MINISTERE
DE L'INDUSTRIE
ET DE LA RECHERCHE

LES BANQUES DE DONNÉES

dans le domaine scientifique et technique

par Edouard Labin

BUREAU
NATIONAL
DE L'INFORMATION
SCIENTIFIQUE
ET TECHNIQUE

LA DOCUMENTATION FRANÇAISE

Remerciements

Le présent ouvrage reproduit une étude sur la technologie des banques de données faite pour le compte conjoint du Bureau National de l'Information Scientifique et Technique (BNIST) et du Bureau de l'Information Scientifique (BIS) de la Délégation à la Recherche sur les Moyens d'Essais (DRME).

Je tiens à exprimer ma reconnaissance à Mr. Jacques MICHEL, Secrétaire Général du BNIST, et Mr. VIELLARD, Directeur du BIS, pour la confiance qu'ils m'ont témoignée et les encouragements constants, aussi amicaux qu'éclairés, qu'ils m'ont prodigués. Aux côtés de Mr. MICHEL, Mme. CORNIER a veillé sur mes ruades avec une patience exemplaire. Les dirigeants et animateurs des 21 banques de données que j'ai visitées, ont supporté avec courtoisie mes inquisitions et m'ont fourni des renseignements complets et illuminants. Qu'ils en soient tous remerciés.

Définitions

On appelle dans ce livre "BANQUE DE DONNEES", un ensemble de données concernant un même domaine de la science et des techniques, mises sur ordinateur et consultables rapidement dans tous les sens. C'est donc essentiellement un outil de DOCUMENTATION AUTOMATISEE, qui complète les centres documentaires destinés à fournir des références bibliographiques.

Par "DONNEE", on entend dans ce livre, une information factuelle bien circonscrite, qui porte toute sa signification en elle-même, c'est-à-dire pour l'intelligence et l'exploitation de laquelle on n'a pas besoin de contexte ou commentaires, et qui peut se ramasser en une expression succincte telle qu'un chiffre, un nom, une marque cochée dans un bordereau, un schéma, une carte, une photo.

TABLE DES MATIERES

PREMIERE PARTIE

ETUDE GÉNÉRALE

DEUXIEME PARTIE

DESCRIPTION DE BANQUES

TROISIEME PARTIE

BIBLIOGRAPHIE

première partie

ETUDE GÉNÉRALE

CHAPITRE 1

Qu'attend-on d'une banque de données ?

1.1 - LES DEFAUTS DE LA DOCUMENTATION LITTERAIRE

Le mouvement récent en direction des Banques de Données est issu de la déception qu'ont provoquée les Centres de Documentation que nous appellerons "littéraires", qui se proposent de fournir des références bibliographiques. Dans la suite, nous désignerons les premiers par l'abréviation "BD", et les seconds par l'abréviation "CDL".

On sait qu'après 18 ans (le premier centre documentaire automatisé, celui du DOD américain, remonte à 1957) d'efforts intensifs et de dépenses considérables, les CDL, tout en rendant des services, n'ont pas répondu aux espoirs ambitieux qu'on avait placés en eux. En gros, on peut dire qu'un CDL monté en vue de consultations aléatoires issues d'un public polymorphe (nous exceptons ainsi les CDL internes aux grandes entreprises et étroitement structurés selon leurs besoins) n'attire toujours pas plus de 5 à 15 % de la clientèle qui devrait logiquement l'utiliser, et que les listes de références qu'il fournit contiennent toujours 50 % de bruit et 50 % de manques. Aucun progrès notable sur cette situation de base n'a été accompli dans les dernières années, sauf l'introduction de la consultation conversationnelle du corpus sur console de visualisation. Cette méthode permet d'améliorer la qualité de l'extraction, mais c'est en faisant travailler davantage le client. En outre, il est très difficile et onéreux de placer des consoles partout, de sorte que le client doit se rendre au centre. Pour ces raisons et d'autres, la recherche conversationnelle d'information "littéraire" n'a pas modifié sensiblement la situation.

