A CHACUN SA BANQUE
(TO EVERY MAN HIS BANK)

Pierre Garneau
Equipe IF
Montréal, P.Q.

RESUME

Les besoins de l'équipe IF a l'Université de Montréal furent le point de départ pour le développement d'un système de programmes pour l'établissement et la gestion de banques d'information. L'accent était mis sur l'optimisation de la recherche d'informations et sur la souplesse d'utilisation. Les programmes permettent de créer une banque, d'entrer et sortir de l'information, et de produire divers types d'index imprimés. La structure interne de la banque est basée sur le principe suivant: chaque mot-clef ne contient pas l'information de l'enregistrement, mais contient un pointeur dans une table unique (ainsi l'espace-mémoire est très réduit). Par ailleurs, les blocs d'information étant reliés entre eux d'une façon dite "séquentielle-indexée", on peut accéder directement à un mot-clef (ainsi, le temps d'exploitation est très réduit). Ce système a été utilisé pour plusieurs banques et s'avère très satisfaisant. (The IF Team at the University of Montreal had to manage several data bases; this provided the starting point for the work described in this paper. A system of programmes had to be set up. The accent was placed on optimising the retrieval procedures and on flexibiliby of use. The programmes allow one to set up a data base, put in and take out information and produce different types of printed indexes. The internal structure of the bank is based on the following principle: each keyword contains no document accession information but contains a pointer in a single table (in this way, the required memory space is extremely limited) On the other hand, the units of information are linked together in an "Indexed-Sequential" file, so one can accede directly to a keyword (so the processing time is reduced). This system has been used for several data bases and is proving very satisfactory.)

A CHACUN SA BANQUE

Après avoir travaillé avec l'équipe IF (1) au développement d'un système de programmes pour la constitution et la gestion d'un thésaurus (2), un besoin s'est fait sentir, au printemps 72, d'un système permettant la conservation et le dépistage de l'information. Il s'agissait plus précisément, de générer et d'entretenir un index inversé à pondérations multiples, et d'y appliquer la coordination des concepts pour la recherche. Prévu à son origine pour une utilisation strictement interne, ce système - après plus d'un an d'utilisation - est en voie de devenir un outil de recherche et de communication utilisé par des personnes d'intérêts divers, pour des applications n'ayant apparemment que peu de choses en commun. Il est assez surprenant de constater l'ampleur du champ d'application qui dépasse de loin la simple recherche documentaire des index inversés, et, d'autre part, l'utilité d'un effort déployé dès le début en vue de généraliser la réponse aux besoins initiaux et de ce fait, de faciliter le maniement des divers programmes composant le système. C'est pourquoi, nous allons en retracer les diverses étapes de conception, de mise sur pied et d'utilisation.

I EMERGENCE DU BESOIN ET CONCEPT INITIAL DU SYSTEME

L'équipe IF, dont le but premier est de résoudre les problèmes de communication et d'information dans l'industrie de la construction, fait partie de la Faculté de l'Aménagement de l'Université de Montréal; c'est donc dans un milieu pluridisciplinaire et non-scientifique (dans le sens conventionel du mot) qu'est né le besoin d'un système de conservation souple et facile d'accès, pouvant répondre partiellement aux besoins d'information sans cesse croissants. Les activités de l'équipe qui ont plus précisement motivé le système incluent: la publication dans les deux langues d'une revue bimestrielle (Industrialisation Forum)

---

(1)L'équipe IF, à la Faculté de l'Aménagement de l'Université de Montréal, s'occupe de problèmes associés à la manipulation de l'information dans le domaine du bâtiment. Elle a établi plusieurs banques de références bilingues, utilisées pour des fins professionnelles et pédagogiques; de plus, elle publie une revue "Industrialisation Forum" (d'où le sigle IF) qui contient un système intégré d'information.

