

Les systèmes documentaires automatisés Cas particulier de la chimie

par Mme Jeannie Persoz
(Association Française de Documentation Automatique
en Chimie*)

La pénétration de la documentation automatisée dans les laboratoires est progressive et régulière même si elle est plus lente que prévu. Précisons tout de suite que son seul but est de sélectionner les documents intéressant le chercheur. C'est donc un « service » ne supprimant pas la lecture et la réflexion mais lui apportant sécurité et gain de temps.

Le principal fonds documentaire en chimie, Chemical Abstracts, fait l'objet en France d'une exploitation automatisée depuis maintenant quatre ans et le nombre des services est en augmentation constante (300 profils en 1972, 500 en 1973, 525 actuellement).

Avec la possibilité récente d'interroger en conversationnel cinq années de C.A. (et d'autres fonds spécialisés) et la généralisation de l'emploi des codes structuraux, les chercheurs et documentalistes vont disposer de nouveaux moyens très puissants. Ils vont donc devoir s'initier plus avant à ces techniques.

C'est précisément le but de cet article que de faire une mise au point claire à l'usage des chercheurs sur la documentation automatisée et ses applications.

A cette occasion, les chercheurs pourront également apprécier le travail important, théorique et pratique, requis pour la mise en forme d'un tel volume d'information. Le côté économique ne leur échappera pas. (Budget annuel de C.A.S. : 22 millions de dollars).

I. Introduction

Deux facteurs essentiels ont concouru au développement de la documentation automatisée dans le domaine scientifique et technique, depuis les années 1960 : les possibilités offertes par l'informatique pour le traitement de l'information et la croissance exponentielle du volume de l'information.

Dans un rapport publié en automne 1973 par l'O.C.D.E., le Professeur Anderla affirme que le taux annuel d'accroissement de la littérature scientifique et technique de 12,5 % ne devrait pas diminuer dans un avenir prévisible, en raison de l'augmentation du nombre de scientifiques et de leur « productivité ». On atteindra donc quatorze millions de documents originaux publiés au cours de l'année 1985, soit sept fois plus qu'en 1970. Les publications produites pendant la décennie 1970-1980 seront en nombre supérieur à la somme de celles accumulées jusqu'en 1970.

Cette croissance se retrouve dans le domaine de la chimie : on a enregistré par exemple, en 1900, la synthèse de 1 000 composés nouveaux, contre 150 000 en 1970.

Chemical Abstracts Service (C.A.S.), dont la mission est de couvrir intégralement la littérature chimique mondiale, a enregistré un taux de croissance moyen annuel du nombre d'extraits de 10 % entre 1971 et 1973. En 1972, C.A.S. a signalé 380 000 documents tirés de 12 000 revues et des bulletins de brevets de vingt-six pays. Ce

* A.F.D.A.C., 88, avenue Kléber, 75116 Paris. Tél. 553.65.19.

taux de croissance n'est d'ailleurs pas uniforme pour tous les domaines traités par C.A.S. (tableau 1) :

Tableau 1.

Domaines	Sections de Chemical Abstracts	Importance en % des extraits	Augmentation en 1972-1971 en %
Biochemistry	1-19	30	15,9
Organic chemistry . . .	20-34	12	10,4
Polymer chemistry . . .	35-46	12	36,0
Applied chemistry . . .	47-64	17	8,8
Analytical and physical chemistry	65-80	29	1,4

II. Caractéristiques des systèmes documentaires automatisés : cas particulier de la chimie

1. L'automatisation de la constitution des fonds documentaires est une nécessité pour accélérer l'entrée des documents et abaisser le coût unitaire d'entrée. Il en résulte souvent une amélioration de l'accès aux documents, index par exemple, qui valorise le fonds. Dans certains cas ce sont des sous-produits de la constitution des fonds qui sont utilisés pour la recherche documentaire automatisée.

2. Ils doivent pouvoir satisfaire les besoins d'utilisateurs très divers : services de recherche et développement, services de propriété industrielle, services de fabrication. Les informations recherchées peuvent être de nature différente; d'autre part les services de recherche sont plus exigeants quant à la pertinence des réponses, tandis que les services de propriété industrielle doivent garantir l'exhaustivité des recherches documentaires.

3. Certains systèmes ont été rapidement opérationnels pour la *diffusion sélective de l'information (D.S.I.)* : elle consiste à signaler au demandeur les documents parus sur son sujet d'intérêt chaque semaine ou chaque mois; mais les besoins sont certainement beaucoup plus importants en *recherche rétrospective (R.R.)* : cette opération consiste à retrouver tous les documents traitant d'un sujet donné, parus depuis 10, 20 ou 50 ans.

