L'information bibliographique en chimie : une nouvelle étape franchie par le CNIC

par A. Déroulède *(Directeur du CNIC, 26 rue Boyer, 75971 Paris Cedex 20).*

Le Centre National de l'Information Chimique (CNIC) met à la disposition de la communauté des chimistes un nouvel outil qui va, d'une part, modifier considérablement les méthodes d'exploitation de l'information bibliographique en chimie et, d'autre part, aider les chercheurs à établir des corrélations entre les structures et les activités des composés chimiques. Ce nouvel outil est le système DARC qui est maintenant adapté aux bases de données de Chemical Abstracts Service. Une telle réalisation est un exploit technique remarquable en soi, qui annonce une ère nouvelle pour la documentation et pour les spécialistes de la chimie fine.

Chaque chimiste est perpétuellement confronté au problème de l'information bibliographique, que ce soit pour se tenir au courant des publications relatives à son domaine de recherche ou que ce soit pour faire le point sur un sujet avant de se lancer dans de nouvelles recherches. Dans le secteur de la chimie, le nombre de revues et d'articles publiés chaque année à travers le monde est tel que les méthodes classiques de la documentation ne suffisent plus. Heureusement, l'informatique avec ses capacités de traitement, de tri et de stockage a permis d'exploiter les fonds bibliographiques sur ordinateurs. Les possibilités de télétraitement, le développement des réseaux de télécommunication interconnectés, les transmissions de données par paquets, donnent les moyens maintenant d'interroger directement « en ligne » ces bases de données à partir d'un terminal relié par une ligne téléphonique à ces ordinateurs appelés « serveurs ». Cependant, pour accéder « en ligne » aux bases de données, il est nécessaire de connaître le logiciel d'interrogation qui diffère d'un serveur à un autre, et les caractéristiques propres de chacune des bases de données.
Et chaque année de nouveaux serveurs sont lancés, le nombre de logiciels augmente et les procédures d'interrogation restent relativement compliquées. C'est pour cela que plus de 95 % des recherches bibliographiques en ligne sont effectuées par des spécialistes de l'information. En chimie, l'interrogation en conversationnel des bases de données bibliographiques est d'autant plus compliquée que la nomenclature chimique est souvent ambiguë, que de nombreux composés chimiques sont cités sous différents noms (quelques substances dans le fonds bibliographique de Chemical Abstracts Service sont citées sous plus de 150 noms dans les différentes publications !) et les règles d'indexation sont complexes.

Le CNIC propose maintenant une méthode originale et efficace pour exploiter l'information bibliographique en chimie. En effet, il ne s'agit pas pour le CNIC de mettre à la disposition du public simplement les mêmes services que ceux offerts par des serveurs étrangers, mais bien d'offrir de meilleurs moyens d'accès à l'information et de faciliter effectivement la tâche des chimistes. Actuellement, le CNIC est le premier organisme à donner accès aux bases de données CAS en utilisant le langage naturel du chimiste, c'est-à-dire en lui permettant de formuler sa question en dessinant la formule développée d'un composé chimique, ou même en décrivant un élément structural caractéristique d'une famille de composés, démarche appelée « recherche par sous-structure ». Ceci est possible grâce au système DARC, inventé par le Professeur J. E. Dubois et qui a été adapté aux bases de données CAS par l'Association pour la Recherche et le Développement en Informatique Chimique (ARDIC). La recherche bibliographique est poursuivie en précisant la question avec des éléments textuels à l'aide du logiciel MISTRAL qui est le langage adopté sur le Centre Serveur Français QUESTEL.

L'interrogation des bases de données CAS à l'aide du système DARC

Les premières bases de données bibliographiques sur lesquelles le système DARC a été adapté proviennent de Chemical Abstracts Service. Ce choix a été fait pour deux raisons. D'abord le fonds bibliographique de Chemical Abstracts est certainement le plus complet en ce qui concerne la chimie, aussi bien fondamentale qu'appliquée. De plus, Chemical Abstracts a constitué un fichier dictionnaire, le « Chemical Registry System », qui contient près de 5 millions de composés répertoriés et qui s'enrichit de 350 000 nouveaux chaque année. A chaque composé est affecté un numéro de registre (Registry Number). Ce fichier dictionnaire contient également la description topologique de la structure des composés.

