DE BERNARD LYON-I
rd du 11 Novembre 1918
69621 VILLEURBANNE

Diplôme d'Etudes Supérieures Specialisées

informatique documentaire

* MEMOIRE DE STAGE

ETUDE DES LANGAGES DOCUMENTAIRES
DANS LA BASE DE DONNEES REPARTIE
MESSIDOR

AUTEUR : Guillemette CLAUSE

DATE : Juin 1980

PLAN DU MEMOIRE DE STAGE

I Introduction

J'ai été accueillie pendant quatre mois au Centre de Documentation de l'Institut National de Recherche en Informatique et Automatique (INRIA). Ce centre de documentation mène actuellement des travaux en collaboration avec les membres du projet pilote SIRIUS, projet étudiant les systèmes d'information répartis.
Les travaux entrepris en commun permettront de constituer une base de données répartie sur le domaine de l'informatique.
L'informatique est, ici, définie au sens large, puisqu'elle comprend les sujets internes à cette science, ses applications et ses problèmes d'utilisation.

A _ Définition d'une base de données répartie.

Une base de données répartie est constituée d'un certain nombre de bases de données reliées à un même réseau de transmission. Un système informatique permet d'utiliser l'ensemble de ces bases de façon globale, comme si elles n'étaient qu'un fichier localisé sur un même site.

B _ Base de données répartie sur l'informatique : le système MESSIDOR.

Le système informatique MESSIDOR permet l'accès interrogatif aux données mises en commun dans la base répartie grâce à l'utilisation d'un langage unique défini au niveau global. Il offre la possibilité de consulter plusieurs fichiers de la base répartie soit simultanément soit successivement.

Les sources d'information qui participeront à MESSIDOR sont :

des bases de données : sur le serveur TELESYSTEMES les fichiers PASCAL du CNRS, TELEDOC du CNET, sur le serveur de l'Agence Spatiale Européenne (ESA) les fichiers NTIS du National Technical Information Service, INSPEC de l'Institution of Electrical Engineers, NASA du Centre d'Information de la NASA.

des fonds documentaires automatisés des bibliothèques de l'INRIA, de l'Institut de Programmation, du Laboratoire de Mathématiques d'Orsay.

des catalogues collectifs : ceux du groupe sectoriel de l'ADBS et des bibliothèques universitaires.

Pour atteindre ces objectifs MESSIDOR assure :

la gestion de la circulation des messages dans le réseau informatique,

la correspondance entre le langage défini au niveau global et les langages utilisés au niveau local, que ce soit le langage de commande ou le langage documentaire.

La correspondance entre langages documentaires est assurée à la fois par la vue globale des données (définition des champs documentaires) et par la normalisation des formats des données (correspondance entre les formats locaux des données et leur format au niveau global).
Chaque type de donnée doit être examiné : auteurs, dates, mots-clefs ...
Les mots-clefs n'ont pas à être normalisés. Ils forment une chaîne de caractères simples et ne soulèvent pas de problèmes de syntaxe pour l'écriture.
Cependant leur utilisation dans la base répartie pose un problème important, explicité dans le paragraphe suivant.

II Compatibilité entre langages documentaires

Dans un système documentaire les mots-clefs ou descripteurs constituent le langage documentaire privilégié pour l'interrogation.
Dans le système MESSIDOR le vocabulaire de la vue globale sera constitué de la réunion des vocabulaires locaux.
Il manquera probablement d'homogénéité en ce qui concerne le niveau de précision, la langue employée, il comportera des synonymes nombreux.
Son utilisation en recherche documentaire exigera de la part de l'utilisateur un travail délicat et l'appel à l'imagination afin de découvrir toutes les formes que

prend un même concept.
Un linguiste, vacataire à l'INRIA, étudia ce problème en 1979. Les conclusions de ses travaux sont résumées ci-dessous.

A- Etude linguistique : réconciliation de lexiques ou de thésaurus.

Cette étude linguistique mit en évidence la nécessité de structurer le vocabulaire global mais l'impossibilité de constituer un thésaurus à partir de cette réunion de mots-clefs provenant des différentes bases.
Il fut proposé de constituer au niveau global un lexique et des classes de similitude, qui réconcilieraient les lexiques (vocabulaire libre ou contrôlé) ou thésaurus des bases locales.
Cette solution promettait d'être coûteuse (comme toutes les solutions aux problèmes linguistiques) en moyens humains et informatiques.
Il fut décidé de ne pas l'étudier davantage, au moins pour quelques temps.
Nous avons cherché dans la littérature les systèmes de classification existants qui nous permettraient de structurer le vocabulaire au niveau global. Cette recherche est présentée ci-dessous.

B. Systèmes de classification automatique

La méthode de classification automatique peut s'appliquer à des documents. Elle est citée ici car elle procède, dans certains cas, à un regroupement préalable du vocabulaire (la méthode de Mr Jackson est un exemple).
La classification automatique s'applique, de manière plus générale, aux données.

1) Méthodes de classification automatique des documents

Différentes méthodes de classification automatique laissent à l'ordinateur la tâche de définir des classes ou structures des mots-clefs, celle de modifier, aussi souvent que nécessaire, la liste des termes afin d'améliorer le système.
Ces méthodes analysent mathématiquement la fréquence de coocurrence des mots les plus importants dans le document lui-même ou dans un résumé.
Des catégories sont ainsi formées. On peut alors grouper les documents selon les mots qu'ils contiennent et classer chaque nouveau document dans un des groupes préexistants.
Il est possible de réviser la structure des groupes à l'aide de nouveaux documents.
Une autre méthode, proche de la précédente, permet de calculer, à partir de la coocurrence des termes, la relation entre chaque nouveau document et ceux qui sont déjà entrés.

On peut exprimer les relations entre termes par un thésaurus interne.

Ces méthodes laissent encore à l'homme le soin de fixer le nombre de termes à utiliser pour l'indexation et la classification.

Il est possible de déterminer, en testant le programme, le nombre optimal de termes à utiliser.

D. M. Jackson dans un article de 1970 (Information Storage and Retrieval, vol. 6 pp. 187-213.) expose une autre méthode. Elle n'est pas basée sur la cooccurrence des termes dans les documents.

A partir de questions parfaitement formulées par l'utilisateur, et de l'ensemble des descriptions des documents on simule des jugements externes grâce aux fonctions de classification et de comparaison. On obtient ainsi, post-facto, une table des documents pertinents.

Les questions sont ensuite reprises et l'on construit pas à pas, une classification qui donnera les plus grands coefficients de ressemblance aux documents jugés pertinents pour chaque question.

Jackson distingue la "pseudo-classification" ainsi obtenue, des classifications obtenues par mesure de la cooccurrence des termes. Elle est en effet formée de classes de termes qui ne représentent pas les concepts des documents utilisés.

Monsieur Diday ainsi que d'autres chercheurs étudient les méthodes de classification automatique. Mais celles-ci sont orientées vers la méthode plus générale de classification des données.

2) Méthodes de classification des données.

L'analyse des données cherche à détecter des classes d'objets tel que deux objets d'une même classe soient plus proches au sens de la mesure de ressemblance que deux objets appartenant à des classes différentes. Monsieur Diday a développé dans ce but la "méthode des nuées dynamiques" et la "méthode des nuées dynamiques séquentialisée". Celles-ci font partie des techniques algorithmiques du type Hall et Ball. Elles permettent l'agrégation de groupes de points que l'on améliorera par itérations successives. Monsieur Diday distingue encore deux approches différentes de la classification automatique, mis à part l'approche par hiérarchie et l'analyse factorielle. Ce sont : la recherche de la partition qui minimise un certain critère, développée par Régnier, Ruspini, Jensen; Les techniques du type Rocchio sur lesquelles ont travaillé Bonner et Hill.

