherche assez large, très porteuse d'idées et d'innovation.

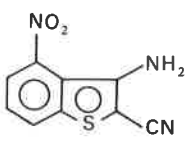
Question : *Ayant au départ une molécule hétérocyclique formée de deux noyaux accolés contenant de l'azote, du soufre et substituée par des groupements amino et nitro, quelles sont les structures voisines possibles déjà étudiées ?*
On entend par recherche de structures voisines un balayage (« browsing ») très large du fichier permettant de « récupérer le maximum d'idées ». La recherche est effectuée sur le fichier Chemdex du serveur S.D.C. (au total 1 600 000 structures).
Formulation : avec l'opérateur *et* (and) :

2/RNO	limite à deux cycles adjacents;
nitro/NF amino/NF	impose nitro et amino;
N_3/MFF O_2/MFF S_1/MFF	impose 3 azotes et 2 oxygènes et 1 soufre dans la formule moléculaire;

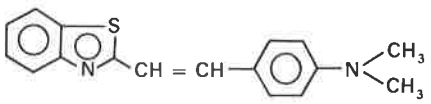
Opérateur *et pas* (and not) :

Cl/MFF Br/MFF	on ne veut pas de chlore et de brome sur la molécule;

Réponses : 21, avec par exemple :

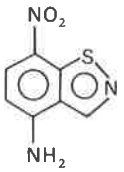


52673-87-7/RN

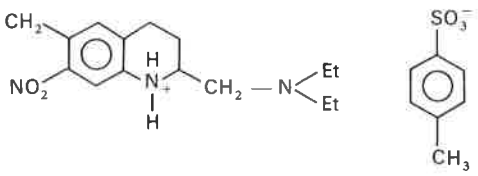


50963-25-2/RN

* *Le fichier des brevets Derwent est une exclusivité du Serveur S.D.C.*



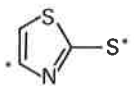
34976-49-3/RN



22989-50-0-RN

c. L'utilisation du codage structural qui, jusqu'alors, n'a reçu qu'une seule application commerciale à grande échelle, le système DARC (dérivé des dictionnaires chimiques CAS).
Actuellement, le fichier CBAC et la totalité du fichier CAS sont interrogeables en DARC *. La recherche sera plus précise que la précédente et on obtiendra soit le RN du composé, soit le dessin de la molécule. Un passage automatique des RN sélectionnés vers le fichier textuel CBAC ou CAS permet de retrouver les références bibliographiques.

Exemple * : déterminer les molécules étudiées dans le fichier CBAC (380 000 structures) et présentant le motif (question) :

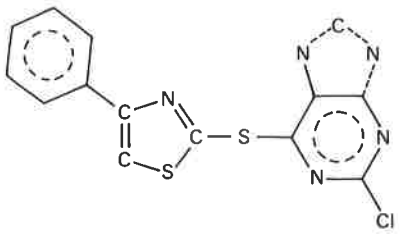


* *position libre pour toute substitution.*

La molécule est décrite, soit à l'aide d'un dessin en utilisant un « crayon électronique », soit en mode textuel, en numérotant les atomes dans n'importe quel ordre, et en décrivant les liaisons et leur nature, puis les atomes autres que le carbone et l'hydrogène enfin, en indiquant les positions substituables. La méthode de codage est très simple. La recherche effectuée est réalisée atome par atome, ce qui la rend plus précise que celle utilisée actuellement dans le système « CAS-on line », développé par CAS lui-même (en fait, CAS va bientôt introduire la recherche atome par atome).
Les structures sont éditées par l'ordinateur, en voici un exemple :

* *Recherche réalisée sur le serveur Télésystèmes.*

C.N.I.C.— — — Système DARC — — — C.N.I.C./C.B.A.C.
RÉPONSE : 18 COMPOSE NUMÉRO : 206168 RN : 038-449-400
1 FRAGMENT 22 ATOMES



FORM. MOL. : C14H8ClN5S2

On obtient les numéros de registre (registry numbers) qui peuvent automatiquement être transférés dans le fichier texte C.B.A.C., pour effectuer ensuite une recherche classique par mots-clé, auteurs,...

