

## **Le système DARC : un outil au service de la recherche**

**par Alain Deroulède, Christian Dutheuil  
et Marie-Christine Jérôme**

*(Centre National de l'Information Chimique, 26, rue Boyer,  
75020 Paris)*

Le Centre National de l'Information Chimique (CNIC) met à la disposition du public plusieurs bases de données sur le centre serveur TÉLÉSYSTÈMES-QUESTEL. Ces bases de données se distinguent de celles qui sont disponibles sur les autres « serveurs » car elles peuvent être interrogées directement « en mode conversationnel » en utilisant la structure moléculaire pour définir les composés chimiques, c'est-à-dire simplement en décrivant la formule développée des molécules. Ceci est possible grâce au système DARC\*.

Le système DARC, inventé par le Professeur J. E. Dubois, est un ensemble de logiciels permettant de décrire la formule développée des molécules dans un langage compréhensible par un ordinateur qui est alors capable d'identifier des composés chimiques et, même, capable de reconnaître tous les composés qui contiennent un élément structural donné, et de les restituer. La description d'une structure, ou d'un de ces éléments (sous-structure) est faite à l'aide de commandes simples :

- acquisition du squelette sous forme alphanumérique ou graphique (dessin par table de digitalisation ou terminal graphique),
- description de la nature des liaisons,
- description de la nature des atomes...

L'une des premières applications du système DARC est la recherche bibliographique automatisée en chimie. La recherche bibliographique par le texte suscite rarement l'enthousiasme des chimistes; en revanche, la nouvelle approche par les structures, provoque un intérêt pour l'accès à l'information en employant le langage du chimiste, d'autant plus que les possibilités vont au-delà de la simple recherche bibliographique.

Le système DARC a été adapté au fichier « Registry Structure File », réalisé par Chemical Abstracts Service (CAS). Ce fichier contient la description mathématique\*\* de toutes les structures définies en correspondance avec leurs numéros de registre (RN).

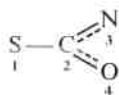
Le RN est un numéro d'identification attribué par CAS à tout composé lors de sa première citation bibliographique. Il est ensuite réutilisé pour indexer les nouvelles citations. Le Chemical Registry System, commencé en 1965, contient actuellement 5,8 millions de substances chimiques et s'enrichit de 350 000 nouvelles structures par an, soit près de 7 000 par semaine.

C'est certainement le dictionnaire chimique le plus complet dans le monde. Le système DARC permet donc d'interroger ce fichier à partir de la description de la formule développée, qui est univoque, compréhensible et adoptée par tous les chimistes. La désignation d'un composé chimique par sa formule développée est considérablement plus simple que par sa nomenclature (règles compliquées, non universelles et modifiées au cours du temps).

\* DARC = Description, Acquisition, Restitution, Conception.

\*\* Sous forme de matrice de connectivité.

Le système DARC permet de retrouver dans les fichiers Chemical Abstracts le numéro de registre (RN) correspondant à une structure. En général, à chaque structure correspond un seul RN. CAS considère les isomères et les dérivés isotopiques d'un composé comme des substances distinctes et à chacun d'eux affecte un RN différent. De même, chaque sel et chaque mélange définis se voient attribuer un RN. Ainsi, à partir de la structure de l'acide ascorbique, on retrouve 34 RN correspondant à différents isomères ou dérivés isotopiques, et 273 RN correspondant à des mélanges définis dans lesquels l'acide ascorbique est l'un des constituants.