Le système DARC est un ensemble de logiciels qui permet de retrouver dans ce fichier dictionnaire un composé avec son numéro de registre à partir de la description de sa structure et mieux encore de reconnaître dans ce fichier dictionnaire tous les composés contenant un élément structural donné, de retrouver leur numéro de registre respectif et de restituer les formules développées correspondant à chacun d'eux.

L'interrogation par composés chimiques des bases de données Chemical Abstracts à l'aide du système DARC est considérablement simplifiée. Il suffit à l'utilisateur de décrire la formule développée de la molécule, ou seulement une sous-structure caractéristique d'une famille de composés, très facilement, sans avoir recours à aucun code. Cette description peut être faite par voie graphique ou alphanumérique. Les atomes sont numérotés dans un ordre quelconque pour décrire le graphe lui-même. La nature des atomes ainsi que celle des liaisons est indiquée, mais le système donne la possibilité de considérer différents atomes dans un même site, différents substituants, et autorise aussi à laisser indéterminée la nature des liaisons. Ainsi, un utilisateur ne limite pas sa recherche à un seul composé.

La recherche est conduite automatiquement, l'utilisateur n'intervenant que pour préciser sa question. Les opérations sont simplifiées du fait qu'un « menu » est affiché et que l'utilisateur est guidé dans chacune de ses commandes.

La recherche par sous-structures aboutit à une liste de numéros de registre. Cette liste est mise en mémoire et la structure des composés retrouvés peut être visualisée sur l'écran d'un terminal graphique. Suivant cette méthode, on peut retrouver tous les composés répertoriés dans la base de données CAS ayant une structure donnée. C'est à partir de cette liste des numéros de registre que l'on procède à la recherche bibliographique proprement dite.

La recherche bibliographique avec le logiciel MISTRAL

La recherche des références bibliographiques citant des composés ainsi retrouvés se fait à l'aide du logiciel MISTRAL. Une seule commande suffit pour introduire tous les numéros de registre. La nomenclature correspondant à chaque numéro de registre peut être éditée. En ce qui concerne la recherche bibliographique, l'utilisateur peut préciser sa question avec des éléments textuels tels que des mots-clés, des noms d'auteurs, etc. L'intersection de la recherche structurale et de la recherche textuelle aboutit à la sélection rapide de toutes les références des publications répondant à la question. Le fait de pouvoir formuler une question par des sous-structures et de travailler d'emblée sur une famille de composés a l'avantage d'aboutir très rapidement à des résultats complets sans avoir à formuler la question composé par composé. Le passage automatique du système DARC au logiciel MISTRAL est un atout considérable, car il permet un gain de temps important dans une recherche en conversationnel et aussi de travailler sur un grand nombre de composés.

Les références bibliographiques retrouvées sont éditées sous différents formats. Elles peuvent comporter les titres, les noms d'auteurs, les mots-clés, les termes d'indexation, et, dans le cas de la base Chemical-Biological Activities (CBAC), les résumés.
Il est également possible de procéder à des recherches bibliographiques directement par le texte. Le logiciel MISTRAL a l'avantage de comprendre les instructions aussi bien en français qu'en anglais. De plus, il peut être pratique de lancer une recherche bibliographique à l'aide d'éléments textuels chaque fois qu'une question ne peut pas être posée par des noms de composés chimiques, et aussi pour traiter des questions portant sur des concepts généraux (comme par exemple : estérification des acides gras, appareillage en chromatographie liquide, etc).

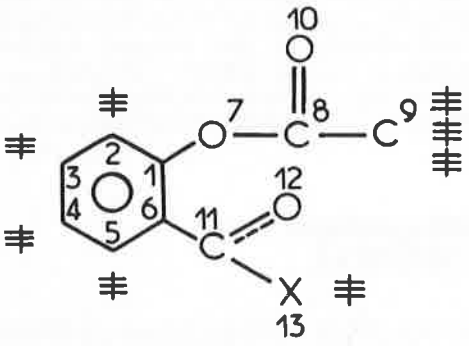
On voit donc que les possibilités de formuler une question, soit par les structures, soit par le texte, et de pouvoir combiner une recherche par sous-structure avec une recherche textuelle, constituent des avantages majeurs. Quelques exemples permettent de mieux réaliser ces avantages :

Exemple 1

Existe-t-il dans la base CBAC des références concernant des publications sur les propriétés des carotènes comme vitamines ?
a) Par le texte, on retrouve 672 références citant les termes contenant la racine « CAROTEN + » et en les combinant avec les termes vitamines ou provitamines, on obtient 84 références.
b) En introduisant la question par le système DARC, à partir de la sous-structure des carotènes, on obtient 122 composés appartenant à la famille des carotènes et qui sont cités dans 1 103 références bibliographiques. En combinant avec le terme vitamine, on obtient 582 références.