Quelle est la cause profonde de cette déficience des CDL, apparemment incurable ? La grande enquête de Crandfield de 1969-1970 a démontré que les erreurs dans la sélection bibliographique commis par les CDL sont dûes :

- pour 15 % aux difficultés techniques et linguistiques, en particulier la non standardisation de la terminologie ;

- pour 55 % aux insuffisances dans l'indexation initiale des documents formant le corpus ;
- pour 30 % à la mauvaise indexation des demandes présentées par les clients.

Les analystes ne perçoivent pas des thèmes qui ont pourtant été traités dans l'article qu'ils examinent, ou ils se trompent sur ceux qu'ils perçoivent, ou ils indiquent des liaisons logiques entre termes qui n'existent pas, ou ils choisissent les mauvais descripteurs. Tantôt ils en mettent trop, pour augmenter la *pertinence* du tri, mais alors ils y aggravent les *manques* ; tantôt ils n'en mettent pas assez, pour *élargir* le tri, mais alors ils y aggravent le *bruit*.

On a essayé de suppléer à l'analyste qui indexe des documents "à la main", c'est-à-dire avec le seul recours de son intelligence et de dictionnaires, par des méthodes d'*indexation automatique*, et des méthodes de *classification automatique*. Les premières tendent essentiellement à faire émaner les descripteurs et le thésaurus non plus d'une concordance "au jugé" entre les mots d'un document et des listes de descripteurs dressées *a priori*, mais de l'analyse statistique et linguistique des documents eux-mêmes, faites par l'ordinateur qui les lit. Les secondes méthodes tendent essentiellement à lever les ambiguïtés sur le sens dans lequel il faut prendre les termes, et à proposer des liaisons logiques correctes entre eux, en les distribuant parmi des classes homogènes et significatives quant à la place de leurs membres dans la hiérarchie générale des notions ; cette classification étant, elle aussi, effectuée automatiquement par l'ordinateur au moyen d'un algorithme basé sur la définition d'une "distance" numérique entre termes (ou de son complément l' "indice de ressemblance"). Or, en dépit d'efforts considérables qui se poursuivent depuis déjà dix ans, et de programmes d'indexation ou de classification automatiques extrêmement élaborés, aucun corpus en service en 1975, dans aucun CDL, n'a été indexé par aucune des 15 à 20 méthodes automatiques qui ont été proposées. Elles n'ont pas davantage servi, dans la pratique, à indexer les demandes des clients. Le mieux que l'on ait tiré de ces travaux, au demeurant fascinants sur le plan de la spéculation intellectu-

elle et de la linguistique, fut d'apporter aux analystes un moyen d'améliorer un peu leurs outils, et spécialement la confection du thésaurus.

On n'a pas réussi jusqu'ici à rendre l'indexation de documents vraiment satisfaisante. Pourquoi ? C'est la question clé qu'il faut placer à l'origine d'une étude en profondeur des Banques de Données.

On a cru pendant longtemps que la qualité de l'indexation était une question de prix. Dans les "Information Analysis Centers" américains, on a constaté qu'une analyse vraiment optimale d'un texte pouvait demander deux ou trois heures à un analyste lui-même spécialiste très averti du domaine. Un centre qui reçoit 10.000 textes nouveaux par an devrait ainsi dépenser 4 millions sur le seul poste des analystes, qui ne représente qu'un cinquième de ses frais. Mais en vérité, même dans ce cas optimum, on n'arrive pas encore à une indexation sans faille.

