

La documentation en génie chimique et l'emploi des bases de données automatisées*

par M. Planchon *, B. Thiriet **
et W. Wigman ***

(* Solvay et Cie, Bruxelles;
** A.F.D.A.C., Paris; *** Koninklijke Shell Laboratorium, Amsterdam)

Les bases de données commercialement accessibles sont évaluées en fonction de leur utilité et de leurs performances dans le domaine du génie chimique.

I. Introduction

Jusque très récemment, le besoin en information accessible par voie automatisée était à peine ressenti par les ingénieurs de génie chimique. Toutefois, cette situation évolue rapidement du fait du développement de systèmes et d'ordinateurs plus efficaces d'une part, ainsi que de la croissance formidable de l'information disponible d'autre part. Pour dominer ce flot d'information, des versions automatisées des plus importants services de résumés et signalements de la littérature ont été préparées par les éditeurs et sont vendues sous forme de bandes magnétiques; en même temps étaient créés des systèmes spécialement

axés sur la recherche de l'information textuelle.

Il est maintenant assez facile pour un ingénieur de trouver un centre de traitement de l'information qui puisse lui assurer par voie automatisée aussi bien des profils de diffusion sélective de l'information courante que des recherches rétrospectives. Quelques centres travaillent pour ce dernier type de service en mode séquentiel répondant à des questions préformulées, mais la majorité emploie maintenant des techniques de recherche interactives faisant appel à des réseaux de transmission et à des télétypes. Ces nouvelles techniques permettent un réel dialogue entre le spécialiste de l'information et l'ordinateur.

Dans la première partie de cet article, nous présenterons rapidement les principales bases de données automatisées relevant du génie chimique et analyserons leurs principaux intérêts sur la base des résultats du traitement de quelques questions tests.

II. Les bases de données existantes

L'ouvrage « Directory to machine-readable data bases and data banks in Europe »

Tableau I

Bases de données existantes

Base de données	Producteur	Date de départ	Nombre de documents par an
C.A.C.	Chemical Abstracts Service	1968	400 000
Compendex	Engineering Index	1969	95 000
Physics Abstracts	I.N.S.P.E.C.	1969	100 000
Computer and Control Abstracts	I.N.S.P.E.C.	1969	25 000
U.S. Government R & D Reports	N.T.I.S.	1961	60 000
A.P.I.-L.I.T.	American Petroleum Institute	1964	15 000
Chemical Market and Electronical Market Abstracts	Predicast	1972	26 000
Chemical Industry Notes	Am. Chem. Soc.	1972	40 000
Pollution Abstracts	Am. Chem. Soc.	1970	6 000
Mechanical Engineering (I.S.M.E.C.)	I.E.E.E. *	1968	40 000
Nuclear Science Abstracts	U.S. Atomic Energy Comm.	1967	90 000
Metals Abstracts (Metadex)	American Soc. of Metals	1966	30 000
S.T.A.R., I.A.A.	N.A.S.A.	1962	55 000
Petroleum Abstracts	University of Tulsa	1965	15 000
Pascal 880	C.N.R.S.	1973	15 000
Source and Citation Index	I.S.I.	1964	375 000
Verfahrenstechnische Berichte	Verlag Chemie	1966	8 000
Dechema	Dechema	1962	10 000
Rapra Abstracts	Rapra	1972	9 000
Central Patents Index	Derwent	1970	90 000
World Patents Index	Derwent	1974	50 000
Comprehensive Data Base Patents	I.F.I.	1950	25 000
A.P.I.-P.A.T.	American Petroleum Institute	1964	8 000

* I.E.E.E. : Institute of Electric and Electronic Engineers.

* Cet article a été préparé à la demande du Groupe de travail « Information et Documentation » de la Fédération Européenne de Génie Chimique.

(Ed. Gordon Pratt, ASLIB-EUSIDIC 1976) cite environ 400 bases de données parmi lesquelles environ 25 peuvent intéresser les spécialistes du génie chimique. Le tableau I analyse ces fonds documentaires en citant l'année de départ du fonds ainsi que le nombre de documents ajoutés annuellement au fichier. La plupart de ces fonds sont accessibles en conversationnel. De plus il existe pour certaines bases des sous-fichiers spécialisés contenant chacun un nombre plus restreint de références.