---

(2)IF Thesaurus of Building Science and Technology Montréal, Université de Montréal, 1972, dont les principes de construction furent présentés à la première conférence publique de l'ACSI, en 1973. Il s'agit d'un thésaurus hiérarchisé, avec un réseau de relations et de niveaux stricts - nécessitant un contrôle rigoureux. Ce thésaurus a été traduit en Français, ce qui nécessitait la préparation de programmes particulièrement souples.

qui fournit une liste de mots-clefs correspondant à chaque article et fiche bibliographique publiés ainsi qu'un index inversé sous la forme de fiches de mots-clefs. Par ailleurs, l'équipe IF était en train de constituer une Banque de Références pour le Ministère de l'Industrie et du Commerce, accompagnée d'une indexation exhaustive, spécifique et pondérée.

Evidemment, il était souhaitable, en même temps, de pouvoir tester le thésaurus dans la pratique, en l'utilisant pour toutes les activités de stockage et de recherche de l'information - ce qui implique l'existence d'outils appropriés.

C'est à partir de ces activités, et de l'expérience de ces problèmes non-numériques inhérente à son domaine, que l'équipe IF a composé un devis initial décrivant les capacités minimales que devrait avoir le système projeté pour résoudre ses problèmes immédiats, car la publication et la gestion manuelle de l'index inversé de la revue IF étaient de plus en plus pénibles, de même que la conservation du parallélisme des deux langues. Quant au thésaurus, comme nous l'avons mentionné, sa structure et son contenu devaient être vérifiés sur une grande quantité d'information, trop grande pour être commodément manipulée à la main, surtout en conjonction avec un thésaurus comptant plus de 5000 termes dans chacune des deux langues.

Pour s'avérer utile dans ces divers cas, le système devait:

- Pouvoir supporter efficacement des banques d'information de dimensions très différentes;

- Pouvoir traiter, à l'indexation et à la recherche, des banques comportant un nombre différent de pondérations;

- Permettre, dans certains cas seulement, d'appliquer à l'indexation le contrôle d'un thésaurus;

- Pouvoir générer des index par mots-clefs, cumulés à partir d'un point précis d'indexation, avec une sélectivité quant aux pondérations restituées;

- Permettre de retrouver la forme non inversée, soit l'indexation courante d'un ou plusieurs enregistrements, qu'elle ait été modifiée ou non;

- Pouvoir supporter une indexation et une recherche dans au moins deux langues, sans dédoublement de l'information.

De même, certaines contraintes se sont imposées spécifiquement pour la recherche dans une banque, cette opération étant la plus

importante et la plus critique (3); cette recherche devait donc permettre de:

- Formuler des requêtes booléennes aussi complexes que nécessaires, sans restrictions sur l'utilisation des trois opérateurs et des parenthèses;

- Limiter la quantité d'information à restituer et préciser une requête précédente;

- Suivre optionnellement les étapes dans la résolution d'une requête pour en connaître les éléments déterminants.

Ces diverses exigences visaient à donner à l'utilisateur la possibilité d'effectuer une recherche précise en vue d'une coordination efficace, et, la capacité d'élargir ou de restreindre son domaine d'investigation (4).

II ELARGISSEMENT DU CONCEPT: COMPORTEMENT FACE AUX UTILISATEURS

Le point déterminant de la conception du système semble avoir été de considérer comme utilisateurs, non seulement les membres de l'équipe IF qui utilisent l'une ou l'autre des banques en particulier mais

---

(3)L'expérience de l'étape précédant la mise sur pied du système avec l'ordinateur lorsqu'on utilisait encore des fiches de carton pour un index inversé, indique qu'il n'est pas possible de travailler d'une façon manuelle et efficace; en effet, dans une banque de dimensions restreintes, chaque mot-clef contient peu d'informations, nécessitant ainsi un grand nombre d'unions de mots-clefs lors de la recherche: le traitement des unions est presqu'impossible en utilisant des fiches. Par contre, dès que la banque est assez remplie pour permettre une utilisation prédominante des intersections (ce qui est possible avec les fiches) la quantité d'informations sur chaque fiche est telle qu'il est difficile de les parcourir rapidement.