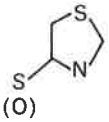
Une différence notable existe entre l'utilisation des dictionnaires chimiques et le fichier DARC. A notre avis, ils sont complémentaires et ne conduisent pas aux mêmes résultats. En effet, DARC restera très précis et ne permettra de retrouver qu'une seule série de structures à partir de la question posée. Les dictionnaires permettront d'autre part de retrouver ce qui est « autour » d'une question, et de repasser éventuellement au DARC pour étudier plus particulièrement une structure et ses substitutions, structure dont on n'aurait pas forcément eu l'idée au départ.

d. « C.A.S. online ». Ceci concerne un fichier structural développé par *Chemical Abstracts.*

La recherche est effectuée par structures et sous structures en utilisant des fragments *. L'ensemble des molécules sélectionnées est ensuite édité sur un écran. Les « registry numbers » sont aussi fournis. La recherche est différente de celle effectuée par le système DARC qui reste plus précis (travail atome par atome). Pour coder la molécule, il faut actuellement déterminer, avec un catalogue, les différents fragments possibles qui la décrivent, puis les introduire au clavier. (CAS indique que cette opération sera automatisée courant 1981).

Exemple partiel de codage : retrouver les molécules ayant comme motif structural le cycle ci-contre; on détermine les fragments :

S*C*N*C*C cycle 5
C*S* S*C*N* C*N*C○S...



(O)

on peut aussi exclure certains fragments :

S*C○S...

Entre autre, le nombre de fragments est limité, et actuellement le codage demande une certaine habitude; cependant, il est prévisible que ce système fera de très rapides progrès.

Il existe aussi des banques spécialisées, telles Pluridata, (RMN et spectrométrie de masse) qui donnent accès à la recherche de spectres par glissements chimiques, ou fragments. De même, Thermodata fournit des données thermodynamiques. Ceci permet d'introduire la notion de banque de données, par opposition aux bases de données. Disons succinctement qu'une base de données contient des signalements bibliographiques, tandis qu'une banque de données renferme des informations précises, chiffrées (données thermodynamiques...).

III. Autres bases de données utilisables en chimie

Elles sont très nombreuses. Nous avons déjà cité Biosis qui traite de la biologie et de la biochimie.
De même, les bases médicales telles que Medline, Excerpta Medica contiennent aussi de nombreuses informations. Les fichiers dérivés de la Société Derwent : Plasdoc, Vetdoc, Ringdoc, CRDS (fichiers de réactions), tiennent une place importante.

Enfin, il existe des fichiers plus spécialisés, tels que : RAPRA, polymères, Paperchem, pollution, agriculture, environnement, nourriture, normes, Pascal ** (pluridisciplinaire mais incomplet dans certains cas); avec une mention très spéciale pour trois fichiers particuliers : celui des contrats NTIS (National Technicals Information Service) des U.S.A., celui des thèses américaines, et celui du SSIE (Smithsonian Science Information Exchange Inc.) qui traite des contrats de recherche en cours aux USA financés par le gouvernement.

Sans entrer dans le détail, car tous ces fichiers ont leurs propres caractéristiques, nous citerons quelques exemples pour que le lecteur ait une certaine idée des résultats obtenus :

• AGRICOLA : serveur S.D.C. :

```
AN  - 789704255
TI  - WEEDS FOR WOOLS : / A REFERENCE CHECKLIST TO 81 BOTANICAL
      ENTITIES REPORTED TO HAVE BEEN USED IN THE NATURAL DYEING OF WOOL
      FIBERS / WEEDS? WILDLINGS, AND OTHER PLANTS, LOCAL TO
      METROPOLITAN WASHINGTON, D.C. - SUBURBAN AND RURAL AREAS ...
      NEAL BOZARTH. -
AU  - BOZARTH, NEAL
SO  - TAKOMA PARK, MD. : (S.N.),, 126 P. ; 28 CM.
PD  - 1977
NO  - INCLUDES INDEXES. ERRATA SLIP INSERTED.
DT  - MONOGRAPH
SN  - US PUB
LA  - ENG
CN  - TP899.B6
PCC - 850500
IT  - DYES AND DYEING
IT  - WOOL.
```