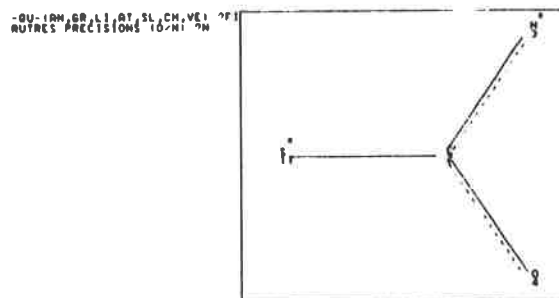


La nature des atomes, puis celle des liaisons sont ensuite définies.

```
-OU- IAN,GR,LI,AT,SL,CH,VEI ^AT
^ATOMES
^C 1
^S 1
^N 3
^O 4
^
-OU- IAN,GR,LI,AT,SL,CH,VEI ^LI
^LIAISONS
^S1 1-2
^TR 2-3-4
^
-OU- IAN,GR,LI,AT,SL,CH,VEI ^SL
^SITES LIBRES
^1 1,3
^
-OU- IAN,GR,LI,AT,SL,CH,VEI ^VE
```

La commande SL (sites libres) permet d'indiquer les possibilités de substitution : 1 substituant possible en position 1 ou 3.

La structure décrite peut être vérifiée.



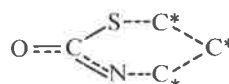
## b) La recherche

Une fois la structure décrite, la recherche est alors effectuée automatiquement par le système DARC, avec la possibilité d'intervenir éventuellement pour apporter des précisions complémentaires. Dans notre exemple, la structure très « ouverte » offrant la possibilité d'un grand nombre de substituants, il est prudent de tester la question sur le fichier échantillon « MINICAS », correspondant à une extraction de 1 % de la base EURECAS (5,6 millions de composés). Le fichier MINICAS est destiné à la mise au point des questions.

La recherche elle-même se fait en deux étapes. La première, (RE) consiste en la sélection automatique des composés candidats comportant des caractéristiques communes de la structure de départ. La seconde permet d'aboutir au résultat final par comparaison des composés candidats avec la structure de départ (recherche atome par atome AA).

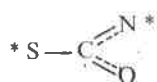
Dans notre exemple, l'interrogation de MINICAS fournit 42 composés. Sur le fichier EURECAS, on peut s'attendre approximativement à 4 200 réponses. La question formulée sur MINICAS peut être transférée et traitée automatiquement sur EURECAS. En fait, on obtient à l'issue de l'étape (AA) 4 551 composés dans le fichier EURECAS. Un tel nombre de composés est trop élevé pour exploiter toute l'information dont ils ont pu faire l'objet, aussi on peut introduire des précisions et la recherche est alors effectuée à partir du sous-fichier constitué par les composés-réponses obtenus à la fin de la recherche initiale. Dans notre exemple, nous pouvons ainsi indiquer que la sous-

structure  $S-C \begin{matrix} \diagup O \\ \diagdown N \end{matrix}$  fait partie d'un cycle thiazine dont on laisse indéterminée la nature des liaisons  $S-C-C-C-N$  et pour lequel on admet la possibilité de substitutions sur les atomes de carbone :



## Recherche par « sous-structure »

Le système DARC permet de reconnaître les composés qui possèdent en commun un élément de structure donné. Ainsi, il est possible de retrouver, en quelques minutes, dans le fichier EURECAS qui contient près de 6 millions de structures, tous les composés qui contiennent la sous-structure suivante :



Les possibilités de substitutions sont représentées par les astérisques. C'est le genre de question difficile pour ne pas dire impossible à exprimer à l'aide de la nomenclature. A une question ainsi posée, le système DARC répond en donnant les 4 551 numéros de registre des composés contenant cette sous-structure et pour chacun restitue la formule développée correspondante (chaque fois, des structures qui n'avaient pas été imaginées au départ sont retrouvées). Très rapidement, une famille de composés, découverts ou ayant fait l'objet de publications depuis 1965, peut ainsi être reconstituée.

Les fichiers « structuraux » interrogeables à l'aide du système DARC sur le serveur TÉLÉSYSTÈMES-QUESTEL sont associés à des fichiers « bibliographiques » qui peuvent être interrogés soit directement à partir d'éléments textuels (mots-clés, mots du titre, noms des auteurs, type du document, etc.) comme sur les autres serveurs, soit à partir des composés retrouvés à l'aide du système DARC. En effet, à l'issue d'une recherche « structurale », tous les numéros de registre des composés retrouvés peuvent être transférés en une seule étape dans les bases de données bibliographiques. Les questions peuvent alors être précisées avec des termes d'indexation, et c'est la combinaison des recherches structurale et textuelle qui permet une sélection rapide de toutes les références qui répondent à la question.