4. En effet, particulièrement pour les recherches d'antériorité menées par les services de propriété industrielle, les documents « chimie » ne se périment pas.

5. Pour un système documentaire donné on ne peut évidemment retrouver à l'interrogation que les informations mémorisées à l'entrée. Celles-ci peuvent être classées en trois groupes :

a. informations permettant de situer le document : auteurs, titre de la revue, référence bibliographique, organisme, langue, pays, sections si le système documentaire comporte une classification, numéro d'accès dans le système; dans le cas des brevets : numéro du brevet, date de priorité, classes auxquelles il est affecté, brevets équivalents;

b. informations textuelles : titre, éventuellement résumé, ou mots-clés d'indexation. L'indexeur peut utiliser le langage naturel, ou choisir les mots-clés dans une liste établie à l'avance dénommée *thésaurus*. Les thésaurus comprennent en général une dizaine de milliers de termes : des relations de synonymie ou de hiérarchie entre les termes sont souvent précisées;

c. informations relatives aux structures chimiques : celles-ci peuvent être décrites par un système de nomenclature. Cependant, les nomenclatures étant peu connues des utilisateurs, et se prêtant mal au traitement automatisé des recherches documentaires sur les structures et les sous-structures, d'autres solutions ont été adoptées. Certains systèmes documentaires utilisent un *système de registre* : les composés définis décrits dans les documents sont mémorisés dans un fichier spécial et on leur affecte un *numéro de registre* (Registry Number : Reg. N°) qui permet de les identifier. Par ailleurs des systèmes de description des structures appelés *codes structuraux* ont été conçus depuis les années 1960, en vue du traitement sur cartes perforées ou en ordinateur. (Voir le complément sur les codes).

Certains codes structuraux sont particulièrement adaptés à la description des formules génériques, souvent appelées *formules de Markush*, rencontrées dans les brevets : ces formules comprennent une partie définie, avec des substituants de nature et de position variables; elles peuvent couvrir des milliers, parfois des centaines de milliers de composés distincts.

La finesse d'indexation d'un système documentaire varie en sens inverse de sa couverture; dans un système très spécialisé, le vocabulaire d'indexation peut être beaucoup plus détaillé que dans un système plus général.

Le nombre et la finesse des informations mémorisées ont une inci-

dence sur le coût d'entrée d'un document : celui-ci peut varier de 10 à 300 F suivant le degré de sophistication du système.

6. Le coût des services automatisés n'est pas négligeable pour l'utilisateur, d'autant moins qu'il s'agit de frais supplémentaires à régler à l'extérieur du laboratoire ou de la société. Ces dépenses doivent pourtant être mises en balance avec les heures de personnel passées à la recherche des documents. La comparaison devient alors beaucoup moins défavorable à la documentation automatisée qui a pour but de fournir au chercheur les références pertinentes, dans les meilleures conditions de rapidité et de précision. Les prix peuvent être abaissés si le système intéresse un grand nombre d'utilisateurs partageant le coût d'entrée des documents et le coût du traitement. C'est dans ce sens que travaillent des organismes collectifs tels que l'A.F.D.A.C., l'I.N.S.E.R.M., l'I.T.F. par exemple, qui doivent évidemment garantir rigoureusement le secret.

7. Les systèmes peuvent paraître difficilement accessibles aux utilisateurs. Force est de reconnaître que le degré de sophistication des systèmes est souvent tel qu'il exige soit la formation des utilisateurs, soit l'assistance de spécialistes, « interfaces » entre systèmes et utilisateurs. Le traitement d'un système documentaire automatisé requiert en fait une coopération entre informaticiens, documentalistes et utilisateurs.

III. Description des principaux systèmes documentaires automatisés

Dans le tableau descriptif 2 sont indiqués l'éditeur, le nom du système avec sa date d'origine, la couverture et le contenu (le nombre de documents analysés correspond à une année), les centres assurant le traitement, pour la diffusion sélective et la recherche rétrospective. En chimie les principaux fonds utilisés sont ceux du C.A.S., par voie automatisée pour la diffusion sélective, par voie manuelle pour la recherche rétrospective en raison de la qualité des index semestriels ou quinquennaux; ceux du Central Patents Index et d'I.F.I. pour les brevets; RINGDOC pour la chimie pharmaceutique.

IV. La politique de l'information en France

En 1972 a été créé le Bureau National pour l'Information Scientifique et Technique (B.N.I.S.T.) qui dépend du Ministère du Développement Industriel et Scientifique. Le B.N.I.S.T. est chargé d'aménager une politique sectorielle de la documentation scientifique et technique, et de coordonner les actions entre les divers secteurs : agriculture, biologie, chimie, électricité et électronique, médecine, métallurgie, nucléaire, pétrole, physique, sciences humaines, etc...