3) Bibliographie

METHODES DE CLASSIFICATION AUTOMATIQUE

day E. et coll. optimisation en classification automatique 1980
IRIA - Rocquencourt (à paraître)

ay E. Sélection typologique de paramètres. 1976
Rapport de recherche n° 188. IRIA. Laboria <02207>

day E. Automatic sequential classification of large tables 1974
Rapport de recherche n° 87 . IRIA-Laboria <02105>

day E. Optimisation en classification automatique et reconnaissance 1972
des formes. RAIRO (nov 1972) V-3 p61-96

day E. Nouvelles méthodes et nouveaux concepts en classification 1972
automatique et reconnaissance des formes.
Thèse d'Etat. 1972 - U. Paris VI. <U150 7189>

day E. Une nouvelle méthode en classification automatique et 1971
reconnaissance des formes : la méthode des nuées dynamiques.
Revue de Statistique Appliquée. Vol XIX n°2 . p 19-33

ay E. La méthode des nuées dynamiques et la reconnaissance
des formes. Cahiers de l'IRIA - Rocquencourt. 1978.

METHODES DE CLASSIFICATION AUTOMATIQUE

chevallier Y. — Classification automatique optimale sous la contrainte d'ordre total. Rapport de recherche IRIA. Laboria n°200 <02373> — 1978

— Analyse de données et informatique. document préliminaire n°2. Fontainebleau du 19 au 30 mars 1979 <H 712 15055> — 1979

Cline N., Sibson R. — The construction of hierarchic and non hierarchic classifications. Computer Journal n°11 p 177-184 — 1968

Cline N., Sibson R. — Mathematical Taxonomy 1971, Londres et New York : Wiley — 1971

Rijsbergen C.J. — Information retrieval Londres - Butterworths - 1975 <V 433 12473> — 1975

an I. C. — Les bases de la classification automatique Paris - Gauthier - Villars. 1970. <V 172 04719> — 1970

A — Linguistique automatique Rocquencourt - IRIA. 1975 <V 398 11142> — 1975

K Jones K., Kay M. — Linguistics and information science New-York, Londres - Academic Press - 1973 <V 330>9031> — 1973

K Jones K. — Automatic keyword classification for information retrieval. London - Butterworths - 1971. <V 165 5438> — 1971

K Jones K., Needham R. — Automatic term classifications and retrieval GBR - Pergamon Press - 1968 (Information Storage and Retrieval vol 4 p. 91-100)

ck Jones K., Current approaches to classification and clump-finding 1967
bon D.M. at the CLRU.
Computer Journal, 1967, vol 10 p. 29

ath P.H.A. Numerical taxinomy - the principles and practice of numerical 1973
l R.R. classification.
San Francisco - Freeman - 1973 <T 353 08225>

nack R.M A review of classification 1971
J. R. Stat. Soc. 1971, A p 351-353

ton G. The SMART retrieval system. experiments in automatic 1971
document processing.
New-Jersey. Prentice-Hall - 1971 <V 103 4571>

son D.M. The construction of retrieval environments and pseudo- 1970
classifications based on external relevance.
in Information Storage and Retrieval (Pergamon Press) 1970
vol 6 p. 187-219.

son D.M. Comparison of classifications 1969
in Numerical Taxonomy (New-York - Academic Press.)
1969 p. 91-111.

yne A.J. Some modern approaches to the classification of 1968
knowledge.
Class. Soc. Bull. 1968 vol 1 n°4 p. 12-17

io H. Automated language processing 1967
New-York - John Wiley and Son - 1967 <V211 1395>

o H. Studies on the reliability and validity of factor analy- 1965
tically derived classification categories.
in Statistical Association Methods for Mechanized Documentation. Washington - D.C. National Bureau of Standards.
1965

le L.B. — Is automatic classification a reasonable application 1965
of statistical analysis of text?
Journal of ACM. Octobre 1965 p. 473-489

le L.B. — Semantic road maps for literature searchers
Journal of ACM. octobre 1961 p. 553-578

'hom R.M. — Contribution to the theorie of clumps - I and II
or-Rhodes A.G.

4) Conclusion

Les techniques appliquées à la classification automatique des documents forment des catégories de mots par analyse de la cooccurrence des termes mais l'unité sémantique de ces catégories n'est pas assurée.
Il faudrait d'autre part, pour appliquer ces méthodes, disposer de toutes les références sur un même ordinateur ce qui n'est pas le cas pour la base répartie.
Les techniques, plus générales, de classification des données sont tournées actuellement, plus vers les applications taxonomiques que linguistiques. Là encore, il faudrait utiliser des critères linguistiques comme ceux décrits par l'étude résumée en II. A.

Il a donc été décidé de laisser en attente le problème de correspondance entre langages documentaires considéré dans son ensemble.
Nous avons limité son étude au cas particulier de la traduction des termes anglais mêlés au vocabulaire global.

C- Systèmes de traduction d'anglais en français

Des bases de données documentaires en langue anglaise seront intégrées dans le système MESSIDOR dont les bases NASA, NTIS, INSPEC de l'ESA.
Une hétérogénéité de langue sera ainsi introduite dans le vocabulaire global qui obligera, dans certains cas, à effectuer une recherche bilingue sur mots-clefs.
Les mots-clefs échangés entre les bases locales anglaises et le niveau global devraient donc être traduits, tout en gardant le terme dans sa langue originale au niveau local.
La recherche bibliographique présentée ci-dessous présente les systèmes de traduction automatique d'anglais en français, qu'ils soient français ou appartiennent à la Communauté Européenne.

1) SYSTRAN (version 1978 de la Commission des Communautés Européennes)

Le système SYSTRAN, mis au point par le Dr Toma aux USA est utilisé depuis 1969 pour la traduction automatique russe vers anglais. La Commission des Communautés Européennes en a acquis une version anglais-français en 1976. Une version améliorée a été fournie en 1978, ainsi, entre autre, qu'une version français-anglais.

a. Outils mis en œuvre

Le système est constitué d'un logiciel et de dictionnaires bilingues construits par les lexicographes ou codeurs de la Commission.
Dans le cas de langues riches en flexions, le dictionnaire ne contient que le radical des mots et sa consultation est précédée d'une analyse morphologique. Le système dispose d'une grammaire simple.

b. domaine d'application
Le dictionnaire du système est spécialisé dans le micro-domaine des sciences et technologie alimentaires et comporte 30 000 entrées.

c. résultats
SYSTRAN est actuellement utilisé par différents organismes dont la NASA, la multinationale Rank Xerox.
Evalué en 1978 par la CEE sur de nombreux critères, le système montre de nombreuses lacunes. Le texte nécessite une post-édition importante afin de corriger les erreurs portant sur les structures de phrase, les accords, le vocabulaire.

d. Perspectives
SYSTRAN devrait développer son logiciel et ses dictionnaires afin d'intéresser un public plus large. Mais l'organisation du Centre de Traduction de la CEE et le travail quotidien auquel il doit faire face limite ce développement.

Ref.. Evaluation by the EEC Commission of the "SYSTRAN" automatic translation system 1978 version. G. Van Slype. Information et Documentation n°4 p. 27-35 (May 79). FRANCE.

- Deuxième évaluation du système de traduction automatique SYSTRAN anglais-français de la Commission des Communautés Européennes.
Loll Rolling.
CCE (Luxembourg).

- Commission des Communautés Européennes
Evaluation du système de traduction automatique SYSTRAN.
Rapport n°5.
Synthèse des évaluations économique et qualitative.
Bureau Marcel Van Dijk. (Paris) 1977. 15p.

- Traduction automatique : les faits.
Loll Rolling
CCE (Luxembourg). 1978. 9p.

2) Système de traduction automatique étudié par le GETA

GETA (Groupe d'Etude pour la Traduction Automatique)
Université de Grenoble I
B.P. 53 - 38 041
Responsable : M Vauquois B.

Le GETA réalise des modèles d'analyse, de transfert et de génération sur diverses langues : russe, français, anglais, portugais.

a. Perspectives

Le GETA associe ses efforts à ceux de plusieurs universités européennes pour élaborer un système de traduction européen unique sous la responsabilité de la Commission. On devrait expérimenter ce système en 1982.
Le GETA va continuer dans les prochaines années, la mise au point d'un système de traduction automatisé de 2e génération. Ce système, considérant la phrase comme un tout, analysera la structure de chacune; puis les mots, groupes de mots et structures seront traduits dans la langue cible. Dans une troisième étape ou étape de synthèse, le texte en langue cible sera généré.
Le futur système réalisera la séparation entre les programmes et les données linguistiques grâce à des programmes universels. Ceux-ci fourniront le langage d'écriture des dictionnaires et des grammaires et généreront des programmes de traduction lorsqu'on leur donnera des données

linguistiques.
Enfin ces nouveaux systèmes de traduction permettront la correction interactive.