Chemical Abstracts qui analyse environ 400 000 références par an est le fonds documentaire le plus important pour le génie chimique. Les sections « Chimie appliquée » contiennent environ 70 000 références par an dont 20 % sont des brevets. D'autres références importantes sont classées dans les sections « Chimie physique ». Pour la recherche de propriétés physico-chimiques telles que points d'ébullition, de fusion, pressions de vapeur, densités, viscosités, conductivités, C.A. s'avère être un fonds particulièrement digne de confiance. A noter cependant que équipements et appareillages n'y sont mentionnés que dans la mesure où ils sont décrits en relation directe avec un procédé chimique particulier, ce qui implique une limitation de la couverture dans ce domaine. Les informations commerciales et technico-économiques ne sont pas reprises même si elles se rapportent à des composés chimiques définis.

Les termes sur lesquels la recherche peut être effectuée dans la version disponible sur bande magnétique, les Chemical Abstracts Condensates, sont : mots du titre, groupes de mots-clés ajoutés lors de l'indexation du document, auteurs, titulaires du brevet et autres données bibliographiques. Ce vocabulaire n'est pas contrôlé (langage naturel) ce qui oblige le spécialiste de l'information à être au courant de la terminologie des auteurs et des indexeurs de CAS lors de la formulation de la recherche.

Engineering Index, dont la version automatisée est appelée COMPENDEX, cite 95 000 nouvelles références par an provenant de 3 500 journaux. Le génie chimique concerne 15 % de ce fichier. Bien qu'il y ait un important recouvrement avec CA, Compendex constitue un assez bon outil supplémentaire pour la recherche d'information sur les aspects équipements et opérations unitaires. Compendex répond bien aux questions larges tandis que Chemical Abstracts Condensates lui est supérieur pour la recherche d'information de nature plus spécifique, par exemple pour alimenter des activités de recherche. Quelques désavantages de Compendex par rapport à CA sont à signaler : les brevets n'y sont pas pris en compte, de même qu'on n'y trouve que peu de publications autres qu'en langue anglaise (journaux allemands notamment). Les termes interrogeables sont non seulement les mots libres du titre mais aussi des descripteurs contrôlés provenant d'un thésaurus de 12 000 termes ainsi que des termes additionnels d'indexation en langage libre. Les résumés des articles figurent sur la bande magnétique.

La version automatisée des VtB-Verfahrens-technische Berichte contient les mêmes informations que le fonds imprimé. Celui-ci couvre principalement les opérations uni-

taires, les procédés chimiques, la conception d'installations et d'autres aspects du génie chimique. Ne sont pas incluses les technologies relatives aux plastiques, textiles et produits pharmaceutiques. 8 000 articles spécialisés sont sélectionnés dans 800 périodiques et autres publications. Les VtB représentent donc une information complémentaire à celle des CA des plus souhaitables pour le spécialiste de génie chimique. Les termes interrogeables sont : mots-clés, auteurs, données bibliographiques. Tous ces termes peuvent être imprimés en plus du texte du résumé.

INSPEC couvre l'information dans le domaine de la physique, de l'électronique et de l'électricité, des ordinateurs et de la régulation. La première et la troisième sections sont particulièrement intéressantes pour le spécialiste du génie chimique ; elles contiennent plus de 100 000 références par an. Les questions relatives aux mesures et à la régulation ou encore à l'emploi d'ordinateurs pour le génie chimique peuvent probablement être résolues le plus aisément avec le fonds documentaire Inspecc. La recherche peut être effectuée sur les titres, des descripteurs contrôlés et des mots libres additionnels. Les résumés figurent également sur la bande magnétique.

Les services de résumé et d'index de l'American Petroleum Institute (15 000 articles de revue et 8 000 brevets) sont probablement le service le plus sophistiqué accessible actuellement sur ordinateur et qui relève du domaine du génie chimique. Tous les documents retenus sont résumés, analysés en profondeur et caractérisés par un nombre important de mots-clés choisis dans un thésaurus spécialement bien conçu et permettant un haut degré de rappel ainsi qu'une excellente pertinence. Les résumés et les indexages demandant beaucoup de main-d'œuvre, le service est relativement cher mais extrêmement fiable pour toutes les questions présentant un rapport même lointain avec le pétrole et les industries pétrochimiques (à l'exception de l'exploration et de la production).

Predicasts est une organisation spécialisée dans l'information sur les marchés et fournissant diverses bases de données dans ce domaine ; certaines d'entre elles donnent des références, d'autres des statistiques. Les références (26 000/an) sont indexées avec un vocabulaire contrôlé complété par des termes libres choisis en partie dans le titre et le résumé du document. Les résumés des documents sélectionnés peuvent être imprimés.

Jusqu'à présent l'essentiel de l'information contenue dans ces fonds documentaires concerne les États-Unis et, seulement une petite partie, le marché européen. Cependant, Chemical Market Abstracts fournit des résumés très complets et un bon indexage.