• Fichier PASCAL, serveur Télésystèmes
Question :
Thiazole *et* nitro *et* acétique : une référence,

```
1520518 C.PASCAL
NO : 80-6-0121190
ET : THE SYNTHESIS AND + +(13)C NMR-SPECTRA OF PYRROLOTHIAZOLES AND THEIR
     PRECURSORS. BROMINE-INDUCED CYCLIZATION OF PYRROLYLTHIOUREAS
AU : GREHN L.
AF : UNIV.UPPSALA, INST. CHEM.,UPPSALA 75121,SWE
DT : TP;LA
SO : CHEM. SCRIPTA; SWE; DA. 1978-1979; VOL. 13; NO 2-3; PP. 78-95; BIBL. 26
     REF.; LOC. CNRS-15135
LA : ENG
FA : SYNTHESE D'ACYL-3 THIOUREIDO-4 PYRROLECARBOXYLATES-2 PAR REACTION
     D'AMINO-4 PYRROLECARBOXYLATES-2 AVEC DES ISOTHIOCYANATES D'ACYLE;
     CYCLISATION DE CES COMPOSES EN PYRROLO (3,2-D)-, -(2,3-D)- ET -(3,4-D)
     THIAZOLES. SPECTRES RMN + + (1)H, RMN + + (13)C
CC : 170.E.07.G
FD : BROME:ACT; CYCLISATION; HETEROCYCLE AZOTE; CYCLE 5 CHAINONS; REACTION
     CATALYTIQUE; COMPOSE BICYCLIQUE; HETEROCYCLE SOUFRE AZOTE; ACETIQUE
     ACIDE;SUB; PYRROLO THIAZOLE(BENZAMIDO ETHOXYCARBONYL METHYL);FIN;
     PYRROLO THIAZOLE(ETHOXYCARBONYLAMINO METHYL NITRO);FIN; PYRROLO
     THIAZOLE(DIMETHYL ETHOXYCARBONYL ETHOXYCARBOXYLAMINO);FIN;
```

** Le terme « fragment » est plus approprié que le terme « écran » en français.*
*** Prend aussi en compte les contrats et thèses françaises (en partie).*

PYRROLECARBOXYLIQUE-2 ACIDE(THIOURE-IDO-4);FIN;ENT;
PYRROLECARBOXYLIQUE-2 ACIDE(BENZOYL-3P THIOUREIDO-4 METHYL-1) ESTER ETHYLE;FIN;ENT; PYRROLECARBOXYLIQUE-2 ACIDE(AMINO-4) ESTER ALKYL;ENT; BENZOIQUE ACIDE,ISOTHIOCYANATE;ENT; FORMIQUE ACIDE(ISOTHIOCYANATO) ESTER ETHYLE;ENT; PHOSPHATE(TRIMETHYL);SUB

• Excerpta Medica (exclusivité du serveur L.I.S.) :
Studies on bancroftian filariasis control with diethylcarbamazine. I. Frequency and nature of drug reactions.
Sundaram R.M.; Koteswara Rao N.; Krishna Rao Ch.; et al.
Reg. Filaria Train. Res. Cent., NICD, Rajahmundry.
J. COMMUN. DIS. (INDIA), 1974, 6/4 (290-300), Coden : JCDSB.
Languages : ENGLISH.
The weekly dosage schedules precipitated lesser drug reactions than the alternate day or daily dose schedules. In the shorter regimens substantial number of reactions persisted till the end of the fifth dose in the longer regimen (weekly) there was a steep fall in the reactions by the second dose itself. The quantum of mild reactions had no relevance to the microfilaria density but the severity of reactions was associated with higher density. Females showed higher percentage of reactions than males. Children below 5 yr of age who harbored low microfilaremia showed least reactions. Pretreatment with antihistamine drug quickened the decline of reactions. Weekly doses are most accepted by the microfilaria carriers than the alternate day or daily schedules. Drug acceptance in total population therapy was poor.