La raison est donc plus profonde. Elle git dans la nature même de l'information scientifique primaire qu'on demande au CDL. Celle-ci charrie, en effet, dans son état naissant, des textes qui sont *intrinsèquement flous*. La pensée d'un auteur est surtout féconde lorsqu'elle *sort* de ces sentiers balisés que l'analyste a pour mission d'y tracer. Il n'est pas jusqu'au sens convenu des mots que le créateur n'ait tendance irrésistiblement à transgresser, comme Claude Bernard déjà l'avait souligné. En plus d'être *flou*, le message informatif est par essence *polymorphe*, ou si l'on veut *polyvalent*. La plupart des documents scientifiques traitent de plusieurs sujets à la fois, et pour la désignation desquels on ne dispose pas d'un vocabulaire convenu, ni même adéquat. Combien d'articles sur une maladie ne sont-ils pas bourrés de renseignements intéressants sur tel composé organique ? Dans un article sur les oscillateurs synchrones que l'auteur avait envisagés comme des recopieurs de fréquence, plusieurs lecteurs avaient vu des stabilisateurs automatiques de phase ; etc.

L'élargissement continu du tissu scientifique, où désormais tout tient à tout, rend de plus en plus ardue la tâche d'isoler et de désigner clairement les thèmes dont traite un article. Il faut ajouter à cela que le rôle de l'auteur n'est pas de songer à toutes les manières possibles d'utiliser ses résultats. Il a écrit sur le comportement d'un enzyme dans tel acide gras parce qu'il s'intéressait à la multiplication cellulaire. Il n'était nullement tenu de prévoir que tel autre chercheur pourrait trouver dans son travail des indications sur les maladies tropicales, et que tel autre pourrait en tirer parti du point de vue du marché des huiles. Le seul devoir de l'auteur est de développer le cheminement de sa pensée à lui, avec toutes ses nuances et tous ses détours. C'est, en fin de compte, pour avoir cette confession qu'on le lit. Mais mieux il accomplit ce devoir, plus il alourdit et rend confuse la tâche de l'analyste indexeur.

Les choses ne se présentent guère mieux du côté de l'analyse de la demande. Celle-ci émane d'un client qui par définition se trouve dans le doute. Il ne sait pas, il ne peut pas savoir, il ne doit pas savoir exactement ce qu'il veut. Le mieux qu'on puisse dire de ses souhaits, c'est qu'il cherche une compagnie intellectuelle, un réconfort, une lumière dans les ténèbres. Il peut la trouver dans une remarque latérale, voire une inflexion, d'un discours qui visait tout autre chose. Quand le meilleur analyste a résumé une demande de documentation en une *équation de recherche*, le client a toujours l'impression qu'on a trahi sa quête, car on l'a réduite à une recherche de titres, alors qu'elle était une recherche d'inspiration. En tant que créateur en puissance, les besoins du client sont aussi essentiellement confus que les écrits des créateurs qui l'ont précédé. Lorsqu'un centre documentaire se voit reprocher par son client "le bruit" ou les manques" dans la liste de référence qu'il lui a livrée, il se sent souvent accusé injustement car il constate qu'il ne s'est trompé que *par rapport à des desiderata que le client n'avait pas formulés*. Et pourtant, le client n'est pas coupable non plus, car il lui était impossible de préciser ces desiderata davantage *tant qu'il n'était pas lui même éclairé* par la littérature. La production et la recherche d'information littéraire appartiennent à ces fameux cercles vicieux de la vie où un dessein ne se définit que quand il est accompli.

On pourrait être tenté de résumer ces difficultés de l'indexation de l'information scientifique de type littéraire en disant que celle-ci est trop subtile et trop complexe pour un service qui devra être finalement confié à l'ordinateur. Ce n'est pas une façon correcte de présenter les choses. L'ordinateur peut être instruit d'exécuter les programmes les plus complexes dès lors que le programmeur sait définir exactement la tâche et coder les éléments de travail à chaque pas. Le problème réside donc dans *l'ambiguité* du matériau, non dans sa *complexité*. On indexe mal, et par suite le CDL répondra mal, parce que l'analyste ne sait pas au juste ce qu'il doit voir dans un texte, et que le client ne sait pas au juste ce qu'il doit demander aux textes.