---

(4)Si la requête ne produit pas une réponse adéquate, il est important de pouvoir retracer les étapes de coordination, afin de savoir à quel moment l'insuffisance (ou l'excédent) des réponses s'introduisait. L'équipe IF conçoit que les processus de réponse devraient être interactifs, c'est-à-dire que l'interlocuteur peut modifier sa question originale à la lumière des réponses reçues entre-temps. Cette approche est le contraire d'une approche qui prônerait la modification des questions d'une façon automatique.

en outre, les groupes ou individus éventuels désireux de générer leurs propres banques. En ce sens, les besoins initiaux de l'équipe IF ont été regardés comme des cas particuliers, à l'origine d'un comportement plus général. Le but du système s'est modifié, afin de permettre à un utilisateur quelconque de constituer facilement son fichier électronique personnel, avec ses caractéristiques propres, pour une utilisation privée ou publique.

A cet effet, la versatilité initiale offerte à l'utilisateur est la capacité de définir:

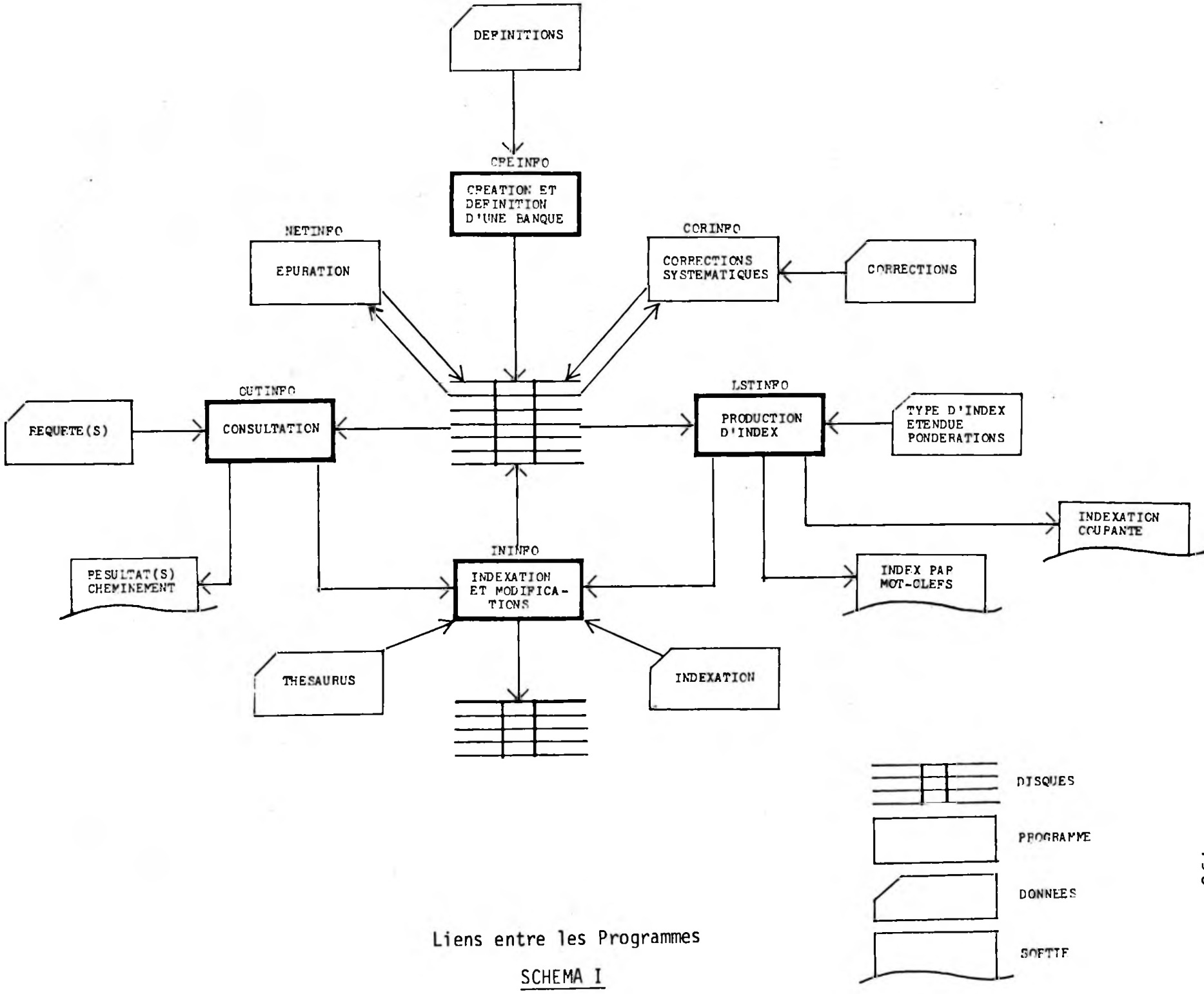
- Le nom de sa banque
- La forme de l'information à retrouver
- La forme de ses mots-clefs
- Le nombre de niveaux de pondération (limité à 63)
- La dimension limite de sa banque (limitée à 144,000 enregistrements)
- Des mots de passe pour assurer la sécurité nécessaire à son information et à son thésaurus s'il y a lieu.