Pour le secteur chimie, le Centre National de l'Information Chimique (C.N.I.C.) créé en 1972 regroupe deux associations : l'Association Française de Documentation Automatique en Chimie (A.F.D.A.C.) créée en 1970 sur l'initiative de l'Union des Industries Chimiques et l'Association pour la Recherche et le Développement de l'Informatique Chimique (A.R.D.I.C.), créée en 1971 par le Professeur J. E. Dubois, qui travaille particulièrement au développement du système DARC basé sur le code topologique DARC (voir le complément sur les codes).

V. Les services de l'A.F.D.A.C. : bandes du Chemical Abstracts Service

Dès sa création, l'A.F.D.A.C. a décidé d'organiser le traitement des bandes C.A.S., en raison de la couverture étendue de ce fonds, de la richesse des informations qu'il contient, et de la grande habitude de Chemical Abstracts qu'ont les chimistes.

Description de Chemical Abstracts

Les documents analysés par C.A.S. sont signalés dans la revue Chemical Abstracts (C.A.). Les extraits avec résumés sont répartis dans 80 sections, et C.A. paraît toutes les semaines : une semaine paraît un fascicule impair comprenant les sections 1 à 34 (Biochemistry et Organic Chemistry); la semaine suivante paraît un fascicule pair comprenant les sections 35 à 80 (Polymers, Applied Chemistry, Analytical and Physical Chemistry).

Un document analysé par C.A.S. subit également d'autres traitements (figure 1) :

a. il est indexé de deux manières différentes. Une indexation rapide fournit les mots-clés des « Keyword Subject Index » situés à la fin des fascicules hebdomadaires. Une indexation plus complète et hiérarchisée fournit les entrées des index semestriels : « General Subject Index » et « Chemical Substance Index ». Résumés et index sont rédigés en langage naturel. Cependant, dans les index semestriels, le langage est contrôlé, sans qu'il soit fait usage d'un thésaurus;

b. les composés chimiques définis, y compris les polymères, décrits dans un document, sont introduits dans le système de registre de C.A.S. : le fichier, commencé en 1965, contient actuellement plus de 2,5 millions de structures. Un numéro de registre ou « Registry Number » est attribué à chaque structure non encore répertoriée.

Tableau 2.