Ref. - Automatic translation and Computer-assisted translation. S. Heriard Dubreuil.
Informatique et Gestion (France) n°107 p. 49-54 (Juin-Juillet

- La traduction automatique à Grenoble - B. Vauquois.
Paris, Dunod, 1975.

- L'évolution des logiciels et des modèles linguistiques pour la traduction automatisée - B. Vauquois.
GETA - Grenoble.

3) Système ROBRA utilisé par le CRAL

Groupe de traduction automatique du Centre de Recherches et d'Applications linguistiques (CRAL) de l'Université Nancy II.
B.P. 33-97 - 54000 - Nancy Cedex. tel (83) 96-10-11
Directeur Mme H. Naïs

Ce groupe expérimente actuellement une chaine de traduction entièrement automatique anglais-français sous la responsabilité de Mr Bourquin.

a. Outils mis en oeuvre

Le logiciel utilisé, ROBRA, conçu par le GETA de Grenoble est implanté sur IBM 360.

b. Domaine d'application

Les textes utilisés sont du micro-domaine de la métallurgie. Le dictionnaire automatisé comporte 2000 entrées.

c. Les étapes du programme de traduction

Le logiciel impose un passage par les différents systèmes :

ATEF analyse morpho-syntaxique

CETA 1 analyse syntaxique

TRANS transfert lexical

CETA 2, 3 transfert et génération syntaxique

SYGMOR Génération morphologique

La dernière étape est réalisé par l'équipe de Grenoble.

d. Les résultats

Cette chaîne de traduction est entièrement automatique. Pratiquement aucune préparation du texte n'est demandée. Les 6 étapes du traitement, énumérées plus haut, traduisent la phrase anglaise en une phrase française souvent compréhensible bien qu'elle contienne encore de nombreuses erreurs.

Ces erreurs pourront dans l'avenir être évitées en complétant le dictionnaire de génération morphologique

du GETA (cause d'erreurs mineures), en complétant également le programme de traitement afin de diminuer les fautes d'analyse.
La prise en compte de l'aspect sémantique améliorerait la qualité de la traduction.
Le système de traduction, notamment au niveau de l'analyse, n'a pu faire abstraction de la langue d'arrivée.
Dans les premières expériences réalisées, la traduction d'un mot prend 2/3 de seconde.

c. Perspectives

Selon le but poursuivi dans les prochaines années, on pourra soit affiner l'analyse syntaxique, en la complétant par une analyse sémantico-syntaxique pour obtenir une traduction plus fine mais plus coûteuse, soit s'orienter vers une aide à la traduction et se limiter à une grammaire plus simple, optimisée et appliquée à un unique couple de langues.

réf. - Groupe de Recherches en Traduction Automatique.
Rapport Scientifique 1979 pp. 11-39.
Centre de Recherches et d'Applications Linguistiques.
Université de Nancy II.

4) TITUS

TITUS est le système de traduction de l'Institut du Textile de France. La version TITUS II fonctionne actuellement, la version TITUS III devait être lancée en 1979-80. TITUS est un système opérationnel de traduction. Il donne des résultats satisfaisants mais impose une lourde préparation du texte à traduire afin de simplifier sa syntaxe.

ref. - The development of the Titus four-language automatic translation method. S. Steiff. Information et Documentation (France) n°4 p20-26 (May 1979).

- Constitution de lexiques multilingues pour traduction automatique. Problèmes posés dans le cas de TITUS. Hubert J. M. Banque des mots (France) - 1979 n°16 pp. 187-196.

5) EURODICAUTOM (banque de terminologie multilingue)
Système du Bureau de la Terminologie, Division de la Traduction, Commission des Communautés Européennes (Luxembourg)
responsable J. A. Bachrach
J. Goetschalckx

La Commission des Communautés Européennes a organisé la banque de données terminologiques en six

langues, EURODICAUTOM, afin d'aider les traducteurs de la Communauté dans leur travail.
EURODICAUTOM résulte de la fusion, en 1969, du système DICAUTOM de l'Université de Bruxelles et du système EUROTERM du Bureau de Terminologie du Marché Commun de Bruxelles.

a- But du système

EURODICAUTOM met à la disposition des utilisateurs des dictionnaires multilingues basés sur une conception phraséologique.
Il fournit la traduction, dans les 6 langues de la Communauté Européenne, des unitermes, des expressions avec éventuellement une définition ou le contexte du ou des termes, ainsi qu'une série d'informations utiles au traducteur.
Celui-ci peut choisir le domaine, parmi tous les sujets couverts par le dictionnaire, où il désire trouver la vedette qu'il recherche.

b- Outils mis en oeuvre

Les fichiers de EURODICAUTOM (fichiers principal, transversal, inversé) sont implantés sur un ordinateur IBM.
Le nombre de mots enregistrés dans la banque est de 600 000.
L'interrogation a lieu en conversationnel ou en différé.

c - Résultats

EURODICAUTOM apporte une aide utile aux traducteurs. Mais quelques inconvénients demeurent : ainsi l'apparition des réponses dans un ordre non satisfaisant amène une perte de temps. Le système possède d'autres lacunes. Ainsi il ne distingue pas si les contextes se correspondent ou non.

d. Perspectives

EURODICAUTOM sera amélioré grâce à une réduction des inconvénients précédemment cités, et grâce à une amélioration du contrôle de la saisie.
Il est nécessaire d'arriver à un accord entre fabricants et utilisateurs de terminologie et de définir les formats informatiques et terminologiques utilisés.

réf. - EURODICAUTOM - J. Goetschalckx pp. 71-75
Translating and the Computer (Londres) - 14 Nov 1978
Amsterdam, North-Holland, 1979.

- EURODICAUTOM Possibilités et limites d'un système automatisé d'aide à la traduction. A. Reichling
V Congrès International de Linguistique Appliquée.
CEE - Août 1978.

6) bibliographie

ANQUES DE TERMINOLOGIE MULTILINGUE AUTOMATISEES

GENERALITE

- Multilingual terminology banks here and abroad ; online terminology dessemination as an aid to translation · S. E. Morton (Carnegie-Mellon University, Pittsburg, PA, USA)
The Information Age in Prospective. Proceedings of the Asis Annual Meeting 1978, New-York, USA, 13-17 Nov 1978.
(White Plains, NY, USA : Knowledge Industry Publications Inc 78 p 242-44)

- Utilisation électronique de la banque de mots . J. P. Vinay
meta, vol 15, n°1-2 March-June 1971, pp 95-104.

- Les Banques Automatisés de Terminologies Multilingues et les Organismes de Normalisation. M. Paré.
in Infoterm Series 3, Premier Symposium d'Infoterm.
Coopération Internationale en Terminologie
Munchen ; Verlag Dokumentation, 1976.

- "Comment les banques automatisés de terminologie multilingues peuvent assister les organismes de normalisation." M. Paré.
Dans Colloque sur la coopération internationale en Terminologie. Avril 1975.

- On-line Generation of Terminological Objects in language translation: an aid in Terminology Processing. Lippman E. O.
Rapport de recherche IBM Sept 1975 n° 5624.

SYSTEMES DE TRADUCTION AUTOMATISEE

GENERALITES

- La traduction assistée par ordinateur : des banques de terminologie aux systèmes interactifs de traduction.
  C. Carstis Greenfield, O. Serain.
  Iria - 1977. (INRIA : 16 243 V530)

- Translating and Computer
  Amsterdam Netherlands ; North-Holland (1979)
  Conférences Fenues à Londres le 14 Nov 1978.
  (INRIA : 16129 V527)

- Handbook of machine translation and machine-aided translation
  H. E. Bruderer.
  . Amsterdam, 1978.

- Progress in documentation, machine translation and machine-aided translation, Hutchins W. J.
  in Journal of Documentation, GBR 1978, vol 34, n° 2 pp 119-159.

- Perspectives d'avenir de la traduction automatique.
  Brinckmann K. H.
  Meta. Canada - 1979. 24, 3 pp 315-325.