1.2 - LE PASSAGE AUX BANQUES DE DONNEES - DEFINITIONS DE BASE

A partir du moment où l'on accepte ce diagnostic, on est nécessairement conduit à chercher des services documentaires dans une autre voie, où on ne souffrirait pas de cette indéfinition foncière des matériaux mêmes qu'on doit ordonner et ficher. C'est ainsi qu'on s'est tourné vers le projet de stocker et fournir non plus des *références bibliographiques*, mais des *données*.

Le moment est propice pour définir aussi clairement que possible le sens dans lequel nous allons prendre ce mot.

> *"Par "DONNEE", nous entendrons une information factuelle bien circonscrite, qui porte toute sa signification en elle-même, c'est-à-dire pour l'intelligence et l'exploitation de laquelle on n'a pas besoin de contexte et commentaires, et qui peut se ramasser en une expression succincte telle qu'un chiffre, un nom, une marque cochée dans un bordereau, un schéma, une carte, une photo".*

On voit que les données ne sont pas forcément quantitatives, mais en fait elles le sont le plus souvent dans les applications.

Les fameuses "propriétés" des corps ou des processus, pour autant qu'elles aient été isolées et identifiées, constituent le prototype des "données", encore que nous trouverons beaucoup de données d'un autre genre. Pour bien comprendre ce concept de "données", et achever de le définir, il importe de distinguer trois notions que nous allons appeler, comme si pour énoncer une donnée il fallait articuler une phrase :

. LE SUJET
. LE PREDICAT
. LES CIRCONSTANCES

Le *sujet* sera naturellement ce sur quoi porte la donnée. L'ancienne rhétorique, pour faire comprendre en quoi le sujet d'une proposition se distingue des attributs que cette proposition lui assigne, enseignait qu'un attribut tel que "grand", "blond", "mou", "bout à 100°", "provoque un empoisonnement", etc..., ne se conçoit pas, ni dans la nature ni dans notre esprit, seul et autonome sans que quelque chose lui serve de support ; par exemple, un homme, une chevelure, la pâte à modeler, l'eau, la strichnine. Ce support est le *sujet* qui, lui, au contraire, possède une existence propre qu'on peut appréhender en dehors de tout attribut. La chose qui est rouge se conçoit sans le rouge, mais le rouge ne se conçoit pas sans quelque chose qui soit rouge. Pour donner des exemples empruntés aux BD, les sujets des données fournies par ces banques seront : des substances, des endroits du sol, de l'air ou de l'océan, des cas d'empoisonnement, des milieux naturels, des fossiles, des étoiles, des processus, etc... sur lesquels on possède des informations ayant la nature de "données".

Le mot d' "attribut" par lequel nous venons de désigner les caractéristiques relevées chez un sujet, pourrait nous convenir s'il n'était trop lié, dans son acception courante, à l'idée de "qualifiant". Or, les banques de données fournissent souvent sur leurs sujets des informations qui n'ont pas la nature d'un "qualifiant", comme par exemple les roches que l'on trouve dans un sous-sol donné. C'est pour avoir un terme plus général, applicable à n'importe quoi que l'on dit d'un sujet : une qualité,

une propriété, une origine, une manière de se comporter, une localisation, que nous introduisons le mot de *prédicat*. Car ce mot est justement défini en linguistique comme désignant "ce qu'on dit d'un sujet". En fait, dans la plupart des cas, les mots "attribut", "propriété", "prédicat", voire "caractéristique", sont interchangeables, et désignent tous des "données". (Pour autant que sont respectées les autres conditions contenues dans la définition de la page 19).

Nous verrons de plus en plus que le produit stocké dans, et livré par une BD se ramène effectivement à une "phrase de données", notion qui s'éclairera progressivement dans la suite. Mais normalement, entre le sujet et le prédicat d'une phrase, il y a un *verbe*, qui est le plus souvent la simple copule "être", mais peut aussi désigner n'importe quel autre type de rapport. C'est par exemple "bouillir" dans le cas de la phrase de données "l'eau *bout* à 100°". Ca peut être "guérir" dans le cas de la phrase de données "l'aspirine (sujet) *guérit* la grippe (prédicat)". Mais on peut ne pas mettre en évidence le verbe, parce qu'il est en général patent d'après la constitution de la banque : quand on sait, par exemple, qu'elle livre des températures d'ébullition, il suffit de mettre en regard *sujet* et *propriété* pour qu'on sache en quoi et comment la propriété s'attache au sujet. Le verbe est en quelque sorte *implicite* dans la totalité d'une BD homogène. Nous verrons toutefois que certaines banques abritent plus d'un type de phrases de données, et dans ce cas les verbes qui nichent dans ces phrases doivent être explicités et distingués.