Chemical Industry Notes, qui est publié par Chemical Abstracts Service, fournit une sélection hebdomadaire d'extraits d'articles orientés vers le monde des affaires et qui intéresse l'industrie chimique, en particulier sur la production, les nouveaux produits, les ventes, etc. Sa version sur bandes magnétiques existe depuis 1972 ; on peut y chercher les mots du titre et ceux du texte de l'extrait : produits chimiques, marques de fabrique, procédés, sociétés, etc. La

sélection s'effectue sur 82 périodiques dont 47 sont américains.

Rapra Abstracts paraît deux fois par mois et couvre l'information technique et commerciale dans le domaine des matières plastiques et du caoutchouc. La version imprimée comprend 4 parties : articles de journaux, comptes rendus de conférence, nouvelles acquisitions de la bibliothèque (livres, normes et documents publicitaires) et brevets. Les trois premières parties existent aussi sur bandes magnétiques portant titre, référence et résumé. Pour faciliter la recherche, des descripteurs choisis dans un thésaurus ont été marqués dans le titre et le résumé ou s'ils manquaient là, ajoutés à la fin du résumé.

Central Patent Index est un service d'information pour les brevets chimiques. 300 000 nouveaux brevets provenant des principaux pays industrialisés sont ajoutés chaque année au fonds documentaire ; 90 000 de ceux-ci représentent de nouvelles inventions qui sont classées dans diverses sections dont l'une est le « Génie chimique » ; elles sont résumées en détail et codées pour la recherche. Les bandes magnétiques fournies aux souscripteurs comprennent un titre étendu du brevet, les classifications des offices des brevets, et deux types de code conçus spécialement pour la recherche : une classification hiérarchique aussi appelée code « manuel », et des codes chimiques par fragmentation. Ce fonds documentaire est le plus complet et le plus à jour dans le domaine des brevets ; la recherche dans le domaine des produits chimiques et des matières plastiques peut s'effectuer soit par les mots du titre, soit par le code manuel hiérarchique, soit par le code chimique par fragmentation et donne souvent des résultats satisfaisants ; dans les autres sections, toutefois, le code manuel est rarement suffisamment distinctif et les mots du titre trop peu nombreux pour que les recherches donnent de bons résultats ; ceci est notamment le cas dans la section « Génie chimique ».

World Patent Index représente une extension de ce service aux brevets non chimiques. Les mots du titre et les codes « IPC » sont les seules possibilités dont on dispose dans ce fichier pour effectuer des recherches par sujet.

I.F.I. Comprehensive Data Base to U.S. Chemical Patents, donne accès à plus de 350 000 brevets U.S. concernant l'industrie chimique et les industries apparentées et remontant à 1950. Des brevets correspondants originaires de pays européens ont été ajoutés lorsqu'ils étaient connus. Les recherches peuvent être effectuées sur la base de tous les mots des titres de brevets étendus, de la classification de l'Office Américain des Brevets et des déposants. Cette base de données est la seule qui donne accès à 25 ans de documentation ; ses inconvénients sont une couverture limitée aux brevets américains et des possibilités de recherche plutôt restreintes. Pour déterminer l'utilité de certaines de ces bases de données disponibles commercialement dans le domaine du génie chimique, onze questions tests ont été mises au point de manière à couvrir quelques-uns des aspects les plus importants du domaine du génie chimique. Ces questions tests sont résumées au tableau II.

Tableau II

Questions tests :

1. Modèles mathématiques de réacteurs chimiques.
2. Données physicochimiques (pression de vapeur, point d'ébullition, température et pression critique, chaleur de combustion, etc...).
3. Formule générale pour les propriétés de transport.
4. Cellules d'électrolyse avec diaphragmes en amiante.
5. Distillation du pétrole brut.
6. Distillation du benzène.
7. Distillation sous vide.
8. Distillation : mesure, régulation et applications des ordinateurs.
9. Transfert de chaleur dans les lits fluidisés.
10. Tuyaux et plaques dans l'équipement pour le génie chimique (échangeurs de chaleur et récipients sous pression).
11. Broyage des matières plastiques.

Ces questions ont fait l'objet de recherches effectuées soit en conversationnel soit en une seule passe sur Chemical Abstracts Condensates et Compendex; certaines d'entre elles ont fait l'objet de recherches complémentaires sur certains autres fichiers mentionnés au tableau I.