Systèmes documentaires automatisés

Éditeur	Fonds	Couverture et contenu	D.S.I.	R.R.
	Chemical titles C.T. 1962	130 000 titres tirés de 650 journaux de chimie.	U.K.C.I.S. (G.B.) D.T.B. (D.K.)	Georgia 1962-1968
Chemical Abstracts Service (U.S.A. Columbus)	Chemical Abstracts Condensates C.A.C. 1968	Couverture de Chemical Abstracts (C.A.) : 390 000 documents tirés de 13 000 revues, brevets de 26 pays, titres, mots-clés des index hebdo- madaires, auteurs, sections, etc...	A.F.D.A.C. (F)	Georgia (U.S.A.) Consoles
	Chemical and Biological Activities C.B.A.C. 1965	37 000 documents : sections 1 à 5 de C.A.; titres, résumés, auteurs, sections, <i>numéros de registre</i> , etc...		Georgia 1965-1971
	Polymer and Science Technology POST 1968	45 000 documents : sections 35 à 46 de C.A.; même contenu que C.B.A.C.	A.F.D.A.C. (F)	
	Integrated Subject File I.S.F. 1967	Index semestriels de C.A. : General Subject Index et Chemical Substance Index. Mots-clés des index semestriels, numéros de registre.		
	C.A. Subject Index Alert C.A.S.I.A. 1973	Même couverture que C.A.C. : mots-clés des index semestriels, numéros de registre, etc...		
C.N.R.S. (Paris)	Pascal	Sciences exactes, biologiques, médicales, sciences de la terre, technologie, sciences humaines, 500 000 articles indexés tirés de 15 000 revues.	C.N.R.S. (F)	
I.S.I. (U.S.A. Philadelphie)	Source and Citation Tapes 1964	Couverture pluridisciplinaire : 350 000 documents tirés de 2 500 revues. Titres, auteurs, organismes, références citées.	I.S.I. (U.S.A.)	
	Index Chemicus Registry System I.C.R.S. 1960	13 000 documents tirés de 200 revues. Titres, auteurs, organismes, mots-clés, codage structural Wiswesser, 150 000 nouveaux composés par an.	I.S.I.	I.S.I.
BIOSIS	Biological Abstracts 1969	Sciences de la vie : 250 000 documents issus de 8 000 revues. Titres, auteurs, mots-clés.	Georgia BIOSIS	
N.L.M. (U.S.A.)	Medlars 1964	Médecine : 250 000 documents issus de 2 800 revues. Thésaurus.	I.N.S.E.R.M. (F)	I.N.S.E.R.M.
A.S.M. (U.S.A.)	Metals Abstracts 1968 Nuclear Science Abstracts N.S.A. 1970	Métallurgie : 25 000 documents issus de 1 200 revues. Thésaurus. Nucléaire : 55 000 documents. Titres, mots-clés. Thésaurus.	CEDOCAR (F) A.F.D.I.N. C.E.A.	CEDOCAR A.F.D.I.N. C.E.A.
I.T.F. (Paris)	TITUS 1971	Textile : 30 000 documents. Thésaurus.	I.T.F. (F)	I.T.F.
A.P.I. (U.S.A.)	A.P.I. 1966	Pétrole : technologie pétrolière et pétrochimie, 25 000 documents. Thésaurus.	I.F.P. (F)	I.F.P.
Engineering Index (U.S.A.)	COMPENDEX	Couverture pluridisciplinaire : sciences de l'ingénieur, 50 000 documents, résumés, mots-clés.	Georgia (U.S.A.)	Georgia
Derwent (Londres)	RINGDOC 1964 PESTDOC 1969 VETDOC 1969	Pharmacie : 45 000 documents. Titres, auteurs, mots- clés (Thésaurus); codage structural. Pesticides : 8 000 documents; même contenu. Sciences vétérinaires : 5 000 documents; même contenu.	Mises à jour trimestrielles	Réservée aux souscripteurs
Derwent (Londres)	Central Patents Index C.P.I. 1963 à 1970 suivant sections	100 000 brevets répartis en 12 sections : plastiques (A), pharmacie (B), agriculture (C), chimie (D), métal- lurgie, nucléaire, pétrole, etc... Thésaurus propre à chaque section. Codage structural pour les sections B, C et E.	Mises à jour trimestrielles	Réservée aux souscripteurs
I.D.C. (Francfort)	I.D.C. 1959	Chimie organique : 40 000 documents tirés de 300 revues, et des sections B, C, et E du C.P.I. Thésaurus. Code structural GREMAS.	I.D.C. (R.F.A.) U.K.C.I.S. (G.B.)	I.D.C. U.K.C.I.S.
I.F.I. (U.S.A., Arlingter)	Uniterm 1950	Brevets américains chimie : 23 000 par an, mots- clés (Thésaurus), système de registre, codage struc- tural, société, classification des brevets.	Mises à jour périodiques	I.F.I.
	Comprehensive Data Base 1950	Même couverture qu'Uniterm. Thésaurus, système de registre et codage structural plus complet que dans Uniterm.		I.F.I. I.F.I.
Rhône-Poulenc	Diapason 1972	Chimie et polymères : 10 000 articles et 30 000 bre- vets tirés du C.P.I. Thésaurus polymères, codage structural Diapason.	Rhône-Poulenc	Rhône-Poulenc

Fabrication, Morphology, and Dynamic Mechanical Properties of a Model Composite System Containing in Situ-Grown Filler

J. L. KÁRDOS, W. L. McDONNELL, and J. RAISONI
Department of Chemical Engineering
Washington University
St. Louis, Missouri

Summary

The technique of in situ crystallization was utilized to fabricate a model composite system in which the filler morphology is variable under constant interface conditions. A butadiene-acrylonitrile copolymer was chosen as the matrix from which acetanilide was crystallized in two distinctly different crystal morphologies for filler volume loadings up to 0.35. At the same volume fraction filler, the shape of the relative modulus-filler loading curve is sensitive to the filler crystallization temperature and, thus, to the filler size and shape. For a given crystallization temperature (25°C), the data in the low-volume loading region (<0.2) can be represented reasonably well with the Mooney equation which indirectly yields a filler aspect ratio in agreement with scanning electron micrographs. When the observed morphological parameters are utilized for the entire filler loading range, the 25°C data are best predicted by the Halpin-Tsai equation in conjunction with lamination theory. In general, filler interlocking at high filler volume loadings (~0.3) causes the experimental modulus to rise significantly higher than can be predicted with existing theory and the known filler morphology. The normalized damping ratio, $(G''/G')_f / (G''/G')_m \phi_f$, rises significantly above 1.0 with increased filler loading at temperatures above T_g . This rise is attributed to changes in the properties of the polymer matrix near and at the interface. The in situ-grown composite system described here also provides an interesting link between the equations used to design structural composite materials and crystalline polymers.

397

Copyright © 1972 by Marcel Dekker, Inc. NO PART of this work may be reproduced or utilized in any form or by any means, electronic or mechanical, including xerography, photocopying, microfilm, and recording, or by any information storage and retrieval system, without the written permission of the publisher.

Extrait C.A.