On sait que le langage courant ne dispose souvent pas de deux mots distincts pour désigner la *nature* et la *valeur* d'un attribut ou d'un sujet. Ainsi, le mot "fortune" désigne à la fois le concept "ensemble des biens possédés" *et* l'un des états possibles de cet ensemble, peu ou beaucoup rempli. Cette pauvreté du vocabulaire est à l'origine de nombreux troubles de l'expression et il importe particulièrement de s'en garder dans le domaine des Banques de Données. Lorsqu'on dit qu'un client cherche "une donnée", on veut dire réellement qu'il cherche la *valeur* qu'a un prédicat d'une certaine catégorie dans un *certain cas d'espèce*. Autrement dit, il cherche un *élément d'un* ensemble, *l'ensemble*

définissant la *nature* du prédicat et l'élément définissant une *valeur* particulière. Vu que le glissement entre l'acception "nature de... et l'acception "valeur de..." d'un même mot est particulièrement gênant dans le domaine des BD, nous prendrons le parti, quand la distinction ne ressortira pas clairement du contexte, d'utiliser les suffixes "*n*" de *nature*. et "*v*" de valeur. Nous aurons ainsi :

Sujet-n	ex. : les métaux
Sujet-v	ex. : le fer
Donnée-n	ex. : la densité
Donnée-v	ex. : 5,3 kg/dm3

Passons maintenant au troisième concept fondamental que j'ai cité plus haut, celui de "*circonstances*". Il est là pour *achever de définir*, soit le sujet, soit le prédicat. Quand je parle d'une tôle de métal, même si je précise chimiquement de quel métal, mon sujet n'est pas encore suffisamment défini. Car on sait que ses propriétés vont dépendre d'une foule de facteurs : sa pureté, son mode de préparation, les traitements thermiques qu'elle a subis, la manière dont on l'a découpée, etc... Dans cet exemple des métaux, un institut américain a compté qu'il fallait spécifier *onze* paramètres sur la genèse d'une pièce métallique avant de pouvoir énoncer valablement un prédicat dessus ; le moins variable de ces paramètres pouvant avoir 4 valeurs distinctes, et le plus variable 56 ! Ce sont des paramètres de ce genre que nous appellerons dans ce rapport les "*circonstances*". Ce mot est choisi exprès très général parce que les précisions qu'il faut ajouter pour définir pleinement un sujet sont de nature extrêmement diverse. Ce peuvent être des dates, des lieux, des états de l'environnement, des marques commerciales, des compositions chimiques, des âges, des labels, etc... etc...

En général, les circonstances à préciser concernent surtout le *sujet* d'une phrase de données. Mais elles peuvent aussi concerner le *prédicat*, par exemple quand il faut indiquer la méthode de mesure pour donner son plein sens à une propriété. Parfois, la couronne des circonstances qu'il faut fixer pour définir pleinement un sujet est tellement touffue que le sujet s'y noie.

Par exemple, dans l'industrie du bâtiment, on rencontre des pièces dont la définition requiert une bonne douzaine de caractéristiques, y compris d'emploi ; et dans cette masse, on ne sait plus si le sujet est la pièce -une plaque, par exemple- ou l'emploi -par exemple, un isolement. Cela pose, comme on le verra, des problèmes spéciaux pour parvenir à ce dont on parle. Dans des domaines non techniques, qui sortent du cadre du présent rapport mais sur lesquels il n'est pas mauvais de jeter un coup d'oeil de temps à autre, cette situation peut aller jusqu'à mettre en cause la notion même de sujet. Ainsi, dans une banque de données sur les exportations ou importations, où faut-il arrêter le sujet : au concept "exportations" tout seul, ou au concept plus circonstancié "exportations de fer", ou au concept encore plus circonstancié "exportations de fer travaillé", ou au concept encore plus circonstancié "exportations de tôles de fer du pays A vers le pays B" ?