- Automatic translation and computer-assisted translation.
  S. Heriard Dubreuil
  Informatique et Gestion (France) n° 107 pp 49-54 (Ja, Juil 79)

- A Survey of approaches and issues in machine-assisted translation systems. W. W Zachary (Systems and Operations Analysis div. of Analytics, Willow Grove, PA, USA)
  Computer and the Humanities vol 13, 1 pp 17-28 (Jan - mars 79)

- Traduction automatisée: du rêve à la réalité.
  Bureau Gestion (France) 1979, 2, 14 p 25.

EXEMPLES DE SYSTEMES DE TRADUCTION AUTOMATISEE

- An artificial intelligence approach to machine translation.
  Y. Wilks
  Stanford University - 1972. (STAN-CS-264-72) - 46p.

- An intelligent analyzer and understander for English. Y. Wilks.
  communication ACM 1975. May 75, vol 18 n°5.

- Computer analysis of basic english as a first step in machine translation.
  J. R. Straub (Univ. of Arizona, Tucson, AZ, USA)
  C.A. Rogers.
  in book: Progress in cybernetics and systems research, Vol V - R. Trappl, F Hanika, F. R Richler - p 482-90 - England - Wiley - 1979.

- Getting started in a natural language translation project.
  R. Earp, J. Chernika (System Sci. Univ of West Florida, Gainsville, FL, USA)
  J. Educ. Data Proces (U.S.A.) vol 15, 2 pp 25-34 - (1978).

SYSTEMES D'INFORMATION MULTILINGUES

- Multilingual information systems : some criteria for the choice of specific techniques. C.P.R. Dubois.
(International Coffee organization, London, WIP 4DD, GBR)
J. Inf. Sci., NLD (79), 1, 1 pp 5-12.

- Experiments in multilingual information retrieval.
G. Salton. Ithaca Cornell University - 1972 (TR-72-154)

- Overcoming the language barrier. — — — — —
3rd European Congress of Information System and networks.
Munchen : Verlag Documentation - 1977.

- Zum Stand der informationserschliessung in den I and D-Systemen der Bundesrepublik Deutschland. (Situation de l'exploitation de l'information dans les systèmes I et D de la République fédérale allemande)
SCHÖN (J.)
Kooperation in der Klassifikation 1. Gesellschaft fur Klassifikation, (1978) pp 96-110.
(Résumés multilingues, thésaurus et résumés traduits automatiquement)

7) Conclusion

Les systèmes de traduction d'anglais en français, mis au point en France ou par la Communauté Européenne sont peu nombreux. On en distingue deux sortes.
D'une part les systèmes automatiques d'aide à la traduction qui se limitent pratiquement à un dictionnaire automatisé (type EURODICAUTOM),
D'autre part les systèmes de traduction automatisée. Dans cette catégorie il en existe très peu de réellement fonctionnels. Ils nécessitent en général une pré-édition (TITUS) ou post-édition (SYSTRAN) importante de la phrase. Le système développé par le CRAL de Nancy est intéressant car d'un bout à l'autre de la chaîne, l'homme n'a pratiquement pas à intervenir.
Mais ce système, comme ceux énumérés dans la bibliographie n'est pas encore réellement fiable. Tous ces systèmes occupent d'importantes places dans la mémoire de l'ordinateur (le système de traduction du CRAL occupe un million d'octets de mémoire virtuelle).
La traduction automatique paraît donc difficilement utilisable dans le cas de Messidor puisqu'il est prévu de localiser la majorité des programmes de gestion du terminal, de traduction des formats et autres programmes de traitement des données, sur des micro-ordinateurs associés aux terminaux. Or la taille mémoire d'un micro-ordinateur n'excède pas 65 000 octets.

Il fut décidé qu'au cours de la deuxième partie de notre stage serait défini un outil pratique destiné à faciliter l'usage de la base répartie sur l'informatique. Il s'agissait d'informer l'utilisateur occasionnel sur les bases disponibles dans Messidor et leur contenu sémantique.

III Formalisation du contenu sémantique des bases de données.

Ce travail a été accompli pour les bases servies sur TELESYSTEMES et sur l'ESA.
Les bases couvrant convenablement l'un des domaines d'intérêt défini par Messidor (l'informatique, ses applications, les problèmes de son utilisation) ont été sélectionnées.
Leur structure en classes sémantiques ont été étudiées afin d'identifier les sous-fichiers homogènes qui présentaient un intérêt.
Dans une deuxième étape ont été précisés d'une part les critères d'accès et d'autre part les critères de choix pour chaque base ou sous-ensemble de base sélectionné.

Ce travail a été réalisé après que nous ayons passé quelques jours au Centre d'Etudes et d'Expérimentation des Systèmes d'Information (CEESI).

Ce séjour nous a permis d'étudier les méthodes mises en oeuvre par la cellule documentaire de ce centre pour créer leur propre base de données REDOSI.
Le double intérêt de cette étude résidait d'une part dans le domaine couvert par cette base, c'est-à-dire les applications de l'informatique, d'autre part dans les méthodes qui seraient employées pour créer des liens entre les mots-clefs.

LA BASE DE DONNEES REDOSI DU CEESI

t : Le Centre d'Etudes et d'Expérimentation des Systèmes d'Information (CEESI) crée actuellement la base de donnée bibliographique REDOSI. Elle sera composée des documents choisis par les différentes équipes ou "programmes" du CEESI et stockés en général dans ce Centre.
Les documents traitent du développement de l'information dans des domaines variés tel que l'information foncière, l'information sur les équipements collectifs, sur le marché rural, sur les données sociales ; l'information pour le milieu urbain, pour les entreprises, pour le grand public ; les catalogues d'information et l'emploi à distance de l'information.

thode : Chaque document est analysé par un documentaliste qui remplit ensuite le bordereau de saisie pour le document.
Les différents champs de ce bordereau contiennent les références bibliographiques précises, le mode d'accès au document, sa localisation, un résumé et éventuellement des mots composés caractéristiques ajoutés par l'indexeur.
Sur le bordereau, un certain nombre de mots composés portant un sens riche, sont soulignés dans le titre français ou anglais

ou dans le résumé et à partir de maintenant systématiquement réduits au singulier.
Ces mots composés constitueront le sous-lexique 6, dans lequel sont entrés toutes les formes que prend l'expression lorsqu'on permute un terme à la fois.
Ces formes permutées sont créées par le documentaliste sur bordereau de saisie avec un point précisant la place réelle du terme inversé dans le mot composé.
Enfin le documentaliste définit une relation de synonymie entre les formes permutées de la même expression.
D'autres relations seront définies, toujours par des méthodes manuelles.

Le sous-lexique uniterme est construit automatiquement par l'ordinateur. Le logiciel utilisé permet d'éliminer du texte et du résumé les mots contenus dans une liste de "mots vides". Les termes présents dans le titre ou le résumé et qui ne sont pas des mots vides entrent dans le sous-lexique "uniterme".

D'autres lexiques : auteurs, pays, nom de système constituent des moyens de recherche documentaire.

Le sous-lexique des mots composés sera utilisé d'une façon

différente.

Il ne pourra pas servir à la récupération des documents par l'intermédiaire de fichiers inverses, contrairement aux unitermes aux noms d'auteurs, de pays, de systèmes.

mais, il constituera une aide à la recherche en permettant à l'utilisateur de connaître les différents contextes sémantiques dans lequel peut se trouver un même terme.

Actuel : Depuis neuf mois (juillet 1979), l'équipe des documentalistes du CEESI travaille à temps complet sur le projet.

Les premiers mois ont permis, grâce à des essais successifs, de choisir les champs à utiliser et de définir leur contenu afin de répondre le mieux possible aux besoins des équipes du CEESI et d'autres futurs utilisateurs.

Des méthodes précises de rédaction des résumés ont été définies pour chaque type de document (environ une dizaine). Elles assurent une homogénéité entre le travail des différents documentalistes.

Des bordereaux de saisie des documents ont été créés.

L'équipe, tout en préparant les documents nouveaux afin de les rentrer dans la base, traite petit à petit le fond ancien.

Un travail important consiste à vérifier tout ce qui est entré dans la base : les documents saisis et listés, les sous-lexiques unitermes et mots composés, afin d'obtenir un produit très fiable.

Mille six cents documents sont déjà entrés dans la base et l'on note une forte diminution du nombre moyen par document, d'unitermes ou de mots composés entrant dans les sous-lexiques.