73124x Fabrication, morphology, and dynamic mechanical properties of a model composite system containing in situ-grown filler. Kardos, J. L.; McDonnell, W. L.; Raison, J. (Dep. Chem. Eng., Washington Univ., St. Louis, Mo.). *J. Macromol. Sci., Phys.* 1972, 6(2), 397-412 (Eng). Acetanilide [103-84-4] was in situ-crystd. in a butadiene-acrylonitrile copolymer - [9003-18-3] matrix in 2 distinctly different CRYST. MORPHOLS. for filler vol. loadings $\leq 35\%$. At the same vol. fraction filler, the shape of the relative modulus-filler loading curve was sensitive to the filler crystn. temp., and to the filler size and shape. For a given crystn. temp. (25°), the data in the low-vol. loading region (<20 vol. % filler) was represented with the Mooney equation which indirectly yielded a filler aspect ratio in agreement with scanning electron micrographs. This 25° data was best predicted by the Halpin-Tsai equation in conjunction with lamination theory when the obsd. morphol. parameters were utilized for the entire filler loading range. Filler interlocking at high filler vol. loadings (~30%) caused the exptl. modulus to rise significantly higher than could be predicted with existing theory and known filler morphol. Normalized damping ratio rose significantly above 1 with increased filler loadings at temps. > the glass transition temp. This rise was attributed to changes in the properties of the polymer matrix near and at the interface.

Extrait C.A.C.

**CAC *076(014)*073124X*S0036* JMAPBR*

FABRICATION, MORPHOLOGY, AND DYNAMIC MECHANICAL PROPERTIES OF A MODEL COMPOSITE SYSTEM CONTAINING IN SITU-GROWN FILLER.

KARDOS, J. L.; MCDONNELL, W. L.; RAISONI, J. (DEP. CHEM. ENG. WASHINGTON UNIV., ST. LOUIS MO.).

J. MACROMOL. SCI., PHYS. (PUB. 000072) 6 (2) 397-412 (EN)

COPOLYMER FILLER CRYSTALS ACETANILIDE CRYSTAL GROWING BUTADIENE COPOLYMER CRYSTAL MORPHOL FABRICATION BUTADIENE COPOLYMER CRYSTAL

(Mots-clés des index hebdomadaires)

Indexation dans les index semestriels

Chemical Substance Index

Acetamide, N-phenyl-[103-84-4],
crystn. of, in butadiene-acrylonitrile copolymer, crystal form of, 73124x.
1,3-Butadiene, polymers
polymer with 2-propene nitrile [9003-18-3] crystn. of acetanilide in, and morphology of product there from, 73124x.
2-Propenenitrile, polymer
polymer with 1,3-butadiene [9003-18-3] crystn. of acetanilide in, and morphology of product there from, 73124x.

General Subject Index

Crystal form
of acetanilide fillers, in butadiene-acrylonitrile polymer, 73124x.

Figure 1

Le numéro de registre est le lien entre toutes les parties du fonds C.A.S. Il apparaît pour le moment dans les résumés des sections 1 à 5 et 35 à 46, et dans les index semestriels pour l'ensemble des documents.

Description des bandes Chemical Abstracts Condensates (C.A.C.) et Polymer Science and Technology (POST)

Tous les documents analysés dans C.A. font l'objet d'un extrait dans les bandes C.A.C. : comme C.A., celles-ci paraissent toutes les semaines ; une semaine paraît une bande C.A.C. 1, la semaine suivante une bande C.A.C. 2. Un extrait C.A.C. comprend le titre, les auteurs, la référence bibliographique, le code abrégé de la revue, le type de publication (journal, brevet, livre), la langue, l'organisme où travaillent les auteurs, la section de C.A., et les mots-clés des « Keyword Subject Index ».

Les bandes POST contiennent les extraits des sections 35 à 46 (Polymers), avec les mêmes informations que sur les bandes C.A.C., et en outre les résumés, les numéros de registre et les formules brutes des composés.

Diffusion sélective : les profils

Un profil est une liste de termes associés entre eux au moyen d'une équation logique : il traduit la question de l'utilisateur et sert de clé de tri pour sélectionner sur une bande les documents répondant à la question.

Les profils de l'A.F.D.A.C. sont traités depuis 1971 par le Centre de Documentation de l'Institut Français du Pétrole, qui a mis au point le programme de lecture de texte PRETEXT II. Celui-ci permet l'utilisation des tronçatures des termes à droite et à gauche (pour retrouver les mots contenant un fragment donné), l'association des termes par des logiques « booléennes » (ET, OU, SAUF), ou par des logiques « syntaxiques » (AVEC, IGNORE) qui permettent de préciser les distances désirées entre les termes.

Nous donnons un exemple de profil, et une référence retrouvée. Le sujet est le suivant : antiinflammatoires nouveaux, non-stéroïdes ; brevets.