## Exemple d'une recherche de composés à l'aide du système DARC

Pour illustrer les procédures d'interrogation à l'aide du système DARC, on peut reprendre l'exemple cité précédemment : recherche de tous les composés qui contiennent l'élément de structure :



Les atomes marqués d'un astérisque peuvent être substitués; les atomes d'hydrogène sont pris en compte automatiquement.

### a) Formulation de la question

Cette sous-structure peut être décrite en mode graphique (terminal graphique) ou en mode alphanumérique comme dans l'exemple traité.

Un numéro est attribué à chaque atome, dans un ordre quelconque. Le graphe est alors décrit en indiquant l'enchaînement des atomes : 1-2-3, 2-4. Un tiret représentant une liaison entre les atomes, et une virgule séparant les numéros des atomes qui ne sont pas directement liés.

```

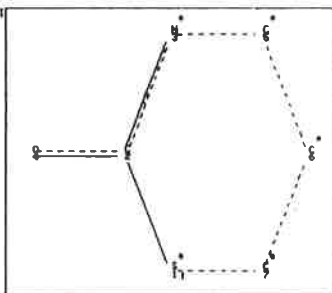
-BT-(QU,RE,AA) 7QU
** QU **
QUESTION PRECEDENTE (O/N) ?O
-BU-(AN,GR,LI,RT,SL,CH,VE) ?GR
ENTREE GRAPHIQUE (O/N) ?N
**GRAPHIE
?3-5-6-7-1
?
-BU-(AN,GR,LI,RT,SL,CH,VE) ?RT
**ATOMES
?C 5,6,7
?
-BU-(AN,GR,LI,RT,SL,CH,VE) ?LI
**LIÉONS
?N 3-5-6-7-1
?
-BU-(AN,GR,LI,RT,SL,CH,VE) ?SL
**SITES LIÉS
?3 5,6,7
?
-BU-(AN,GR,LI,RT,SL,CH,VE) ?VE

```

```

-BU-(AN,GR,LI,RT,SL,CH,VE) ?FI
AUTRES PRÉCISIONS (O/N) ?N

```



```

-BT-(QU,RE,AA) 7RE
** RE **
RÉSULTAT 37
** RE ** NOMBRE DE RÉPONSE(S) : 37 **
-BT-(QU,RE,AA) 7AA
** AA **
** AA ** NOMBRE DE RÉPONSE(S) : 1 **
-BT-(QU,RE,AA) 7FI
** FI **
CMD (BA,ST,RN,BI) 7BA
** BA **
TEMPS ÉCOULÉ SUR "MINICAS" : 18 23
BASE (CASC/MINICAS/EURECAS/UPCAB) ?EURE

```

```

** BASE EURECAS /CNIC : 4 683 962 STRUCTURES
DERNIER RN : 88329-16-2
CMD (BA,ST,RN,BI) ?BT
** BT **
-BT-(QU,RE,AA) 7QU
** QU **
QUESTION PRÉCÉDENTE (O/N) ?O
-BU-(AN,GR,LI,RT,SL,CH,VE) ?FI
AUTRES PRÉCISIONS (O/N) ?N

```

```

-BT-(QU,RE,AA) 7RE
** RE **
RÉSULTAT 3849
** RE ** NOMBRE DE RÉPONSE(S) : 3849 **
-BT-(QU,RE,AA) 7AA
** AA **
** AA ** NOMBRE DE RÉPONSE(S) : 137 **

