Titane : serveur national des bases de données en chimie

André Baldy* *ingénieur de recherche hors classe*

Summary : ***Titane : national chemistry databases server***

The national chemical databases server is the result of a cooperation between the CNRS and the Ministry. 12 millions molecules, 2 300 users under contract in the academic wold. Help, training, help for new databases development, are the tools given to chemists.

Mots clés : ***Bases de données en chimie.***
Key-words : *Chemical databases.*

Partenariat stratégique

Le Centre d'Ingénierie et de Modélisation Moléculaire (CIMM) a deux pôles essentiels d'activité : la modélisation moléculaire et les bases de données en chimie. C'est ce deuxième aspect que nous nous proposons de développer.

Le CIMM entend favoriser le développement de l'utilisation des grandes bases de données chimiques dans les laboratoires, et souhaite apporter un appui logistique aux équipes voulant développer ou valoriser leurs propres bases de données.

Cette idée découle de la constatation que la recherche française en chimie dépend, de façon cruciale, de l'utilisation des grandes bases de données (de structures, de réactions, de bibliographies...). Il était donc primordial d'intégrer ces bases de données dans l'environnement de travail du chercheur scientifique.

En mai 1994, au sein du GDR 1093 « Traitement informatique de la connaissance en chimie organique », était prise l'initiative d'une enquête (réalisée en collaboration avec Françoise Girard de Polytechnique, Claude Laurenço et Jacques Coste de Montpellier) auprès de laboratoires intéressés par la synthèse organique, afin d'évaluer l'intérêt pour les systèmes d'information moderne en chimie.

Cette initiative s'intégrait au plan d'action 1994-1996 du département des Sciences chimiques du CNRS.

- Accord CNRS-MDL, en septembre 1995, reconduit en 1996 et conforté par le ministère en 1997, 1998 et 1999.
- Accord ministère-CNRS-Beilstein, en novembre 1996, reconduit en 1997, 1998 et 1999.
- Accord ministère-CNRS-Cambridge, en avril 1998, stipulant que le CIMM devient le centre national affilié et assume la gestion complète de ces bases gérées par le département des Sciences chimiques du CNRS depuis 1994, reconduit en 1999.

Vitalité du système

Un serveur national, de type G30 IBM, a été installé à Marseille en décembre 1995 et a subi, depuis, les extensions nécessaires à l'exploitation et à la gestion de ces bases. Deux stations RISC 6000 assurent les fonctions de sécurité Fire-Wall et de serveur de listes.

En ce début d'année, un nouveau serveur vient conforter ce complexe.

Pour assurer à l'information le degré de sécurité voulu, il faut protéger les systèmes utilisés contre les manœuvres captatoires ou les intrusions susceptibles de la trahir, de l'altérer ou de la détruire. Nous y veillons et avons demandé à chaque utilisateur de signer une charte de sécurité.

Un poste d'ingénieur d'étude CNRS et un poste de technicien de l'enseignement supérieur viennent, tout récemment, d'assurer la pérennité de l'opération.

Actuellement, plus de 2 300 accès aux différentes bases sont ouverts sous contrat, au travers de près de 1 400 logins, douze millions de molécules sont accessibles, plus de mille personnes sur 29 villes ont reçu une formation.

Les formations académiques et industrielles sont possibles au CIMM ou sur site, en premier niveau ou en niveau avancé.

Ces actions de formation faisaient partie des priorités fortes du département, et se plaçaient au cœur de son plan de formation. En juin 1997, le Bureau national de formation du CNRS nous accordait un label national avec mise en œuvre déconcentrée en région Provence.

Ce transfert des connaissances, par les personnes formées, est assuré suivant le principe de la démultiplication de la formation, ou essaimage.

* Faculté des sciences, ESA Q6009/CIMM case D62, avenue Normandie Niemen, 13397 Marseille Cedex 20. Tél./Fax : 04.91.28.81.93.
E-mail : baldy@titane.u-3mrs.fr

Bases de données disponibles

Des journées d'informations et de négociations nous ont conduits chronologiquement à mettre à la disposition des chercheurs académiques français les bases suivantes, regroupées en quatre contrats selon leurs distributeurs :

1. Bases de MDL : bases spécialisées réactionnelles et moléculaires.
2. Chirbase : aide à l'optimisation des séparations chirales par chromatographie.
3. Bases de Cambridge : base de structures 3D.
4. Bases de Beilstein : bases à vocation exhaustive de structures, propriétés chimiques et physiques, et de réactions.