On interroge par les termes relatifs aux antiinflammatoires : ANTIINFLAMMATORY, ANTIINFLAMMATION, INFLAMMATION INHIBITOR, ANTIPHLOGISTIC, les termes NEW, NONSTEROID, brevets (type de publication : P) et on exclut la section 32 relative aux stéroïdes. D'où la formulation du profil :

```
001 TXT ANTIINFLAM*
002 TXT ANTIPHLOGISTIC
003 TXT INFLAMMATION
004 TXT INHIBIT*
005 TXT NEW
006 TXT NON STEROID*
007 TXT NONSTEROID*
008 TXT NON-STEROID*
009 CPU P
010 SEC 032
(001 OU 002 OU (003 ET 004))
ET (005 OU 006 OU 007 OU 008 OU 009)
ET NON 010
```

Sur la référence retrouvée, sont éditées les informations suivantes (figure 2) :

- 1 : référence dans C.A. : volume, fascicule, numéro d'extrait ;
- 2 : section de C.A. : 025 (Noncondensed Aromatic Compounds) ;
- 3 : code des brevets espagnols ;
- 4 : titre ;
- 5 : organisme ;
- 6 : identification du brevet : pays, numéro, classes de la classification internationale, dates de publication et de priorité ;
- 7 : mots-clés du Keyword Subject Index ;
- 8 : termes du profil retrouvés dans l'extrait.

**CAC *078(005)*029455Y*025* SPXXAD*

- ① *EPSILON.-AMINOCAPROIC ACID SALTS. PUB.IN 375,998. (LA
 ② BORATORIOS BAMA, S. A.).
 ③ SPAN. 375998 (CL. C 07C, A 61K), (PUB. 160572) ZZ, A
 PPL. 290170; 5 PP.
 ④ *CAPROATE AMINO SALICYLATE ANTIINFLAMMATION *INFLAMMAT
 ION INHIBITOR AMINOCAPROIC SALICYLATE *ANTIINFLAMMATORY
 AMINOCAPROIC SALICYLATE SALT
 ⑤ /ANTIINFLAM*/INFLAM*/*INHIBIT*/P

PROFIL : A13005

CARTE : 003

Institut Français du Pétrole 1 et 4, avenue de Bois-Préau 92-Rueil-Malmaison FRANCE

Figure 2.

L'A.F.D.A.C. traite actuellement 520 profils répartis comme l'indique le tableau 3 :

Tableau 3.

	C.A.C. 1	C.A.C. 2	POST	Total
Entreprises	135	223	73	431
Organismes de recherche (C.E.A., CEDOCAR, I.N.R.A., I.N.S.E.R.M.)	28	41	5	74
Laboratoires universitaires et C.N.R.S.	5	10	—	15
TOTAL	168	274	78	520

Le prix d'un profil dépend du nombre de termes de recherche et du nombre de références retrouvées. Il peut varier de 350 à 4 000 F par an suivant son importance. Le prix moyen annuel (20 termes de recherche, 20 références par quinzaine) est de 750 F pour C.A.C. 1 et C.A.C. 2, 930 F pour POST.

Recherche rétrospective

Comme nous l'avons déjà dit, la recherche rétrospective suscite certainement plus de demandes que la diffusion sélective. Pour les fonds du C.A.S., la recherche rétrospective automatisée ne peut remonter au-delà de 1968 avec les bandes C.A.C., au-delà de 1967 avec les bandes I.S.F. Mais ces dernières ne sont encore exploitées systématiquement par aucun centre. Seule l'Université de Georgia a entrepris une étude approfondie de ces bandes et montré que leur traitement serait très coûteux en raison de l'abondance des informations qu'elles contiennent. Actuellement, seules les bandes C.A.C. sont exploitées à des fins de recherche rétrospective.

a) Fichier séquentiel

L'Université de Georgia a constitué un fichier de recherche rétrospective à partir des bandes C.A.C. en les organisant séquentiellement, par semestre, dans l'ordre des numéros d'extraits. Les questions sont traitées par lots : Georgia en a traité 2 800 en cinq ans ; l'A.F.D.A.C. en a fait traiter une trentaine depuis un an. Le prix d'une recherche est de 250 F par semestre et par fonds interrogé (C.A.C. 1 ou C.A.C. 2). Le délai d'obtention des réponses varie entre six semaines et deux mois.

b) Fichier inversé

Pour remédier à la lenteur et au prix de ce procédé, deux organismes ont entrepris de restructurer les informations contenues dans les bandes C.A.C. en constituant un fichier « inversé » : chaque terme, auteurs, organisme, numéro de section, mots-clés par exemple, constitue une rubrique dans laquelle sont répertoriés tous les numéros d'extraits contenant ce terme. Une telle organisation permet l'interrogation du fichier en conversationnel, avec une console spéciale reliée au fichier central par une liaison téléphonique.