Cette tendance s'accentue avec l'augmentation du nombre de documents enregistrés.

Les trois cent cinquante documents ont généré 3000 termes dans les sous-lexiques unitermes et mots composés. Actuellement ces sous-lexiques comportent 16 000 termes provenant de 1600 documents.

<u>pectives</u> : Dans les prochaines semaines, lorsque le lexique aura été soigneusement vérifié, les premiers travaux linguistiques pourront commencer.

Ainsi, les adjectifs de nationalité seront réduits aux noms de pays correspondants. On rapprochera les substantifs des verbes, on mettra en évidence les racines ainsi que les contraires des mots.

On fera, éventuellement, appel au logiciel PIAF, afin de réduire les termes au singulier et d'extraire la racine des mots composant les sous-lexiques unitermes et mots composés.

L'Equipe du CEESI envisage de créer alors (d'ici deux ou trois mois environ) des liens autre que la synonymie, dans l'ensemble des uniterms et mots composés.
Ces liens ne définiront pas des relations aussi structurées qu'un thésaurus mais leur affichage sur le terminal facilitera la recherche documentaire.

L'Equipe documentaliste du CEESI poursuit donc actuellement la création de la base de donnée bibliographique RENOSI implantée sur le serveur Télésystèmes.
Elle contiendra les références et analyses de documents ayant, en général, rapport avec les systèmes d'information et leurs applications dans tous les domaines.
Elle sera ouverte aux usagers extérieurs. Deux mille documents sont déjà entrés dans la base grâce à des méthodes de saisie maintenant bien définies.
Tandis que la constitution de la base se poursuit, les sous-lexiques de mots clefs sont mis au point grâce à différentes corrections et réductions de plusieurs termes à un seul.

La création ultérieure de relations sémantiques entre mots clefs améliorera la recherche documentaire.
Ces relations n'ont pas encore été choisies et définies par les documentalistes mais il est clair qu'elles auront pour but d'aider l'usager à formuler sa question en lui indiquant un environnement sémantique plus large que celui des synonymes. Mais il ne semble pas être question d'une structure aussi bien définie que celle d'un thésaurus ou d'une classification.

Conclusion : application des méthodes décrites au système Messidor

La base de donnée REDOSI présente les points particuliers suivants:
Elle couvre le domaine des applications de l'informatique;
Le lexique des mots-clefs (unitermes) et le fichier inverse associé sont constitués automatiquement;
des aides à l'utilisation seront mises en place manuellement.
Ce sont: l'affichage des mots composés ou contexte des unitermes, la définition de liens sémantiques plus larges qu'une simple synonymie.
Ces méthodes offrent une possibilité de traiter le vocabulaire.

Elles pourraient donc être employées dans une base locale du système MESSIDOR.

Utilisées par toutes les bases de MESSIDOR, elle donnerait naissance à une liste d'unitermes de milliers de mots. Des aides à l'utilisation prévues, mais encore non précisées par le CEESI, seraient indispensables mais probablement difficiles à réaliser.

D'autre part les méthodes du CEESI ne sont pas conçues, et ne présenteront probablement pas d'intérêt pour caractériser par quelques descripteurs, les domaines décrits par les bases de données.

Il a été décidé, en conséquence, de décrire dans un premier temps le contenu sémantique des fichiers documentaires par la méthode traditionnelle d'indexation manuelle.

B - Choix d'une méthode : l'indexation manuelle

Le travail d'indexation nous a permis d'attribuer à chaque fichier un petit nombre de descripteurs, regroupés ensuite dans un court lexique.
Il fut nécessaire, dans une première phase, d'analyser les domaines couverts par les bases.

1) Analyse du contenu sémantique des fichiers

Divers outils furent mis en oeuvre lors de cette phase d'analyse :

- le manuel de l'utilisateur, constitué par chaque serveur, comporte une description des bases disponibles sous forme de la liste des domaines couverts.
Le nom de l'organisme qui constitue le fichier est important car il révèle le thème central de celui-ci.

- La classification par sujets est parfois associée à ces renseignements. On la trouve également sur les bulletins analytiques lorsqu'ils existent. Elle est précieuse car c'est en fait une indexation à plusieurs niveaux de la base. Elle aide aussi à situer l'importance relative des chapitres (selon les sous-divisions qu'ils contiennent). Cette indication pourrait induire en erreur si elle n'était confirmée par d'autres renseignements.

On peut, par exemple, parcourir un certain nombre de bulletins analytiques, observer le nombre de références par sujet, la fréquence d'apparition des mots-clefs. Enfin l'interrogation en conversationnel à l'aide de mots-clefs et la recherche sur le lexique par voisinage alphabétique seront utilisés surtout si les "outils papier" précédents ne sont pas disponibles.

Le contenu sémantique de la base, précisé par ces divers moyens est alors décrit par des mots-clefs.

2) Outils de description du contenu sémantique

Les notions essentielles, dégagées par l'analyse, sont soit gardées dans leur expression originale, soit transformées, soit supprimées, soit remplacées par un terme plus générique après consultation des

- dictionnaires spécialisés ou encyclopédies (Encyclopédie Internationale des Sciences et techniques, Presses de la Cité; Encyclopaedic Dictionary of Mathématics for Engineers and Applied Scientists, Sneddon; Encyclopedic of Computer Science, Van Nostrand, Reinhold)
- classifications existantes en informatique (par exemple le Glossaire Européen pour la Recherche en Informatique et Automatique GINDEX, la classification du CNRS, la classification INSPEC...

- sommaires ou tables des matières de revues spécialisées en informatique ou mathématiques (Computing Reviews, International Computer Bibliography ...)
- thésaurus (INSPEC thesaurus, Thesaurus of Engineering and Scientific Terms de l'Engineers Joint Council)

Nous nous sommes beaucoup référés au Macro-Thésaurus des Sciences et Techniques du BNIST car sa polyvalence permet de mieux définir les domaines externes à l'informatique proprement dite.
Grâce à ces différents documents, mis à ma disposition par le Centre de documentation de l'INRIA, j'ai réalisé l'indexation des fichiers choisis sur TELESYSTEMES et sur l'ESA.

3) La démarche d'indexation

Elle a été déterminée par le but poursuivi : offrir à l'utilisateur de la base de données réparti sur l'informatique un choix de fichiers ou sous-ensemble de fichiers pouvant l'intéresser, accompagné d'un système lui permettant de sélectionner rapidement ceux qui répondront le mieux à son problème.

L'indexation a donc été faite dans l'optique particulière des futurs utilisateurs, c'est-à-dire des scientifiques spécialistes en informatique, des spécialistes en application

ou en utilisation informatique.
N'ont été indexés que les sujets susceptibles de les intéresser de façon particulière et représentés par un nombre important de documents.
Pour satisfaire un besoin d'information rapide sur chaque base, le nombre de descripteurs a été volontairement limité. Ceux-ci indiquent les macro-domaines représentés dans la base. Cependant plus le sujet décrit est central dans les préoccupations de l'utilisateur de Messidor, plus les mots-clefs seront nombreux et précis.

L'ensemble des descripteurs est présenté dans un lexique et dans l'index alphabétique.
Le vocabulaire n'est pas hiérarchisé, car si le degré de précision est hétérogène, selon les disciplines, il est à peu près homogène dans un même sujet.
La forme du vocabulaire n'est pas fixée. Dans le lexique on trouve aussi bien des uniterme que des groupes de mots coordonnés par des prépositions ou juxtaposés avec des parenthèses.
Dans un premier temps, nous avons préféré cette hétérogénéité à des ambiguïtés possibles sur les termes.

L'index alphabétique des descripteurs constitue un premier outil très simple pour permettre un choix rapide parmi les bases de données. Sa visualisation est rapide.