• Fichier de la Chambre de Commerce de Paris : (serveur GCAM) :

PR CCIP DE PARIS.
AN 1980.
IO ISIS-C.
IS 6-150.
CE FRANCE.
AU FERRIERE (C).
TI LES INDUSTRIES AGRO ALIMENTAIRES, PETROLE VERT DE LA FRANCE.
GR - CO - COOPERATION DISTRIBUTION CONSOMMATION NG, MARS 1980, PP 24-33.
MA AGRO ALIMENTAIRE.
RE ANALYSE DE L'EVOLUTION RECENTE DE LA SITUATION ACTUELLE DES INDUSTRIES AGRO ALIMENTAIRES. MOUVEMENT DE RESTRUCTURATION ET DE CONCENTRATION POUR ARRIVER A AUGMENTER LE SOLDE POSITIF DE NOTRE BALANCE COMMERCIALE.
NA PERIODIQUE.

Conclusion

En tenant compte de la brièveté de cet exposé, nous essaierons de présenter dans cette conclusion les points qui nous paraissent les plus importants :

• le coût.
Le travail en ligne ne coûte pas cher, contrairement à la pensée générale; le prix d'une bibliographie se situera, pour un chimiste, aux alentours de celui de 3 à 4 litres d'acétonitrile pour chromatographie ! Cependant, nous avons constaté une réticence marquée des chercheurs à utiliser ce moyen qui se place au même niveau que la RMN, l'IR, etc. Ceci provient sans doute de la méconnaissance complète du coût de l'information, ainsi que des grandes lois bibliométriques modernes et des facteurs d'innovation.

• les performances.
Elles sont toujours supérieures au travail fait à la main, à condition que les personnes qui effectuent le travail soient compétentes !

• les possibilités.
Elles sont innombrables et permettent d'aborder un axe de recherche avec le maximum de renseignements, et souvent, de situer son effort par rapport au contexte international. Les seuils d'investissements minimaux sont alors abordés, puisque la connaissance du nombre de personnes travaillant sur le sujet, des contrats (au moins U.S.), des thèses et des techniques mises en jeu, est accessible.

• le suivi de la littérature, par le biais des DSI (Diffusion Sélective de l'Information) est immédiat.

• Enfin, ne pas considérer ces outils comme de simples relais bibliographiques.
En fait, ce sont de véritables outils d'aide à la décision. Le Japon, en créant ses propres bases de données, adaptées à ses possibilités exportatrices, s'est doté d'une arme redoutable, fer de lance de son économie. Les U.S.A., en possédant les meilleures bases de données du monde, peuvent ainsi créer les meilleurs modèles économétriques (même de la France !) et « gérer » en partie l'avenir.

• Au niveau des moyens, l'enseignement de l'utilisation de l'information doit avoir une place privilégiée. En France, par exemple, enseigne-t-on comment utiliser *Chemical Abstracts* ? A partir d'informations pertinentes, essayer d'aboutir à des innovations ? L'information scientifique, technique et économique est un tout. Elle ne représente que la partie apparente d'un système où l'information doit conduire à la réalisation. En effet, rien ne sert de se doter des systèmes les plus sophistiqués si, à la base, on ne peut pas utiliser à des fins de production, les données qui nous seront fournies. De même, pourquoi se battre sur le problème philosophique de la liberté d'accès à l'information, ou sur les barrières possibles de cette dernière, si on ne sait pas (ou si on ne peut pas) utiliser celle-ci. L'enjeu se situe bien au niveau des choix *possibles cohérents*, que l'on peut déduire par ces techniques, en évitant une dispersion des hommes et des moyens vers des voies sans issue.

(1) Pour les personnes qui seraient intéressées par le problème de l'information chimique, nous conseillons :
• Cours de formation permanente de l'I.P.S.O.I.
• Cours de formation permanente des agents CNRS de la région Provence-Côte d'Azur.
• Cours dans les entreprises, et audit interne du Centre de Recherche Rétrospective de Marseille, Université Aix-Marseille III.
• Cours sur les CA, et leur interrogation « online ».
• Cours de formation du Centre National de l'Information Chimique, 26 rue Boyer, 75020 Paris. Tél. : 797.29.29.
• Pour plus ample information : Tél. : (91) 63.03.15.