Un organisme américain, System Development Corporation, diffuse le système CHEMCON (Chemical Abstracts on-line) : l'A.F.D.A.C. aura très prochainement accès au système CHEMCON et proposera des recherches rétrospectives au tarif expérimental de 400 F par question jusqu'en décembre 1974.

L'E.S.R.O. diffuse déjà, en conversationnel, les systèmes documentaires de la N.A.S.A., METADEX (métallurgie), COMPENDEX,

Nuclear Science Abstracts, etc. L'E.S.R.O. termine actuellement l'inversion des bandes C.A.C. pour l'usage d'un groupement d'entrées hollandaises. L'accès à ce système est en cours de discussion.

VI. Conclusion

La documentation automatisée est déjà une réalité; de nombreux systèmes fonctionnent déjà de manière opérationnelle. Dans les grandes entreprises de la chimie, comme l'a montré une enquête récente lancée par l'A.F.D.A.C., environ 50 % des recherches documentaires sont effectuées par voie automatisée. Cependant des actions importantes doivent encore être entreprises, essentiellement dans deux directions :

l'information et la formation des utilisateurs;

le développement de la recherche rétrospective, en particulier sur les structures chimiques.

Des moyens importants doivent être mis en œuvre, ce qui exige une coopération nationale et internationale, pour éviter les doubles emplois, diminuer les coûts d'entrée, de traitement, et d'accès aux systèmes.

Complément sur les codes structuraux

Dans le domaine de la documentation chimie, comme nous l'avons déjà dit, se pose le problème de la description des structures : composés définis et formules de Markush. Pour remédier aux insuffisances de la nomenclature, différents systèmes de codage des structures chimiques ont été conçus ces quinze dernières années. Nous les classerons en trois types : codes fragmentaires, codes topologiques, codes linéaires.

I. Codes fragmentaires

Les codes fragmentaires décrivent seulement certains motifs structuraux des molécules, ou fragments, définis lors de la conception du code; les fragments décrivent les cycles, les chaînes, les fonctions, les substituants et leurs positions relatives.

Le codage n'est pas biunivoque, c'est-à-dire qu'à partir des fragments décrivant une structure, on ne peut pas reconstituer entièrement la formule développée. Il en résulte que toute modification ou utilisation, non prévue initialement, du fichier des structures codées nécessitera que l'on remonte aux formules développées.

Le codage ne peut être complètement automatisé, sauf s'il est précédé d'un codage topologique.

Les codes fragmentaires ont été conçus en vue de recherches documentaires de structures ou sous-structures portant sur les composés définis et surtout sur les formules de Markush.

Citons, comme principaux codes fragmentaires :

1. les codes RING (code général, code stéroïdes, code peptides) utilisés pour la littérature pharmaceutique (RINGDOC) et pour les brevets des sections FARMDOC et AGDOC du Central Patents Index. Ces codes ont été conçus pour des cartes perforées. Le code général comprend 350 descripteurs;
2. le code GREMAS utilisé par l'I.D.C. (Internationale Dokumentationsgesellschaft für Chemie, Francfort), pour la littérature et les brevets de la chimie organique; il comporte environ 3 000 descripteurs;
3. le code Diapason utilisé par Rhône-Poulenc pour la littérature et les brevets de la chimie organique; conçu pour l'ordinateur, ce code comprend théoriquement un nombre illimité de fragments;
4. le code I.F.I./Du Pont utilisé par I.F.I. dans le fonds Comprehensive Data Base, qui signale vingt années de brevets américains, et comporte 9 000 fragments.

II. Codes topologiques

Les codes topologiques décrivent tous les atomes et toutes les liaisons des composés définis : le codage obtenu est donc biunivoque.

Le codage peut être automatisé : il est possible d'entrer les données relatives aux structures définies en ordinateur, afin que celui-ci génère automatiquement le codage.

Le codage topologique réserve l'avenir : toutes les modifications ou nouvelles utilisations des structures codées pourront être réalisées automatiquement, à partir du codage topologique mémorisé. Par exemple, le codage topologique permet le transcodage d'un code topologique à un autre, ou d'un code topologique à un code fragmentaire. En effet, à partir d'un code topologique, on peut générer automatiquement un système d'écrans dans lesquels peuvent être inclus les motifs structuraux d'un code fragmentaire donné : ces écrans permettent les tris préliminaires à la recherche documentaire.