Lorsque l'identification complète d'un sujet porteur de données requiert beaucoup de circonstances enchevêtrées, il n'y a pas que l'organisateur de la banque qui est gêné. Le client l'est tout autant, parce qu'il lui est difficile de cheminer à travers cette forêt pour parvenir à un sujet bien circonscrit. Cela pose des problèmes que nous retrouverons plus bas. D'autres problèmes graves suscités par la prolifération des circonstances, et que nous reverrons aussi, concernent la création des fichiers inverses.

Jusqu'à nouvel ordre, nous admettrons qu'il n'y a pas de doute sur l'identité des choses que nous désignons par les trois termes fondamentaux : *sujet*, *circonstances* et *prédicat*. C'est tout de même le cas le plus fréquent.

Ces définitions font ressortir clairement la structure générale d'une BD. Une BD loge en ordinateur des phrases de données comportant des sujets-v d'une certaine classe de sujets-n, sujets entourés de circonstances-v appartenant à une ou plusieurs classes de circonstances-n, et des prédicats-n d'une certaine classe, éventuellement entourés aussi de circonstances, avec, entre les deux parties de la phrase, un ou plusieurs verbes qui, restant toujours les mêmes pour une même banque, ne sont pas

explicités. Cette phrase établit entre ses quatre parties un lien tel que, quand on désigne des valeurs particulières de trois d'entre elles, la valeur de la quatrième est déterminée et peut sortir automatiquement, à condition toutefois que parmi les trois parties désignées figure au moins un sujet-v, ou au moins un prédicat-v.

"L'équation de demande" ne contient plus, comme en CDL, des mots-clés, mais la désignation en clair, ou plus ou moins codée, de ces parties imposées de la phrase. C'est là une différence marquante et capitale avec la documentation littéraire.

Lorsqu'on spécifie le sujet-v et qu'on demande un prédicat-v nous dirons qu'on interroge la banque *en direct*. Lorsque, au contraire, on spécifie un prédicat-v et qu'on demande le sujet-v qui l'admet, nous dirons qu'on interroge la banque *en inverse*.

1.3 - POURQUOI UNE BD DEVRAIT ETRE MEILLEURE QU'UN CDL

On comprend dès lors pourquoi il est légitime d'espérer que l'information d'une BD pourra être identifiée et stockée sans erreur, et fournie ensuite sur demande sans bruit ni manque. C'est parce que, en réalité, on sait d'avance énormément de choses sur elle. Si je demande à une banque de données géologiques la teneur en calcaire de tel sous-sol, je sais déjà que je vais m'intéresser à un endroit parfaitement déterminé de l'univers ; je connais non seulement mon sujet-n, mais encore mon sujet-v. Je sais ensuite que je m'intéresse au calcaire, pas à l'argile, ni à l'eau, ni à l'humus, ni aux mollusques, etc... Je sais encore qu'à propos du calcaire, je m'intéresse à une teneur, pas à sa nature, ni à sa provenance, ni à ses efforts, ni à sa densité, ni à son pouvoir filtrant, etc... En fait, je sais d'avance la cellule précise de l'ensemble des connaissances, même dans le seul domaine où je me situe, peut avoir des millions de telles cellules. Je connais le sujet-v, je connais toutes les circonstances-v, je connais le prédicat-n : la seule chose qui me manque encore est sa valeur, le prédicat-v. Rien d'étonnant que je puisse l'obtenir de la banque *à coup sûr*, si tant est qu'elle le possède. On voit combien la situation est plus favorable que dans la recherche de références bibliographiques, où l'on en savait si peu sur ce que l'on voulait.