4) - INDEX ALPHABETIQUE DES MOTS-CLEFS .

magnétisme	(P/I4.I40)-(8A)
matériaux semi-conducteurs	(P.I4.I45)
mathématiques	(I)
mathématiques de l'informatique	(P/I4)-(P/I4.II0)-(8C)
mécanique	(P/I4.I30)-(8A)
médecine	(P/I4.3I0)
métrologie	(P/I4.I45)
optique	(P/I4.I30)-(8A)
optique électronique	(P/I4.I45)
physique	(E)-(T)-(P/I4)-(P/I4.I30)-(8A)
physique des particules	(8A)
probabilités	(P/I4.II0)
propriétés électriques des matériaux	(P/I4.I45)
psychologie	(P/I4.390)
psycho-pathologie	(P/I4.390)
rapport de recherche	(6)
reprographie	(P/I4.I0I)
sciences de l'espace	(I)
stastistiques	(P/I4.II0)
tehnique nucléaire	(E)
télécommunications	(T)-(P/I4)-(P/I4.I45)-(8B)
théorie de la commande	(8C)
théorie des systèmes	(P/I4.II0)-(8C)

Code des bases :

P	PASCAL
T	TELEDOC
E	EDF
1	NASA
4	COMPENDEX
6	NTIS
8	INSPEC { 8A section A d'INSPEC ; 8B " B ; 8C " C
14	PASCAL § 14.XXX section XXX de PASCAL

5) perspectives

Le lexique et l'index alphabétique des mots-clefs ne sont pas définitifs puisque d'autres bases devront être décrites par la suite. On pourra alors envisager de donner une forme plus précise aux termes utilisés.
On pourrait envisager de développer le vocabulaire afin d'aider à situer un sujet très précis dans le domaine qui lui correspond. Mais un tel vocabulaire ne comprendrait pas moins de 500 à 1000 mots (le nôtre en compte une 50ne
L'utilisateur pourrait poser sa question une première fois sous une forme large afin de sélectionner la ou les "bases pertinentes" puis adresser sa question au sous-ensemble de bases choisi lors de la première étape.
Mais le nombre actuel de bases participant au système Mesidor ne justifie sûrement pas la lourde mise en oeuvre d'un tel outil.

C - Application aux bases de TELESYSTEMES

TELESYSTEMES propose actuellement un accès en conversationnel à 13 bases de données.
Compte tenu de notre centre d'intérêt, nous en avons retenu 3 :

la Base PASCAL fournie par le CNRS
la Base TELEDOC fournie par le CNET
la Base EDF fournie par l'Electricité de France

1) Choix d'une base

Le choix d'une base est effectué dès que la connection avec le logiciel MISTRAL a été obtenue et peut être modifiée en cours de session par la procédure :

.. BA △ [nom de la base]

(△ figure un espace)

2) Choix d'une sous-partie dans la base

La structure des bases et le logiciel MISTRAL ne permettent pas de délimiter dans chaque base des sous-parties indépendantes. Chaque opération élémentaire de recherche porte donc sur l'ensemble du fichier interrogé.
Il est cependant possible d'utiliser, dans certaines bases, l'information contenue dans une classification par matières.

C'est le cas des bases PASCAL et TELEDOC. Chaque document enregistré comporte un champs "chapitre" ou "code de classement".
Un fichier inverse permet de sélectionner des documents selon leur code de classement.
Mais il s'agit là d'une démarche de recherche documentaire pour laquelle l'ordinateur doit parcourir l'ensemble du fichier ou sous-fichier inverse.
On peut ensuite réaliser l'intersection entre le set ainsi obtenu et un autre ensemble de documents sélectionnés. A chaque set il faudra réitérer l'intersection entre la question posée (mot-clef, étape de recherche, ou combinaison de ces éléments ...) et le tri sur le code de classement.

Exemple sur PASCAL

```
?165? /CH ET 6

  TERME MULTISENS 165?:  2
       RESULTAT  50337
  *7*  RESULTAT    561
  PROCEDURE, OU ETAPE DE RECHERCHE  8

?(INFORMATIQUE OU ORDINATEUR?) ET 110I /CH

  TERME MULTISENS INFORMATIQUE:  2
       RESULTAT  13358
  TERME MULTISENS ORDINATEUR?:  3
       RESULTAT  27621
  *4*  RESULTAT     69
  PROCEDURE, OU ETAPE DE RECHERCHE  5
```

3) Etude par base

a- BASE PASCAL

Cette base de donnée à vocation pluridisciplinaire est composée de 50 sections qui définissent 50 domaines différents des sciences et techniques.

Code de classement et recherche par fichier inverse

Le plan de classement PASCAL codifie chaque section par trois chiffres. Le caractère alphanumérique qui suit précise le chapitre.
D'autres caractères viennent s'ajouter lorsque l'on descend dans les niveaux inférieurs du plan de classement mais seuls sont pris en considération dans le sous-fichier inverse des codes de classement (CH) les 4 premiers caractères.

Séléction d'un chapitre de section

Le logiciel QUESTEL permet donc de sélectionner globalement un chapitre par exemple "Sciences économiques et problèmes de gestion" 110I en utilisant la procédure suivante:

" 110I $_{\Delta}$ /CH "

(le suffixe CH précise le sous-lexique consulté tandis que le lexique n'a pas été précisé en préfixe puisque c'est le lexique implicite (BI)).

Sélection d'une section

Il semblerait à priori possible de sélectionner tous les documents d'une section de la base PASCAL par exemple la section 130 en utilisant le masque "?" ou la troncature "+" sur le quatrième caractère. Mais on se heurte alors, pour certaines sections, à un problème de débordement de la zone de travail car le nombre de références est trop important.
C'est le cas pour les sections:

110 (analyse numérique, informatique, automatique, statistique et probabilités, recherche opérationnelle, gestion, économie)

130 (Physique mathématique, optique, acoustique, mécanique, chaleur)

145 (Electronique)

Voici en exemple la section 130:

```
?130+ /CH

**ER 37 DEBORDEMENT : FICHIER DE MANOEUVRE
PROCEDURE, OU ETAPE DE RECHERCHE  5
```

Les sections 101 "information, documentation"
140 "électrotechnique"
peuvent être sélectionnées mais le fichier inverse n'a pas été conçu pour sélectionner des sections.

Remarque: L'emploi de la troncature à droite provoque l'affichage par QUESTEL des termes générés qu'il faut ensuite sélectionner par la commande "S $_{\Delta}$ TT"

Il est conseillé d'éviter cette perte de temps par l'emploi de l'option de traitement automatique des multisens ?..OP MS AU

b. BASE TELEDOC

Sur cette base sont chargées les analyses du bulletin signalétique des télécommunications. Les documents indexés peuvent traiter de sujets variés mais l'intérêt central de la base reste les télécommunications.

La classification

Une classification existe sur la base, dont le sommaire figure en tête du bulletin analytique. Mais elle est imparfaite, malgré les remaniements partiels dont elle a été l'objet.
Ainsi le chapitre 10.115 "machines mathématiques-informatique" a du être divisé, redivisé en sous-parties par suite du développement qu'a pris cette science.
Il est, par ailleurs, fréquent que les documents informatiques soient classés dans d'autres chapitres.

Le code de classification

Le code de classement se compose de deux chiffres, d'un point puis d'un certain nombre de chiffres qui augmente en descendant dans la hiérarchie. ex "10.1020"

Le fichier inverse CC

Les codes de classement sont interrogeables en conversationnel grâce au fichier inverse CC. La procédure est:

```
?/CC 10.115

  *1*  RESULTAT     105
PROCEDURE: DU ETAPE DE RECHERCHE  2
```

Cette procédure est à utiliser avec prudence. La longueur des codes n'est pas homogène. De plus une catégorie par exemple 10.102 ne contient pas les sous-catégories par exemple 10.1020. L'emploi de la troncature "+" est donc nécessaire. ex /CC␣10.102+

C- Base EDF

La base EDF, constitué par l'Electricité de France, comprend environ 200.000 documents. Comme dans la base TELEDOC, on trouve des documents sur les sujets les plus variés mais la majorité des références concernent l'Énergie et les sciences qui lui sont associées. Le bulletin signalétique associé est "Documentation Technique EDF".

La classification

Il existe une classification par domaine dont les grandes lignes sont données dans la table idéologique, en tête de la Documentation Technique. Elle ne peut servir actuellement à interroger. Mais il est prévu, dans un proche avenir, de charger les documents dans la base avec un code de classification à 7 chiffres.