Les codes topologiques ont été surtout conçus pour automatiser l'entrée des structures définies lors de la constitution des fichiers. Pour les recherches de structures, l'interrogation se fait d'abord au moyen d'écrans qui permettent d'éliminer au moins 99,5 % des structures du fichier. Pour améliorer la pertinence des réponses, la recherche peut alors être poursuivie au moyen de la topologie mais ceci a rarement lieu en raison du coût très élevé d'une telle opération. Le choix des écrans est très important puisqu'il conditionne en grande partie le coût des systèmes.

Citons, comme principaux codes topologiques :

1. les matrices de connectivité de Chemical Abstracts Service; le système de registre de C.A.S. comprend actuellement 2,5 millions de composés définis et polymères; il est utilisé par C.A.S. essentiellement pour l'attribution des numéros de registre aux composés chimiques décrits dans la littérature. C.A.S. a également développé un système d'écrans mais en pratique ne fait pas de recherches de sous-structures. C.A.S. a fourni des échantillons de son fichier à quelques centres qui ont développé leurs propres systèmes d'écrans pour des expérimentations. L'Université de Georgia a fait des tests sur 178 000 structures, Lynch à l'Université de Sheffield sur 60 000 structures, Lefkovitz à l'Université de Pennsylvanie sur 1,25 millions de structures. A Paris, le Professeur Dubois dispose de 500 000 structures;

2. le code topologique de l'I.D.C. conçu pour automatiser l'entrée des structures définies. Les formules de Markush sont codées manuellement en GREMAS. Le fichier de l'I.D.C. comporte 1,5 million de composés définis et formules de Markush.

Le code topologique de l'I.D.C. peut être généré automatiquement à partir des matrices de connectivité de C.A.S. Le code GREMAS et le code RING peuvent être également générés automatiquement à partir du code topologique.

L'I.D.C. effectue les recherches documentaires au moyen du code GREMAS. Il a parfois recours à la topologie si la recherche GREMAS apporte un bruit excessif.

L'I.D.C. a entrepris des études sur le codage topologique des formules de Markush. Une solution a été exposée, mais il semble qu'elle ne soit pas encore appliquée opérationnellement;

3. le code DARC, développé en France par le Professeur Dubois. Le transcodage de 500 000 structures C.A.S. en DARC est en cours; des écrans DARC sont ensuite générés automatiquement, et des procédures de restitution sont en cours de développement.

III. Codes linéaires

Nous citerons seulement le code Wiswesser, exploité par I.S.I. (Institute for Scientific Information, Philadelphie) dans le fonds Index Chemicus Registry System qui comprend un million de composés.

Le code s'apparente aux codes topologiques car il est biunivoque. La notation Wiswesser est très concise, mais le grand nombre de règles mises en jeu semble rendre l'automatisation du codage difficile. Des programmes ont été développés pour générer à partir de la notation Wiswesser une matrice de connectivité, puis les codes RING.

La figure 3 montre le codage d'une structure dans différents codes.

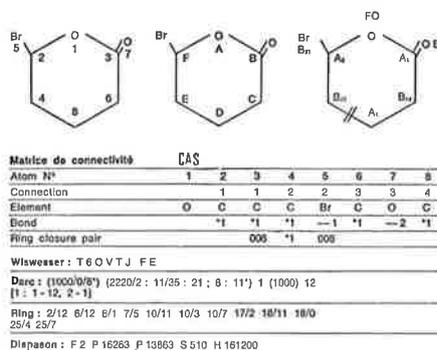


Figure 3.

IV. Conclusion

Certes le chimiste ou le documentaliste reste perplexe devant la multiplicité des codes et souhaiterait une solution universelle permettant de satisfaire tous les besoins. Est-ce une utopie?

La topologie est assurément une voie d'avenir car elle permet la compatibilité entre systèmes grâce au transcodage, mais elle n'est pas encore en mesure de prendre en compte les formules de Markush. Par ailleurs, la constitution de larges fichiers de structures nécessite un financement important, impliquant inévitablement une coopération nationale et internationale toujours difficile à mettre en œuvre.

Enfin, il faut être conscient des difficultés entraînées par le lancement d'un nouveau système. Pour l'utilisateur c'est être contraint d'apprendre une nouvelle langue. En outre, l'existence de systèmes documentaires opérationnels depuis 5 voire 10 ans au moyen de codes fragmentaires, ont entraîné la création de fichiers très importants, ce qui contribue également à freiner l'évolution vers de nouveaux systèmes.

PS

A l'heure où ces lignes vont paraître, nous voudrions ajouter que les profils sont maintenant traités, avec le programme PRETEXT II, par l'A.F.D.A.C. qui dispose d'un terminal relié à l'ordinateur de FRANLAB. Par ailleurs, le service de recherche rétrospective en conversationnel sur le système CHEMCON fonctionne depuis mai 1974 et se développe rapidement.