Au fond, si on est en droit d'attendre d'être parfaitement renseigné par une banque de données, c'est parce qu'on lui demande peu de renseignements. On bénéficie d'un théorême fondamental de la théorie de recherche, qui a l'air d'une lapalissade mais n'en est pas une : à savoir qu'*une recherche aboutit d'autant mieux qu'on en sait davantage d'avance sur la chose recherchée*. Si je demande au radar de me dire tout ce qu'il y a dans le vaste ciel, il présentera beaucoup de bruit et de manques. Mais si je lui demande tous les avions qui se dirigent vers tel secteur à une altitude et avec une vitesse données, il me fournira un joli tableau bien net.

Un théorême voisin est contenu dans l'énoncé que voici : *l'ordinateur est qualifié pour traiter le connu, l'inconnu restant l'apanage de l'homme*. Et de fait, dans toutes ses applications réussies, l'ordinateur n'a d'autre fonction que d'exécuter à supervitesse des travaux dont tous les éléments étaient parfaitement connus, en vertu d'une science antérieurement acquise par l'homme. La documentation automatique confirme pleinement cette loi. Elle ne marche pas bien dans le cas "littéraire" parce qu'elle traite alors de la science en train de se faire. Dans le cas des données, au contraire, on exploite la science déjà faite. Les données ne sont pas entourées de flou parce que des décennies, voire des siècles de recherches l'ont dissipé et ont permis de dégager des notions et des résultats ponctuels, nets, bien circonscrits, bien définis, dont la signification fait partie du bagage intellectuel de tous, et n'est plus contestée.

Le même théorême se vérifie aussi négativement. On constate en effet que même dans les BD, chaque fois il y a un trouble sur la nature d'un sujet ou d'un prédicat, le service de la banque commence à présenter les défauts classiques de bruits et de manques qui affectent les CDL. En particulier, lorsque l'ensemble des circonstances est si touffu qu'on se trompe sur celles qui entourent le sujet-v auquel un client s'intéresse, la réponse de la banque peut indiquer une donnée-v fausse appartenant à un autre sujet-v du client (manque).

Il importe de remarquer que dans le mécanisme documentaire qui vient d'être décrit, *c'est au demandeur qu'incombe la charge*

de cette qualité du renseignement qui distingue les BD des CDL. C'est lui qui doit avoir sur la chose, objet de sa demande, *toutes les informations moins une*. C'est à lui aussi qu'incombe le soin de formuler sa question en termes non ambigus et immédiatement opérationnels. Il peut le faire, en principe, parce que dans les BD, *il n'y a pas de descripteurs*, termes résumés plus ou moins artificiels. Les mots de la phrase de données qu'un client demande à une banque de compléter à l'endroit d'une lacune, sont ceux-là mêmes par lesquels on désigne les sujets, les circonstances et les prédicats dans le langage ordinaire. On peut dire qu'avec les BD, la *réponse ressemble à la demande*, tandis qu'avec les CDL, elle en diffère structuralement. C'est là une différence frappante entre les deux types de centres documentaires.

Cette différence est liée à un autre trait essentiel des BD, qui se fait sentir à tous les échelons de leur organisation : *le champ d'une BD doit en principe recouvrir tout le domaine qui peut être embrassé par un même type d'interrogation*, ce type lui-même étant assez simple pour pouvoir être formulé sans apprentissage préalable par tout technicien du domaine, dans le langage naturel de celui-ci. On peut dire, en simplifiant beaucoup mais sans faire violence à la nature des choses, qu'il doit y avoir autant de banques de données qu'il y a de types de questions factuelles que l'esprit peut poser : "une thématique - une banque", et inversement.

En principe, une BD ne peut pas répondre mal si elle est correctement interrogée. Elle ne donne lieu ni à bruit ni à manque, plus exactement, pas à un manque par rapport au corpus qu'elle possède. Ses deux défauts possibles sont :

- d'une part, que ce corpus soit insuffisant, en ce sens que des données pertinentes n'y ont pas été enregistrées. Nous appellerons ce défaut un "*trou*".

  d'autre part, que les données siégeant dans la banque soient incorrectes. Nous appellerons ce défaut une "*in consistance*".