Après ces définitions, qui seront conservées jusqu'à la destruction de la banque, l'utilisateur dispose d'un ensemble restreint de programmes d'usage général, pouvant interagir les uns sur les autres pour produire les comportements désirés. Chacun de ces programmes, prévu pour être utilisé autant de façon conversationnelle qu'en traitement par lots - lorsqu'il s'agit d'une quantité importante de données - offre en outre divers formats visant à faciliter la communication à l'entrée et à la sortie, communication à la fois avec des humains et avec d'autres programmes. Cette caractéristique rend le système extensible en permettant à un utilisateur de "brancher" ses propres programmes au système, pour générer des données ou traiter des résultats.

Le schéma I illustre les différents programmes et le transfert de l'information entre eux et avec l'extérieur.

III STRUCTURE ET COMPORTEMENT INTERNES

Les comportements internes du système, et son efficacité dans les divers cas, sont directement dérivés d'un petit nombre de décisions de base qui ont constitué des critères de design. La première de ces décisions est de mettre l'accent sur la recherche de l'information, c'est-à-dire faciliter la réponse aux questions. Ceci implique tout d'abord qu'il y ait une structure informationnelle visant à rendre l'accès, et les opérations boléennes, les plus efficaces possible, quitte à dépenser plus d'énergie à mettre la structure sur pied à l'entrée. D'ailleurs, la suite montrera que la structure choisie n'entraine pas une dépense exagérée à l'entrée puisqu'elle n'exige aucun déplacement d'information déjà enregistrée; entrer d'autres informations se réduit à une union