Exemple 2

Dans certains cas, la question ne peut pas être formulée par le texte. Ainsi, lorsque la nature de certains atomes et de plusieurs liaisons est laissée indéterminée, que la nature et le nombre de substituants sont variables, etc. Par exemple, la recherche des composés contenant la sous-structure dessinée ci-dessous. Cette formule correspond à une famille de composés pouvant avoir différents substituants sur les atomes 2, 3, 4, 5 et 9, X désignant un atome quelconque admettant aussi des substituants. Dans la base CBAC, parmi 420 000 composés, on en retrouve 301 contenant cette sous-structure.



Ainsi, on constate que la formulation d'une question par les sous-structures aboutit à des résultats plus complets que lorsque la question est définie uniquement avec des éléments textuels. D'autre part, le fait de pouvoir travailler sur des familles de composés change considérablement les méthodes d'exploitation de l'information en chimie. Les chercheurs peuvent avoir plus facilement une vue d'ensemble sur des résultats publiés sur de nombreux composés et, de là, établir des corrélations entre les structures et les activités.

De même, les documentalistes chargés de retrouver des publications sur des brevets connaissent bien l'importance de l'exhaustivité dans une recherche d'antériorité de brevets, et apprécient cette nouvelle possibilité de recherche par sous-structure sans avoir recours à des codes compliqués, ou à utiliser la nomenclature qui nécessite une préparation longue et minutieuse avec de nombreux risques d'erreurs et d'omissions ou encore à reprendre la même stratégie de recherche d'un composé à un autre.

Enfin, la visualisation sur l'écran d'un terminal graphique ou l'édition sur une table traçante des structures moléculaires des composés retrouvés apporte un avantage supplémentaire. Une telle présentation des résultats est très stimulante car l'utilisateur instantanément peut retrouver des composés dont il n'imaginait pas l'existence au départ de la recherche par sous-structure.

L'accès aux bases de données proposées par le CNIC

Les bases de données, dont le CNIC a la responsabilité, sont chargées sur le serveur QUESTEL. Pour pouvoir interroger ce serveur il faut avoir un numéro d'accès qui correspond à un contrat stipulant des conditions d'utilisation de ce type de service. De plus, l'utilisateur doit disposer d'un terminal alphanumérique ou graphique qui est connecté au serveur en France par le réseau TRANSPAC, en Europe par EURONET et depuis les États-Unis par Tymnet et Télénet.

Aussi simple que puisse être l'utilisation du système DARC, il n'en reste pas moins vrai que c'est un outil très complexe et très puissant. Afin que les utilisateurs sachent exploiter toutes les possibilités apportées à la fois par DARC, MISTRAL et par la combinaison des deux, il est demandé aux utilisateurs de suivre une session de formation de deux jours, organisée au CNIC ou dans le cadre des entreprises et centres de recherches. Au cours de ces stages les utilisateurs apprennent à se familiariser avec les logiciels d'interrogation, les caractéristiques des bases bibliographiques, les stratégies de recherche (préparation des questions et leur traitement au terminal). Après ces sessions, les utilisateurs peuvent faire appel à une assistance technique assurée par les ingénieurs du CNIC. Le temps moyen pour « traiter » une question, en mode interactif, est de l'ordre de 20 minutes. Les chimistes et les documentalistes qui connaissent les temps nécessaires pour effectuer une recherche qui veut être exhaustive apprécient vite la valeur d'un tel service. Il est important de réaliser que ce nouveau système d'interrogation contenant la recherche par structures et la recherche textuelle avec passage automatique de l'une à l'autre constitue un exploit technique remarquable réalisé pour la première fois sur des fichiers bibliographiques d'un tel volume. Des présentations ont été faites en France, en Angleterre, en République Fédérale Allemande, en Hollande, en Espagne, au Portugal et dans plusieurs villes des États-Unis d'Amérique. Les réactions sont très encourageantes. Il faut que la communauté des chimistes en France profite de la chance d'avoir à sa disposition un tel outil, car les premiers à l'utiliser seront certainement avantagés. Actuellement un effort est fait pour développer l'innovation dans tous les domaines très particulièrement en haute technologie et également dans le domaine de la chimie fine. Les services proposés par le CNIC peuvent et doivent contribuer heureusement dans ce sens.