```

```

-BT-(QU,RE,AA) 7HI
** HI **
** L S ** 137 STRUCTURE(S) 137 RN
-BT-(QU,RE,AA) 7VI
** VI **

```

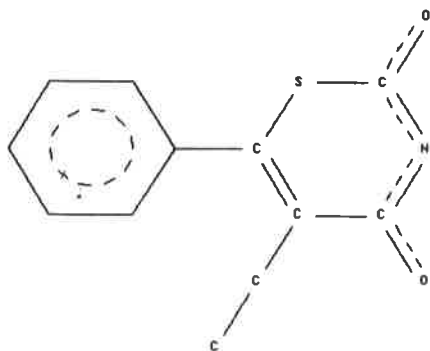
Dans ce cas là, le nombre de réponses est ramené à 1 dans MINICAS et 137 dans EURECAS. Les structures des composés retrouvés peuvent être visualisées sur l'écran d'un terminal graphique et l'édition en différé de ces structures peut être commandée.

RÉPONSE À VISUALISER (0 POUR FINI) ?1

```

C.N.I.C.-----S Y S T E M E D A R C-----C.N.I.C. / EURECAS
RÉPONSE 1 1 COMPOSE NUMERO 1 45703 RN 1 549-36-0
1 FRAGMENT 16 ATOMES

```

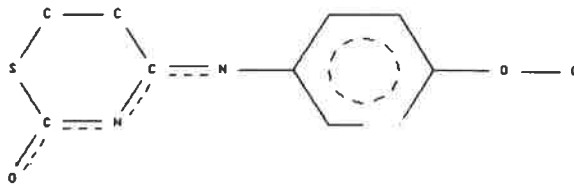


FORM. MOL. : C12H11N2O2S  
?A

```

C.N.I.C.-----S Y S T E M E D A R C-----C.N.I.C. / EURECAS
RÉPONSE 1 58 COMPOSE NUMERO 1 1170626 RN 1 21427-74-7
1 FRAGMENT 16 ATOMES

```

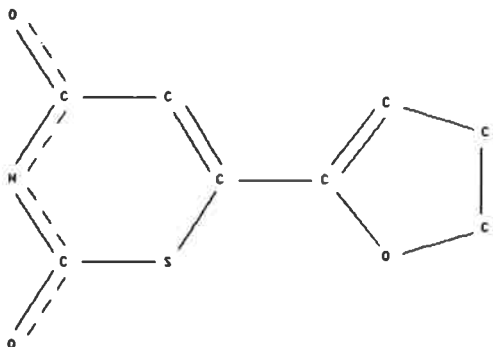


FORM. MOL. : C11H12N2O2S  
?A

```

C.N.I.C.-----S Y S T E M E D A R C-----C.N.I.C. / EURECAS
RÉPONSE 1 188 COMPOSE NUMERO 1 5812962 RN 1 54491-68-8
1 FRAGMENT 13 ATOMES

```

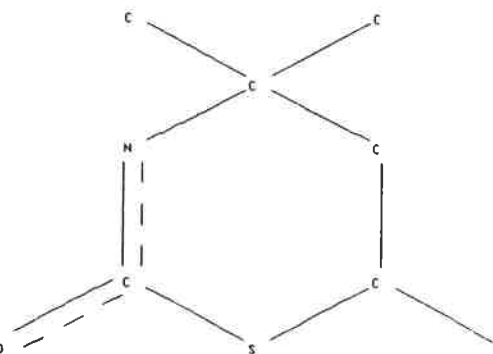


FORM. MOL. : C8H5N2O3S  
?A

```

C.N.I.C.-----S Y S T E M E D A R C-----C.N.I.C. / EURECAS
RÉPONSE 1 137 COMPOSE NUMERO 1 5210569 RN 1 79696-62-1
1 FRAGMENT 18 ATOMES

```



FORM. MOL. : C7H13N2O3S  
?F

```

-ST-(QU,RE,AA) 7FI
** FI **
CMD (BA,ST,RN,BI) ?BI
** BI **
TEMPS ÉCOULÉ SUR "EURECAS" : 9 38
# DARC # VOUS REMERCIÉ, AU REVOIR
*QUESTEL* 8102