III. Principaux usages de l'information automatisée

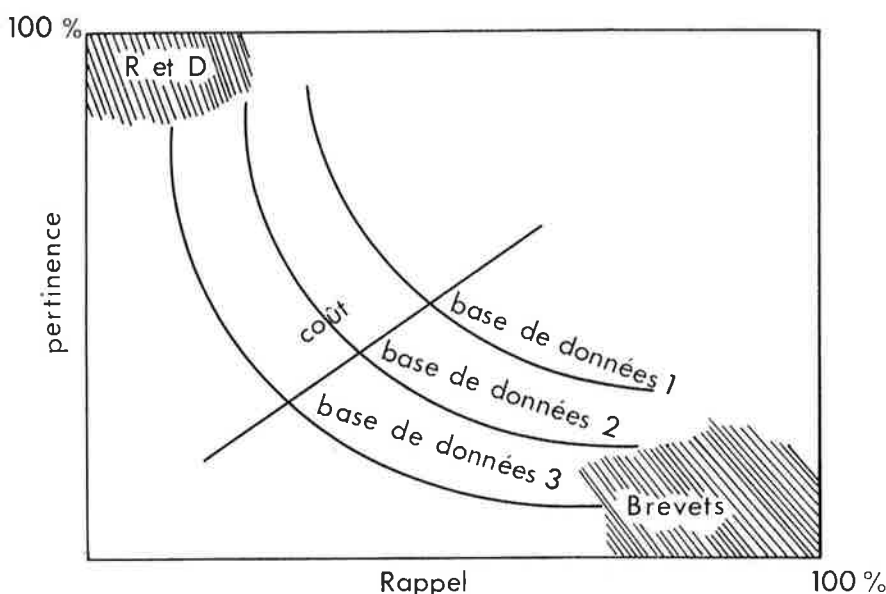
Les questions que l'on peut poser à ces différentes bases de données peuvent différer considérablement selon la fonction de la personne qui les pose. Des types de demandeurs et de questions sont résumés au tableau III ci-dessous.

Tableau III

Besoins	R et D	Planning	Production	Vente	Brevets
Données techniques	×		×		
Références bibliographiques	×				× (1)
Études de marché, informations économiques et financières		×		×	
Synthèses et revues	×	×		×	
Réglementations		×	×		
Diagrammes et plans	(2)		(2)		(2)

(1) L'exhaustivité qui est fréquemment nécessaire dans le domaine des brevets, spécialement pour des recherches d'infractions, est obtenue rarement avec les fichiers accessibles en conversationnel. A.P.I. constitue une exception car il permet de retourner jusqu'en 1964 et fournit un indexage en profondeur.

(2) Les diagrammes et dessins ne peuvent jamais être cherchés comme tels dans les fichiers actuels.



Pour la fourniture de l'information courante (dissémination sélective de l'information), un profil décrivant le domaine d'intérêt du demandeur est comparé à intervalles réguliers au contenu des mises à jour du fichier. Ceci constitue un traitement en un seul lot. Un grand nombre de centres d'information fournissent ce type de service. Pour la recherche rétrospective, le traitement en lot peut aussi être utilisé, mais la recherche en conversationnel gagne rapidement du terrain. La vitesse de recherche est un avantage majeur ainsi que la possibilité d'avoir l'utilisateur final présent lors du dialogue avec l'ordinateur; comme le nombre de réponses est annoncé immédiatement,

la recherche peut être modifiée en cours de route et l'objectif redéfini. La recherche en conversationnel est devenue faisable depuis que les progrès significatifs effectués dans les réseaux de télécommunication ont permis la connexion à longue distance d'un grand nombre d'utilisateurs avec les plus importants centres d'ordinateurs traitant l'information bibliographique (L.M.S., S.D.C. en Californie, E.S.A./S.D.S. à Frascati). L'efficacité d'une recherche peut être exprimée en termes de pertinence (pourcentage de réponses pertinentes par rapport au nombre total de réponses) et de rappel (pourcentage de réponses pertinentes comparées au nombre total de documents pertinents présents dans un fichier). Chaque base de données est caractérisée par une courbe particulière dans un graphique donnant la pertinence en fonction du rappel (voir figure 1). Grossièrement, cette courbe se déplace vers le haut et vers la droite quand le coût de production de la base de données croît. Des bases de données indexées en profondeur comme A.P.I. seront situées dans le coin en haut à droite, par contre des bases de données ne comportant que les titres et les noms d'auteurs se trouveront dans le coin en bas à gauche. En ajustant la stratégie de recherche, on peut faire se déplacer le point caractéristique d'une recherche donnée le long de la courbe de la base de données utilisée. On constate généralement que, par exemple, les utilisateurs finaux travaillant dans le domaine de la recherche et du développement peuvent souvent être satisfaits par une haute pertinence, même si le rappel est relativement bas; toutefois, à l'opposé, les chercheurs du domaine des brevets exigeront un haut rappel même au coût d'une pertinence assez basse.