Liens entre les Programmes

SCHEMA I

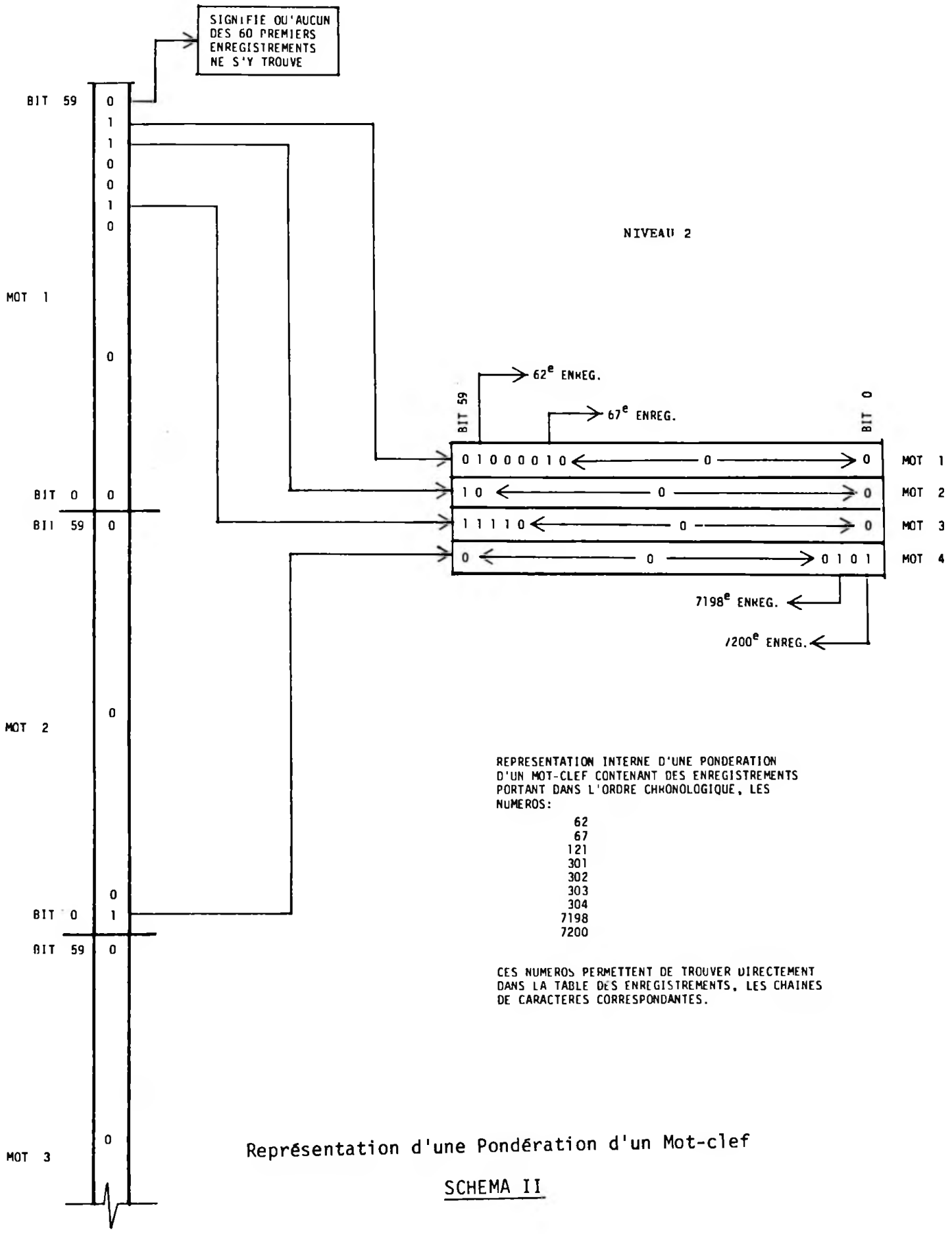
simple.

Le second critère considéré est de viser l'efficacité maximale pour une indexation complexe; en d'autres mots, les banques les plus importantes relieraient chaque enregistrement (chaque document) à un grand nombre de mots-clefs. Il s'ensuit que, dans la forme inversée, les mots-clefs ne doivent pas contenir l'information de l'enregistrement lui-même mais un pointeur dans une table unique où chaque enregistrement apparait une seule fois; cette table permet en outre de conserver l'ordre chronologique d'indexation des enregistrements, nécessaire pour produire les index cumulatifs. Par ailleurs, les pointeurs apparaissant sous les mots-clefs, tout en rendant la banque très compacte, nous permettent d'effectuer complètement une requête, de connaître le nombre d'enregistrements qui la satisfont et même, de la préciser ou de l'élargir à volonté, sans jamais avoir à manipuler l'information de l'enregistrement elle-même: l'information ne sera manipulée qu'au moment de l'impression, et alors, un seul enregistrement à la fois. Ceci nous permet d'utiliser dans une requête des ensembles contenant jusqu'à une banque entière (comme avec l'opérateur de négation) et même d'imprimer un tel résultat sans influence sur la mémoire nécessaire au programme.

Tandis que cette seconde décision concernait l'efficacité du point de vue espace, la troisième concerne le temps d'exécution: Pour permettre une utilisation puissante de la coordination des concepts, une banque doit pouvoir contenir des identificateurs communs englobant une grande partie des enregistrements d'une banque (par exemple, les pays, la langue d'origine, ou les années d'édition). L'utilisation d'un tel identificateur ne doit pas ajouter des coûts prohibitifs à la requête; le temps d'exécution du programme doit être fonction du résultat des opérations plutôt que de leurs "opérandes", c'est-à-dire que la structure des ensembles doit permettre, le plus possible, de ne pas chercher une intersection ou une union là où il n'y en a pas (le schéma II et les explications qui suivent permettront d'en juger.