```

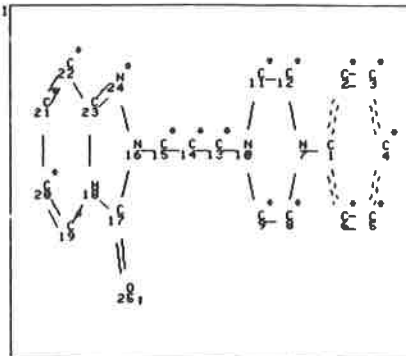
La liste des numéros de registre des 137 composés retrouvés peut être transférée dans la partie bibliographique (EUCAS) de la base EURECAS. EUCAS est constitué de plusieurs fichiers qui correspondent aux périodes des index cumulatifs de CAS : 1967-1971 : EUCAS 67, 1972-1876 : EUCAS 72, 1977-1981 : EUCAS 77, et 1982 : EUCAS 82. Les composés retrouvés ici par DARC sont connus pour leur application en photographie. En combinant leur liste avec le mot-clé « Photosensitive » on retrouve dans le fichier EUCAS 1977-1981 les références de deux brevets.

```

INFO.  RE MU.  BASE ?
? BA EUCAS77
BASE COMMERCIELE EUCAS77
COMMANDE, OU ETAPE DE RECHERCHE 1
? PP DARI LS
00 RN TRAITES 0 RESULTAT 11
00 RN ABSENTS 0 RESULTAT 11
00 RN ABSENTS 0 EDITION (O/N) ?
?M
120 RN TRAITES 0 RESULTAT 14
00 RN ABSENTS 0 RESULTAT 16
COMMANDE, OU ETAPE DE RECHERCHE 2
?S ET PHOTOSENSITIVE
020 RESULTAT 2
COMMANDE, OU ETAPE DE RECHERCHE 1
? VI MAX
1- 1438379 C Chic-Acs
CA 022110150091
TI Therapily degradable light-sensitive material
PA Fuji Photo Film Co., Ltd
SU Safa H. Seki 7/12/72, 1557654, JA
SD 874-9
TT - 02-73-5, 4154-07-4, 18512-85-9, 24800 02-C, 25003 02-5, 679
63-79-5,
47963-00-0, 47963-01-9, 47963-02-0, 73320-74-2, 73320-75-3,
73320-76-4, 73320-77-5, 73320-78-6, 73320-79-7, 73320-80-8,
73320-81-9, 73320-82-0, 73320-83-1, 73320-84-2, 73320-85-3,
73320-86-4, 73320-87-5, 73320-88-6, 73320-89-7, 73320-90-8,
73320-91-9, 73320-92-0, 73320-93-1, 73320-94-2, 73320-95-3,
73320-96-4, 73320-97-5, 73320-98-6, 73320-99-7, (color toner(s)), for pho
tographic
Material(21)
- ketones,ones and misc,di-, (benzothiazine derivs, (1), (2))
- Photothorgraphy, (photosensitive compns for, contg
benzothiazine)one color toner(s)
-2 764884 C Chic-Acs
CA 009120171869
TI Heat-developable light-sensitive material
PA Fuji Photo Film Co., Ltd
SU Safa H. Seki 8/2/72, 275431, JA
SD 874-9
TT - 147-93-3, (reaction effil, with Et chloroformate)
- 141-41-3, (1), with thioalicylic acid)
- 2489-05-6, (photothorog copying compns contg, benzothiaz
ine)one
der(s) toner(s) for)
18512-85-9, 47963-70-5, 47963-00-0, 47963-01-9, 47963-02-0,
(photothorog materials contg, as toner)
- Photographic emulsions,heat-developable, (contg benzothiaz
ine)one
der(s) toner(s)
- Photothorgraphy, (photosensitive materials for, (2))
COMMANDE, OU ETAPE DE RECHERCHE 3
? ST BA