BASE	CLASSIFICATION PAR MATIERE	SELECTION PAR LE CODE MATIERE		
		QUESTION	FICHIER INVERSE	REMARQUES
PASCAL	Plan de classement PASCAL	XXXY$_{\Delta}$/CH	(CH) sous-fichier du fichier inverse implicite (BI)	x Sélection d'une section exe 110+ pose des problèmes de débordement de fichier (110+, 130+, 145+) des problèmes de temps de réponse importants x Sélection d'un chapitre exe 110A possible : c'est l'utilisation normale du sous-fichier CH.
TELEDOC	Classification des Télécommunications	/CC$_{\Delta}$XX[.(X)$^{n\,pos}$]	fichier inverse (CC)	x Sélection à n'importe quel niveau grâce l'emploi de la troncature à droite. ex 10.1+ (Il existe 14 classes XX.)
EDF	Classification de l'EDF par domaines	code à 7 caractères	n'existe pas	

note X : caractère numérique

Y : caractère alphanumérique

[ ] : facultatif

INDEXATION DES BASES SELECTIONNEES SUR TELESYSTEMES

BASE	MOTS-CLEFS
PASCAL (P)	Information Documentation Informatique Automatique Analyse numérique Mathématiques de l'informatique Physique Electricité Electronique Electro-technique Télécommunications
TELEDOC (T)	Télécommunications Electronique Physique
E.D.F. (E)	Energie Technique nucleaire Physique Electricité Electronique Automatisme Electro-technique

5) Conclusion sur le serveur TELESYSTEMES

L'utilisation du fichier inverse "code de classement" ou sous-fichier "chapitre" est le seul moyen disponible pour choisir un domaine d'intérêt.
Cela ne simplifie pas le travail de l'ordinateur. L'ensemble des codes des documents contenus dans le domaine d'intérêt choisi doit en effet être transféré dans la zone de travail où s'opère l'intersection avec la question posée.
Il apparait donc préférable d'utiliser les fichiers de TELESYSTEMES tel qu'ils ont été définis par les serveurs, sans espérer accéder à des sous-fichiers dont la couverture serait moins large.
Nous avons donc caractérisé les fichiers EDF et TELEDOC, chacun considéré dans son ensemble, par un groupe de mots-clefs. La base PASCAL a également été indexée globalement.

D- Application aux bases de l'ESA

L'ESA (Agence Spatiale Européenne) propose l'accès à environ 20 bases de données.
Cinq d'entre elles présentent un intérêt particulier pour nous puisqu'elles traitent de façon développée de l'informatique ou des domaines qui lui sont étroitement liés. Ce sont les bases.

NASA (1) fournisseur : NASA Scientific and Technical Information office (Washington)
COMPENDEX (4) " Engineering Index Incorporation (New-York)
NTIS (6) " National Technical Information Service (Springfield)
INSPEC (8) " Institution of Electrical Engineers (Londres)
PASCAL (14) " CNRS (Paris)

Citons également la base de donnée factuelle sur les composants électroniques ELECOMPS (5) fournisseur: ESA (Frascati).

1) Choix de la Base

L'utilisateur, une fois connecté à l'ESA, choisit la base qu'il désire interroger par la procédure:

BEGIN[n° de la base]

exemple pour INSPEC (8):

? BEGIN8

remarque : le langage utilisé dans cette étude sur l'ESA est le langage RECON, appelé par la procédure :

```
◆
? ..SET SESAME OFF
..SET SESAME OFF ACCEPTED
```

2) Choix de la sous-partie d'une base

Il existe dans le logiciel RECON de l'ESA deux méthodes pour aborder un domaine d'intérêt ou un sujet large.

a. Recherche sur le fichier inverse des codes de classement

A chacun des cinq fichiers documentaires cités plus haut est associé une classification hiérarchique spécifique. Le champs documentaire qui contient le code matière porte des noms divers selon le fichier considéré ("CC", "CA", "CF") mais la procédure d'interrogation est toujours la même.

S cc = [code de classement]

ou #cc = [" "]

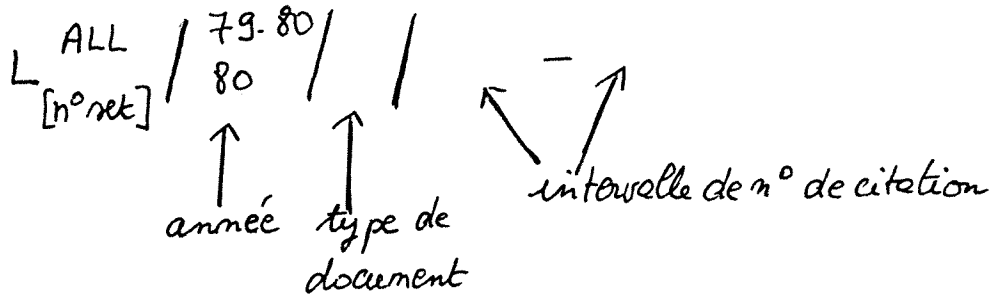
L'intersection de l'ensemble des citations récupérées par cette procédure avec un autre ensemble de références (ou set) permet de retenir dans ce dernier, seulement les documents indexés dans le domaine d'intérêt choisi. Dans le logiciel RECON les opérations logiques (ET, OU, SAUF) n'interviennent qu'entre groupes de document déjà sélectionnés et désignés par leur n° de set.

exemple :

```
FILE 01: NASA: 1962-80.08
                SET  ITEMS  DESCRIPTION
                ---  -----  -----------
? SCC=63
                 1    4190  CC=63
? 3COMPUTER?
                 2   78061  COMPUTER?
? #INFORMATION?
                 3   14970  INFORMATION?
? $1*2
                 4     873  1*2
? $1*3
                 5     255  1*3
? $(2+3)*1
                 6    1075  (2+3)*1
```

b) Procédure de limitation

Il existe une procédure "LIMIT" définie par RECON pour l'ensemble des fichiers de l'ESA. Sa forme est :



Elle permet de trier, parmi les documents sélectionnés, ceux qui correspondent aux caractéristiques définies de la façon indiquée ci-dessus.

L'ordinateur ne consulte pas un fichier inverse mais balaye les champs définis dans chaque document (ce qui n'exige pas le transfert en mémoire de travail de toutes les références contenues dans la classe choisie). Il est plus efficace ou économique de faire porter la limite sur un set plutôt que sur l'ensemble d'une recherche documentaire car dans le premier cas c'est le nombre limité de citations contenues dans le set qui est balayé.

Dans le deuxième cas (LALL/...) l'ordinateur doit balayer les citations associées à chaque mot-clef.

(levée de limitation par la procédure "LALL" ou "L/")

La procédure LIMIT ne permet pas de sélectionner des domaines dans les fichiers NASA (1), COMPENDEX (4), NTIS (6) En effet les n° de citation ("accession number") n'apportent pas d'indications sur la matière traitée dans le document.
Les cas particuliers du fichier INSPEC et PASCAL sont traités ci-dessous.

3) Etude de quelques bases

a- Fichier INSPEC (8)

La commande $L^{ALL}_{[n°set]}/XX/^{A}_{B}_{C}$ ou $L^{ALL}_{[n°set]}\|^{A}_{B}_{C}$ (sans préciser l'année)

↑ année

permet de garder dans la réponse uniquement les citations appartenant à la section A, B, ou C du fichier INSPEC.
Ces trois sections correspondent aux trois abstracts journals d'INSPEC :
A "Physics Abstracts"
B "Electrical and Electronic Abstracts"
C "Computer and Control Abstracts"

exemple : limitation du set n°1 à la section B

```
? L1/ALL/B
                8        0 1/ALL/B
? L1//B
                9        0 1//B
```

limitation des réponses à 1979 et à la section C

```
? LALL/79/C
              LIMIT ALL ALL/79/C
? #CC=1140C
                5        0 CC=1140C
? #CC=C1140C
                6      261 CC=C1140C
```

b- fichier PASCAL (14)

Le numéro de citation attribué à chaque référence lors de son entrée dans la base a changé ces dernières années - l'information sur le domaine du document est toujours incluse dans ce numéro de citation mais d'une façon moins précise.
Cette information n'est d'ailleurs pas utilisable.

La commande LIMIT ne permet d'effectuer qu'une sélection très vague dans le fichier PASCAL entre les sections qui traitent de biologie (section 300 à 390) et toutes les autres.