Une conséquence du problème de la coordination des concepts est le soin qui doit être apporté à l'analyse syntaxique des requêtes; étant donnée l'importance relative d'une opération supplémentaire ou d'un résultat intermédiaire dans le cas où nous manipulons des ensembles qui peuvent être énormes, l'énergie consacrée à simplifier l'exécution d'une requête complexe devient éminemment rentable.

La représentation, sur disque magnétique, d'une banque d'information apparait comme suit: les blocs d'information correspondant aux mots-clefs sont reliés ensembles par une structure dite "séquentiel-indexé"; ce type de fichier permet d'accéder directement à un mot-clef (par exemple, pour une requête) en parcourant un nombre restreint de niveaux de sous-index, ou de passer tous les mots-clefs par ordre alphabétique (par exemple, pour un index) en suivant un chaînage prévu à cet effet.



REPRESENTATION INTERNE D'UNE PONDERATION D'UN MOT-CLEF CONTENANT DES ENREGISTREMENTS PORTANT DANS L'ORDRE CHRONOLOGIQUE, LES NUMEROS:

62
67
121
301
302
303
304
7198
7200

CES NUMEROS PERMETTENT DE TROUVER DIRECTEMENT DANS LA TABLE DES ENREGISTREMENTS, LES CHAINES DE CARACTERES CORRESPONDANTES.

Représentation d'une Pondération d'un Mot-clef

SCHEMA II

A CHACUN SA BANQUE

Dans le fichier ainsi structuré, on trouve, en plus des mots-clefs, des synonymes ou des traductions, c'est-à-dire des mots-clefs spéciaux, ne contenant aucune autre information qu'un pointeur vers un autre mot-clef auquel accéder à leur place.

On trouve aussi des pseudo-mots-clefs inaccessibles aux utilisateurs, utilisés par le système pour conserver la définition de la banque, la table segmentée des enregistrements, etc...

Pour ce qui est des mots-clefs eux-mêmes, le schéma II montre la structure correspondant à une pondération d'un mot-clef; un premier niveau de mots de mémoire contient un "bit" pour un ensemble de soixante enregistrements chronologiquement consécutifs (l'ordinateur utilisé ayant des mots de soixante "bits"). Ce "bit" sera "1" si un ou plusieurs parmi ces enregistrements est représenté sous cette pondération de ce mot-clef; sinon, le "bit" sera "0". Dans le cas positif, il y aura un mot d'un second niveau ou chaque "bit" représentera la présence ou l'absence d'un enregistrement, toujours par ordre chronologique. L'avantage net de cette structure est d'exploiter au mieux les instructions de base de l'ordinateur, en effectuant l'union ou l'intersection "bit" à "bit" de deux mots de mémoire. Par exemple, si on veut trouver l'intersection de deux ensembles, l'intersection de leur premier niveau nous dira immédiatement s'il peut y avoir une intersection et dans quels intervalles de soixante enregistrements: la position des "bits" nous dira où dans la table, se trouvent les enregistrements correspondants, et le nombre de "bits" '1' qui précède nous dira où chercher les mots du deuxième niveau qui sont représentés par ce "bit". On voit que l'intersection de toute la banque avec un mot-clef contenant un seul enregistrement nécessitera seulement deux instructions d'intersection: une pour chaque niveau.

Les utilisateurs n'ont évidemment pas à être conscients de cette structure.

La programmation de ce système a été effectuée sur la CDC 6600 et la CDC Cyber 74 successivement, toutes deux au Centre de Calcul de L'U. de M.; elle se compose d'environ trois mille énoncés des langages "Fortran" et "Compass".

IV BILAN DU TRAVAIL FOURNI

Après plus d'un an d'utilisation, il ressort que le système a de plus en plus d'utilisateurs, tant parmi les étudiants que parmi les professeurs, secrétaires et personnes extérieures à l'Université; l'entraînement, du moins à l'utilisation conversationnelle des programmes, ne pose aucun problème.

Le meilleur test de l'efficacité des programmes a été fait sur

une banque où l'équipe IF a indexé trois mille enregistrements en utilisant plus de six-mille-cinq-cent mots-clefs différents; cette banque utilise environ 2,775,000 caractères sur disque. Une requête complexe (environ 35 opérations booléennes) coûte environ $0.80 incluant le chargement et l'initialisation du programme.