```

-OU-(AN,GR,LI,AT,SL,CH,VE) ?FI  
AUTRES PRECISIONS (O/N) ?N



```

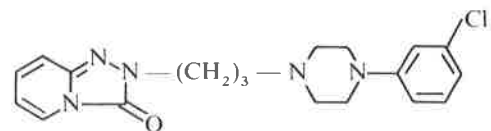
-ST-(QU,RE,AA) ?RE
RE RE 04
RESULTAT 396
PROCHAINE LISTE 1500
CONTINUER (O/N)?N
RESULTAT 396
(0 RE 00 NBRE DE REPONSE(S) 396 00
-ST-(QU,RE,AA) ?AA
15 REPONSE(S) SUR 100, ESTIMATION FINALE 1000
20 REPONSE(S) SUR 200, ESTIMATION FINALE 1400
29 REPONSE(S) SUR 300, ESTIMATION FINALE 1800
(0 AA 00 NBRE DE REPONSE(S) 39 00

```

C.N.I.C.-----S Y S T E M E D A R C-----C.N.I.C. / EURECAS  
REPONSE 1 12 COMPOSE NUMERO 1187286 RN ; 19794-93-5  
1 FRAGMENT 26 ATOMES

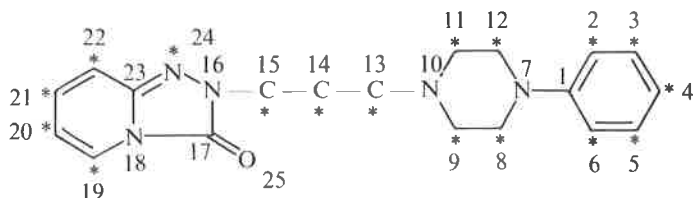
### Exemple d'une recherche bibliographique à l'aide du système DARC et du logiciel QUESTEL :

La trazodone est un psychotrope commercialisé sous différents noms et différentes formes. La nomenclature CAS de la trazodone dans le 9<sup>e</sup> Index Cumulatif est assez compliquée : 1,2,4-triazolo (4,3-a) pyridine-3(2H)-one, 2-(3)-[4-(3-chlorophenyl)-1-piperaziny] propyl.



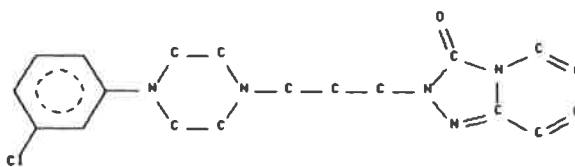
Parmi les composés dont la structure moléculaire est proche de celle de la trazodone : quels sont ceux qui ont fait l'objet de travaux et de publications faisant état de propriétés de tranquillisants et neuroleptiques ?

Afin que cette recherche soit exhaustive, la structure est décrite en permettant de nombreuses substitutions :



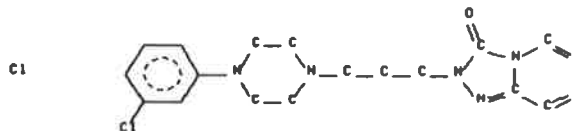
La recherche structurale sur le fichier EURECAS aboutit à 39 réponses (39 RN) dont on peut visualiser les structures. La durée de cette recherche comprenant la formulation de la question et l'édition en ligne de quelques structures est inférieure à 18-minutes.

La liste des RN retrouvés est transférée dans le fichier bibliographique EUCAS 77. Il s'avère que sur 39 composés, 34 ne sont pas cités dans cette portion du fichier, et donc ont été cités avant 1977. Les 5 composés répertoriés depuis 1977 sont cités dans 76 références. En combinant cette liste avec les mots-clés ANTIDEPRESSANTS ou PSYCHOTROPICS, on retrouve



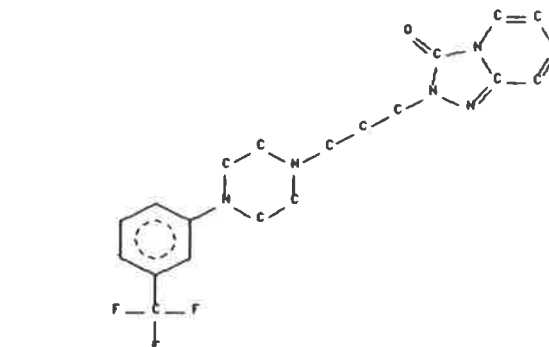
FORM. MOL. :C19H22ClN5O  
?S

C.N.I.C.-----S Y S T E M E D A R C-----C.N.I.C. / EURECAS  
REPONSE 1 16 COMPOSE NUMERO 1486320 RN ; 26332-39-2  
2 FRAGMENTS 27 ATOMES