Toutes ces bases possèdent une interface graphique et sont interrogeables par mots clefs et par structure chimique ainsi que par sous-structure.

MDL

Ces bases peuvent être regroupées en deux catégories : les bases réactionnelles, interrogeables ensemble sous une même interface comme une base unique, et les bases moléculaires thématiques.

Les bases réactionnelles

Cette base est constituée de différentes composantes dont *Cheminform, Theilheimer, Current Litterature File, Core, Chiras, Metalysis, Comprehensive Heterocyclic Chemistry* (version électronique des manuels *Comprehensive Heterocyclic Chemistry* publiés en 1984), *OrgSyn* (version électronique de la revue *Organic Synthesis*), *Reaccs-JSM* (version électronique de la revue *Journal of Synthetic Methods*) et *Spore* (réactions faisant intervenir une phase solide). Cette base contient plus de 800 000 réactions chimiques filtrées novatrices ou sélectives. Il s'agit d'une base réactionnelle de type exemplaire : elle a pour vocation de rassembler un exemple de chaque type de réaction. Elle est donc très utile pour des études de réactivité : rechercher l'effet du changement des conditions d'un type de réaction sur le rendement, rechercher l'effet des substituants sur la réactivité d'une famille de composés, l'utilisation de groupements protecteurs...

Les bases moléculaires

- *La base NCI* du National Cancer Institute contient les molécules dont l'activité cancérogène a été testée.
- *Avalaible Chemical Directory* est le catalogue électronique des produits commerciaux. Cette base permet de retrouver les prix et les adresses des fournisseurs pour plus de 120 000 molécules. Elle fournit aussi une structure 3D pour la plupart des produits commerciaux.
- *MDL Drug Data Report* (version électronique du Drug Data Report) et *Comprehensive Medicinal Chemistry* (version électronique des manuels Comprehensive Medicinal Chemistry), contenant des informations de type structure-activité sur plus de 30 000 molécules ; la structures tridimensionnelle des molécules est incluse. Cette base est dédiée aux molécules actuellement en cours d'études et dont l'activité biologique n'est pas encore déterminée de manière certaine, ainsi que les molécules déjà commercialisées.
- *Metabolite* contient les réactions métaboliques publiées. Il s'agit principalement des informations sur des substances employées en médecine, mais aussi en agriculture, industrie chimique, et des contaminants de l'environnement.

ENSSPICAM Chirbase Office

- *ChirBase :* cette base utilise l'interface ISIS de MDL et est un complément à ce groupe de bases. Elle contient les caractéristiques de colonnes chromatographiques chirales, et permet d'optimiser la séparation d'énantiomères chiraux par chromatographie.

Cambridge

- *Cambridge Structural Database :* structures cristallines de plus de 180 000 substances organiques. Toutes les structures ont été déterminées par des techniques de diffraction de neutrons ou de rayons X. Chaque entrée comprend les références bibliographiques et les informations expérimentales de la structure du cristal, la structure 2D et 3D de la molécule, et leur organisation dans la maille cristalline.
- *Protein Data Bank :* archive de structures tridimensionnelles de macromolécules biologiques déterminées expérimentalement.
- *Interface Iso Star :* informations sur les interactions préférentielles de centaines de groupements chimiques. Cette information a été obtenue d'après les interactions dans les cristaux de petites molécules (base CSD), les interactions entre des protéines et de petits ligands (base PDB), et les interactions de paires de molécules en phase gazeuses (données sur les orbitales moléculaires).

Beilstein

- *Cross Fire Plus Abstracts :* la plus grande base de données de propriétés chimiques. Elle contient la structure, les propriétés physiques et chimiques répertoriées et les références bibliographiques associées à ces propriétés pour plus de 7 millions de molécules, ainsi que plus de 10 millions de réactions organiques. Cette base à vocation exhaustive indexe toutes les données publiées dans les principaux journaux de chimie organique. Etant donné son étendue, elle se prête mieux à des recherches assez précises : retrouver rapidement les propriétés connues d'une molécule, rechercher des molécules répondant à des critères sur leur réactivité ou leurs propriétés physique ou chimique, retrouver les conditions d'une réaction assez précise ou les réactions de préparation d'un composé donné. Elle peut également être très utile pour commencer une bibliographie.
- *Gmelin* : 1 million de composés organométalliques et inorganiques.

Résultats d'enquête

Une enquête diffusée à l'ensemble des utilisateurs (au travers de près de 1 400 logins) nous permet, par l'étude des 222 réponses reçues, de dégager les points forts suivants :

1. Répartition de l'utilisation des bases.
2. Détail de l'utilisation des autres bases.
3. Informations recherchées.
4. Répartition des souhaits de nouvelles bases.
5. Bases de données utilisées pour l'enseignement.