On observa que ce sont, dans les deux cas, des défauts de la banque, ou de la science, non de l'algorithme de tri.

1.4 - L'ESPOIR D'UNE CLIENTELE ELARGIE

A la vérité, l'intérêt pour les BD ne fut pas stimulé seulement par l'espoir qu'on pourrait en extraire l'information d'une façon plus performante que d'un CDL. On escomptait aussi qu'elle intéresserait davantage de gens.

Le moment est en effet venu de préciser un point important. Quand, plus haut, on a mentionné le flou essentiel d'un document qu'on analyse pour l'insérer dans un corpus, ainsi que d'une demande de références bibliographiques, on se plaçait implicitement dans l'univers de la recherche scientifique ou du développement technique fondamental, où les principaux clients sont des savants, chercheurs ou ingénieurs de recherches. Ce sont eux qui, comme il a été dit plus haut, cherchent surtout dans la littérature des idées, des inspirations, des précédents. Mais les besoins d'information de cette sorte sont loin d'être les seuls. Ils sont même exceptionnels. Ils sont peut-être stratégiques pour le progrès des sciences, mais relativement rares et certainement pas primordiaux pour la bonne marche de l'économie. Celle-ci a le plus souvent besoin précisément de ce que nous appelons ici des données, bien plus que de savoir ce qui s'est dit sur un certain thème. Les chercheurs fondamentaux ont, eux aussi, grand besoin de données. En fait, l'enquête conduite en 1965 par le National Bureau of Standards américain, enquête qui a précédé la création dans son sein du "NATIONAL STANDARD REFERENCE DATA SYSTEM", a conclu que les chercheurs des laboratoires de science fondamentale, quand ils ont besoin d'information, ont besoin *deux fois sur trois*, non pas de littérature, mais de données. Et outre les chercheurs deux fois sur trois, une énorme population a besoin de données neuf fois sur dix :

- les ingénieurs et techniciens des diverses phases du développement et de la fabrication ;
- les commerçants ou technico-commerciaux ;
- les concepteurs de projets ou plans ;

- les administrateurs de divers échelons ;
  les médecins ;
- les pharmaciens ;
- les avocats ;
- les architectes ;
- les promoteurs et entrepreneurs ;
- les administrateurs d'innombrables institutions publiques régissant le patrimoine collectif de l'eau, de l'air, du sous-sol ;
  les météorologues ;
- les astronomes ;
- les douaniers ;
- sans compter, hors du territoire des techniques dérivées des sciences dites exactes, tous les économistes, statisticiens, psychologues, pédagogues et sociologues ;
- sans compter, encore plus largement, les gens qui ont besoin de renseignements factuels pour des activités prosaïques telles que trouver quelles sociétés en Allemagne s'occupent de dépôt de couches magnétiques sur des cartons, qui fait quoi, qui a dit quoi, qui a participé à quoi, toutes questions ressortissant à des fichiers automatisés qui s'organisent au fond et s'interrogent comme des BD.

On arrive donc à la conclusion que les BD pourraient, en principe, avoir un marché considérablement plus étendu que les CDL, puisqu'elles devraient, en principe, intéresser tous ceux qui travaillent sur des questions complexes en utilisant l'esprit et les connaissances acquises, ce que fait pratiquement tout le monde dans notre société hautement industrialisée. Selon cette vision, tout ce qui est de la nature d'une donnée devrait être immédiatement disponible au téléphone à quiconque en a besoin, au moyen d'un immense réseau de banques fonctionnant en SVP, et qui constituerait le trésor intellectuel le plus précieux de nos sociétés. Ce rêve n'est pas absurde en soi. Il faudrait seulement que la société comprenne que le stock de ses connaissances est aussi important que son stock d'or...