FORM. MOL. :C19H22ClN5O.C1H  
?S

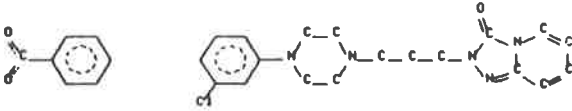
C.N.I.C.-----S Y S T E M E D A R C-----C.N.I.C. / EURECAS  
REPONSE 1 30 COMPOSE NUMERO 1644004 RN ; 70056-43-6  
1 FRAGMENT 29 ATOMES



FORM. MOL. :C20H22F3N5O  
?

27 références de travaux publiés citant les composés de la famille de la trazodone.

```
-ST-(QU,RE,AA) 7MI
88 MI 88
* L 6 8 39 STRUCTURE(S) 39 RN
-ST-(QU,RE,AA) 7VI
88 VI 88
C.N.I.C.-----S Y S T E M E D A R C-----C.N.I.C. / EURECAS
REPONSE 1 11 COMPOSE NUMERO : 1871872 RN ; 19778-38-4
2 FRAGMENTS 36 ATOMES
```



```
BASE CONNECTEE: EUCAS77
COMMANDE, OU ETAPE DE RECHERCHE 1
?..PP DARC L6
34 RN ABSENTS * EDITION (0/N) ?
?N
*1* RESULTAT 76
COMMANDE, OU ETAPE DE RECHERCHE 2
?1 ET (ANTIDEPRESSANTS OU PSYCHOTROPICS)
*2* RESULTAT 27
COMMANDE, OU ETAPE DE RECHERCHE 3
?..VI MAX
-1- 2184238 C.Cnic-Acs
CA : 095(25)2149B1
TI : Effect of antidepressants from various groups on cholinergic
structures
of the brain.
AU : Mashkovskii M.D.; Roshchina L.F.
SO : Zh. Nevropatol. Psikhiatr. im. S. S. Korsakova (ZNPiAP,004445
88); 81;
VB1(7); P.1047-51; (in RS)
CC : 81-5
IT : - 50-47-5; 50-48-6; 50-49-7; 51-12-7; 315-80-0; 1668-19-5; 10
065-57-3;
10347-81-6; 16154-78-2; 18464-39-6; 19794-93-5; 24219-97-4;
24526-64-5; 24667-93-4; 27312-93-2; 46817-91-8; 53734-79-5;
(anticholinergic activity of(1), antidepressant activity in r
elation
to)
- Antidepressants; (11)
- Brain; (cholinergic structures of, antidepressants effect o
n)
```

L'analyse des références permet de constater que seuls la trazodone base, son chlorhydrate et un dérivé fluoré sont connus pour leurs propriétés neuroleptiques et que 6 brevets ont été cités par CAS.

Le suivi d'une question peut être effectué sur les mises à jour des fichiers bibliographiques (bi-mensuelles) et structuraux (EPCAS, mensuel).

Ces exemples donnent une idée des possibilités offertes d'une part par le système DARC, et d'autre part, par la combinaison d'une recherche structurale et d'une recherche textuelle. Cette nouvelle approche permet d'obtenir des résultats complets et précis avec une excellente efficacité.

# MESUCORA PHYSIQUE

# 82

## EXPOSITION INTERNATIONALE DE LA RECHERCHE A L'AUTOMATISME



**6-11 DECEMBRE 1982**  
**PARIS · PORTE DE VERSAILLES · FRANCE**

Même lieu, mêmes dates :



Exposition Internationale de l'Équipement Électrique



Commissariat Général SEPIC  
40, rue du Colisée, 75381 Paris Cedex 08  
Tél. : (1) 359.10.30 - Télex 640 450 F SEPIC