Bien que n'ayant pas CAS sur notre serveur, il nous a semblé utile, à titre comparatif, de l'inclure dans notre enquête. Les autres bases et serveurs cités sont les

seuls qui ont été mentionnés par nos utilisateurs. Ces résultats indiquent donc les bases qu'utilisent ou ont utilisées les laboratoires abonnés au serveur Titane.

Répartition de l'utilisation des bases

La base Beilstein est largement la plus utilisée (*figure 1*) car elle est, avec CAS, la plus polyvalente. Elle est consultée aussi bien pour des propriétés de molécules, pour une ébauche de bibliographie, que pour l'aide à l'élaboration d'une synthèse.

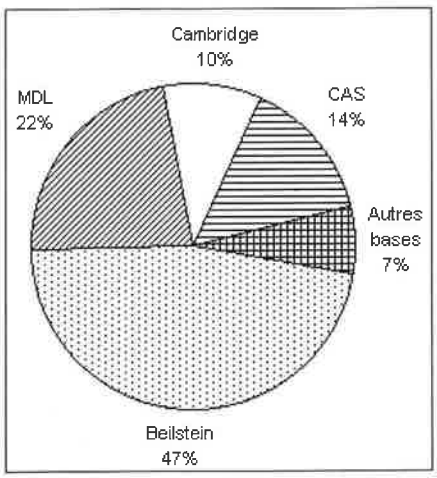

***Figure 1** - Répartition de l'utilisation des bases.*

Le handicap de la base CAS est le mode de facturation des serveurs qui l'abritent : selon le temps de consultation. Il est difficile de l'inclure dans le budget prévisionnel d'un laboratoire, et l'accès ne peut être laissé à la libre disposition sous peine de surcoûts importants. En effet, même avec les réductions académiques, une recherche de réaction par structure ou sous-structure dans CASREACT coûte rarement moins de 300 francs. Or, il s'agit du mode de recherche le plus utilisé par nos abonnés. Le forfait pour la consultation du Beilstein, pour quatre postes de consultation et pour une université, se monte en 1999 à 16 000 F HT.

Les bases spécifiques s'adressent à un nombre plus restreint d'utilisateurs, correspondant à leur spécialité, ce qui explique qu'elles soient consultées moins régulièrement. Mais elles n'en sont pas moins capitales, rassemblant en une seule source, pratique à manipuler, une grande quantité d'informations sur un domaine précis. D'ailleurs, la demande en nouvelles bases de données concerne surtout des bases spécialisées.

Détail de l'utilisation des autres bases

Les bases consultées en plus de Titane (*figure 2*) sont principalement bibliographiques (CAS 33 % + Current Contents 20 % afin de compléter une bibliographie) ou biochimiques. Il est vrai que ce domaine est peu couvert par les bases servies par Titane.

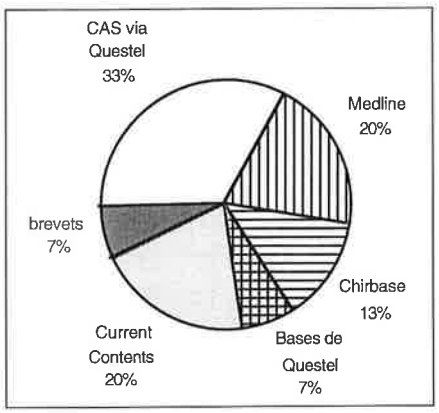

***Figure 2** - Détail de l'utilisation des autres bases.*

On peut noter la présence à ce titre de Medline, principale base biochimique disponible gratuitement sur un serveur aux États-Unis, malgré une localisation posant de sérieux problèmes d'encombrement réseau. Les Currents Contents ont également été fréquemment cités comme sources d'information. Permettant un suivi de la littérature, ils sont un complément utile aux bases bibliographiques.

En outre, on peut noter que les bases citées sont, à part CAS, gratuites ou à paiement forfaitaire ; et l'utilisation de CAS via Questel est le résultat de réductions avantageuses de la part de ce serveur. Il semble donc bien que le problème du financement soit pour les laboratoires le facteur déterminant dans le choix des sources d'information.

Informations recherchées

La vocation première de chaque type de bases se retrouve dans la façon dont elles sont utilisées (*figure 3*). Les recherches dans CAS se font avant tout par auteur, et par molécule. Mais le champ d'investigation s'étend à toutes les propriétés. Les recherches structurales, beaucoup plus chères, restent minoritaires.

Beilstein est également utilisé comme base bibliographique, avec des recherches par auteurs, molécules ayant des propriétés particulières, réactions. Son principal mode d'interrogation reste, de loin, la molécule. Il est vrai que c'est le mode d'interrogation le plus riche. Il permet, soit de rechercher les propriétés connues d'une molécule dont sa réactivité, soit de rechercher et d'évaluer les voies de synthèses qui peuvent conduire à cette molécule.

Quand aux bases MDL, ce sont la base ACD, permettant de rechercher les produits commerciaux, et l'ensemble des bases réactionnelles qui sont les plus consultées. Malgré le fait que les bases réactionnelles soient de type exemplaire et non exhaustive, 47 % des utilisateurs font régulièrement des recherches de réaction exacte. Ce mode de recherche n'est pourtant pas le plus approprié pour

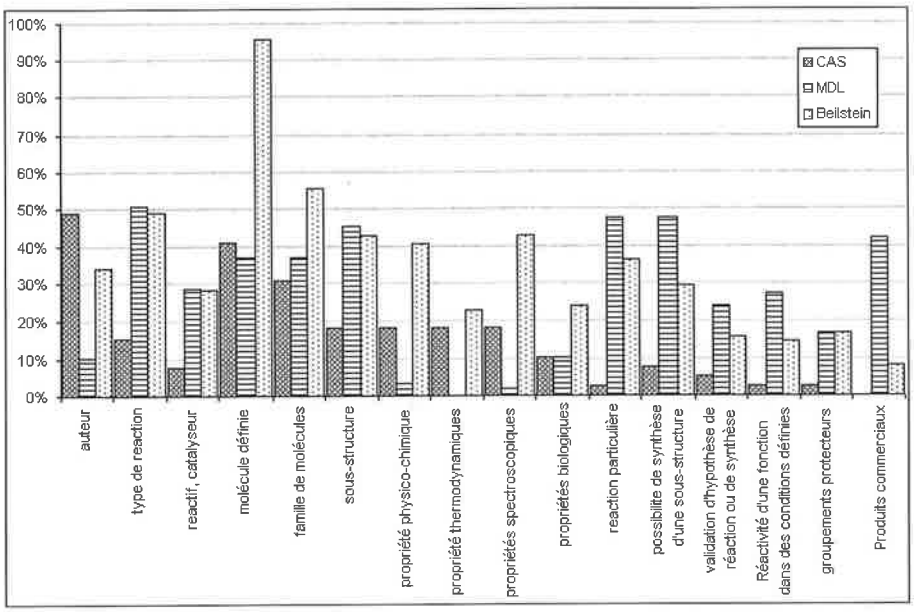

***Figure 3** - Informations recherchées dans les principales bases concernant la chimie organique.*

ce type de bases. En effet, ces bases sont essentiellement adaptées à l'étude de la réactivité des fonctions. Dans cette optique, il est plus intéressant de pouvoir comparer un nombre raisonnable de réactions très différentes, couvrant un large domaine de conditions expérimentales. Il sera ainsi plus aisé d'étudier l'effet des substituants et des conditions opératoires sur la réactivité. Par opposition, les bases dites exhaustives ont pour vocation de regrouper toutes les informations disponibles dans la littérature dans le domaine qu'elles couvrent.

Répartition des souhaits de nouvelles bases

Les souhaits concernant de nouvelles bases reflètent la diversité des domaines d'activité des laboratoires (*figure 4*). On note une forte préférence pour les bases dans le domaine de la synthèse organique, mais les deux bases les plus citées sont aussi les seules bases pour lesquelles une documentation était accessible. La plupart des utilisateurs ne connaissent pas d'autres bases de données en dehors de celles qu'ils consultent sur Titane ou les bases bibliographiques les plus répandues dans les bibliothèques (CAS, Current Contents, Science Citation Index...). Nous avons vu, au début, que 7 % des utilisateurs seulement consultaient d'autres bases, essentiellement CAS, Current Contents ou MedLine. Il est probable qu'une meilleure appréciation des besoins des laboratoires serait obtenue en leur proposant une documentation sur une large collection de bases spécialisées.

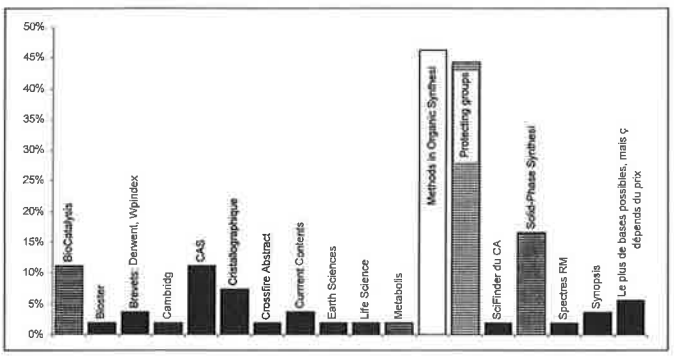

Figure 4 - Répartition des souhaits de nouvelles bases.

Bases de données utilisées pour l'enseignement

48 % des utilisateurs ayant répondu à cette enquête se déclarent concernés par l'enseignement. Les statistiques suivantes portent sur ce sous-ensemble :

- Il semblerait qu'une forte proportion des utilisateurs enseignants utilise les bases de données de MDL, et surtout Beilstein (*figure 5*). Les bases de données peuvent être l'objet même de l'enseignement (nous pouvons accorder ponctuellement des postes supplémentaires), ou tout simplement l'outil pour la préparation de cours ou d'exercices.

- Ce sont essentiellement les filières universitaires qui bénéficient de ces outils.

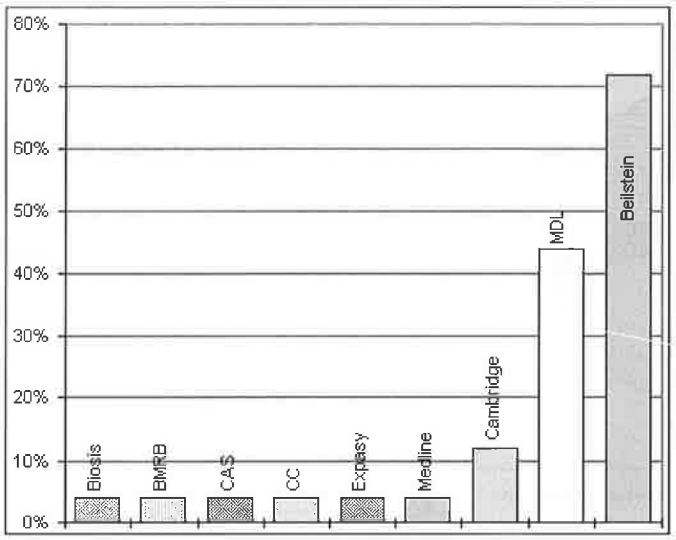

***Figure 5** - Bases de données utilisées pour l'enseignement.*

Conclusion

Afin de ne pas alourdir cet article, nous passerons sous silence toutes les conditions techniques d'exploitation mais, nous pouvons signaler que le nombre total de connections significatives (connexions sur le serveur ayant été suivie d'une ou plusieurs requêtes) sur MDL pour 1997 a été de 17 902 et de 42 176 pour Beilstein. Soit un total de connections pour l'année dernière de 60 078. Sur les six premiers mois de 1998, un taux de croissance de 10 % est relevé pour MDL et un taux de 41 % pour Beilstein (suite logique des formations effectuées).

Actuellement :

MDL : 644 utilisateurs pour 115 postes attribués.

Beilstein : 1 162 utilisateurs pour 304 postes attribués.

Cambridge : 73 postes d'accès sur le serveur et 31 CD distribués.

Le surcoût sur la recherche est réel et lourd (voir les conditions d'abonnement à l'adresse « http://vulcain.u-3mrs.fr/stt », rubrique « souscription aux bases »), il est indispensable de l'inclure dans les prévisions au même titre que les autres postes dans les budgets prévisionnels des laboratoires.

Les conclusions que nous pouvons tirer de cette enquête sont les suivantes :

- En général, la recherche des molécules s'effectue sur CAS et Beilstein (93 % des cas). Beilstein étant utilisé pour la recherche des molécules mais aussi pour la recherche de réactions.

- Nous notons les demandes importantes sur les nouvelles bases suivantes :

. Pour Beilstein : Autonom, Abstract
. Protecting groups
. Mos
. Biocatalysis

- Une journée de réflexion, prévue pour la fin de l'année, devrait nous permettre de mieux situer les besoins exprimés. Des contacts sont en cours avec d'autres fournisseurs de bases.

Remerciements

Que Kirstin Ruiz, directeur technique, et Pierre Vatton, administrateur système depuis Grenoble, soient ici vivement remerciés pour leur efficacité compétente et créatrice.

Merci à Michel Arbelot pour son implication dans le développement de nouvelles bases de données.