- Pour n'obtenir que les documents répondant à la question posée et indexés dans une "section biologique" on emploie la commande: $L^{ALL}_{[n°set]}/B$
- Pour éliminer du set tous les documents indexés dans une "section biologique": $L^{ALL}_{[n°set]}/P$

Les 3 tableaux suivants explicitent pour chaque base :

- la forme du code de classement et son utilisation en recherche automatisée.

- le contenu sémantique des bases et éventuellement des sous-bases.

- le contenu sémantique, les techniques d'accès par le langage RECON (ASE) et QUESTEL (TELESYSTÈMES) aux parties du fichier PASCAL intéressant l'informatique au sens large.

- le contenu sémantique, l'accès aux grandes classes de la section C "Computer and Control Abstracts" du fichier INSPEC.

BASE		CLASSIFICATION PAR MATIERE	SELECTION PAR LE CODE MATIERE	
			QUESTION	REMARQUES
NASA	1	Nasa Subject Categories	#CC = XX	xsélection d'une classe ex "31" possible mais sujet trop étroit Il existe 74 classes xSélection d'un ensemble de classes "3?" provoque un débordement n'est pas intéressant (les grands sujets ne sont pas limités par le chiffre des dizaines et peuvent chevaucher deux dizaines)ex : Mathematical and Computer Science s'étend des classes 59 à 67.
OMPENDEX	4	Card-a-lert subjects	# CA = XXX	x sélection d'une classe ex 92? possible Ces classes représentent une grosse quantité de documents (de 30 000 à 100 000) - Il en existe 38. x sélection d'un ensemble de classes 9? provoque un débordement
'TIS	6	Classement NTIS (il existe une correspondance avec le cosati subject categories)	#CF = [X]XY	x sélection d'un chapitre ex "62?" possible x sélection d'une rubrique de chapitre "62B" possible

BASE		CLASSIFICATION DE MATIERE	SELECTION PAR LE CODE MATIERE QUESTION	REMARQUES
INSPEC	8	INSPEC Sectional classification	CC = YXXXX[Y]	x sélection possible à chaque niveau de la hiérarchie par usage de la troncature "?" ex: C 1 4 0 C C 1 4 0 ? C 1 ? C ? Dans "Computer and Control abstracts" il existe 7 grandes classes "CX?" Dans "Physics abstracts" : 8 grandes classes AX?
PASCAL	14	Plan de Classement PASCAL	CC = XXX[.Y.XX.Y XX.Y.X.Y] section chapitre sous-catégories	x la sélection d'une section entière n'est pas toujours possible par exemple la sélection de: 110? 130? } provoque un débordement 145? (On peut limiter par section et par année : 191 ∧ (80) seulement section 191 en 1980 191 ∧ (7?) depuis 1970 191 ∧ [dans les années récentes) x sélection par chapitre et sous-chapitre ex: 110.A 110.C.05.B? etc...

Note : X : caractère numérique

Y : caractère alphanumérique

[ ] : caractère facultatif

INDEXATION DES BASES OU SOUS-BASES SÉLECTIONNÉES SUR L'ESA

BASE	ACCÈS A LA SOUS-BASE	SOUS-BASE	MOTS CLÉS
SA 1			aéronautique astronautique mathématique informatique sciences de l'espace
APENDEX 4			génie
ECOMPS 5] ngue de nnée			composants électroniques (caractéristiques)
IS 6			rapport de recherche
SPEC 8	LALL//A L[n°set]//A	A. Physics abstracts	physique physique des particules électricité magnétisme optique acoustique chaleur mécanique
	LALL//B L[n°set]/B	B. Electrical and Electronic abstracts	électricité électronique télécommunications
	LALL//C L[n°set]//C	C. Computer and control abstracts	informatique (théorie) informatique (matériel) informatique (logiciel) informatique (applications) théorie des systèmes théorie de la commande génie automatique automatique (applications) analyse
SCAL 14	LALL//B L[n°set]//B	sections biologiques (n°300 à 330)	
	LALL// L[n°set]//B	PASCAL, sauf les sections biologiques	information documentation informatique automatique analyse numérique mathématique de l'information physique électricité électro-technique électronique télécommunications

JDEXATION DES SECTIONS OU DE LEURS SOUS-ENSEMBLES SELECTIONNES

UR PASCAL-ESA ET PASCAL-TELESYSTEMES.

ACCES A LA CLASSE OU SOUS-CLASSE	CLASSE OU SOUS-CLASSE	MOTS-CLEFS
101 +$_{\Delta}$ /CH #CC=101?	101	documentation édition information informatique audio-visuel reprographie
débordement "	110	informatique automatique théorie des systèmes analyse numérique statistiques probabilités combinatoire logique mathématique mathématiques de l'informatique économie gestion
débordement "	130	physique optique acoustique chaleur mécanique
140? $_{\Delta}$ /CH #CC=140?	140	électricité magnétisme électrotechnique
débordement "	145	électronique électronique nucléaire optique électronique métrologie matériaux semi-conducteurs propriétés électriques des matériaux électromagnétisme télécommunications
224C $_{\Delta}$ /CH #CC=224.C?	224.C	Géologie Géologie mathématique informatique
#CC=224.C.06.?	224.C.06	Géologie Géologie mathématique informatique

SUITE DE

NDEXATION DES SECTIONS OU DE LEUR SOUS-ENSEMBLES SELECTIONNES

R PASCAL-ESA ET PASCAL-TELESYSTEMES

ACCES A LA SECTION OU SOUS-SECTION	SECTION OU SOUS-SECTION	MOTS-CLEFS
226A △ ICH #CC = 226.A ?	226.A	hydrologie informatique
#CC = 226.A.03	226.A.03	hydrologie informatique
310A △ ICH #CC=310.A ?	310.A	informatique médecine biologie gestion hospitalière
#CC=310.A.05.?	310.A.05	informatique médecine biologie gestion hospitalière
390B △ ICH #CC = 390.B ?	390.B	psychologie psychopathologie informatique
#CC = 390.B.02.D	390.B.02.D	psychologie psychopathologie informatique

notes : TEL. procédure sur Télésystèmes

ESA " l'ESA

△ espace

DEXATION DE LA SECTION C "COMPUTER AND CONTROL ABSTRACTS"

U FICHIER INSPEC (ESA)

E DE LA CLASSE	MOTS-CLEFS
?	
?	théorie des systèmes mathématiques théorie de la commande
?	génie automatique automatique (application)
?	analyse numérique informatique (théorie)
?	informatique (matériel)
?	informatique (logiciel)
?	informatique (applications)

5) Conclusion sur le serveur ESA

Le logiciel RECON de l'ESA fournit deux méthodes pour sélectionner des documents dans un domaine d'intérêt choisi par l'utilisateur :

La méthode de recherche par le fichier inverse du classement par matière. Cette méthode ne se distingue pas de la méthode de recherche par mots-clefs si ce n'est qu'elle permet de définir mieux des champs sémantiques très larges. Cette possibilité est offerte sur la majorité des fichiers de l'ESA et sur tous ceux que nous avons retenus.

Une deuxième fonction "LIMIT" est définie sur les fichiers INSPEC et PASCAL. Elle procède par balayage d'un set de documents.

S'il est possible, sur l'ASE, de limiter à un domaine sémantique les documents sélectionnés, on ne peut obtenir, pour la recherche, l'accès à un sous-fichier particulier.

E. Conclusion sur les deux serveurs TELESYSTEMES ET ESA

Les deux systèmes de documentation TELESYSTEMES et ESA offrent aux utilisateurs des bases de données dont certaines

disposent d'une information polyvalente et d'un nombre très élevé de références.
La mise en place de moyens permettant d'interroger un sous-fichier d'une base de données n'a pas été faite ni sur l'ESA ni sur TELESYSTEMES mais elle se révélera probablement utile dans les prochaines années car le nombre de documents par base va en croissant rapidement.

Le classement par matière interrogeable en conversationnel sur les deux serveurs n'offre par le service précédemment décrit mais il n'est pas dénué d'intérêt. C'est en fait un macro-vocabulaire organisé.
TELESYSTEMES et l'ESA permettent de l'utiliser grâce à un fichier inverse des codes de classement. l'ESA offre, sur les fichiers PASCAL et INSPEC une deuxième procédure pour limiter à un groupe sémantique très vaste, un set de références déja sélectionnées, ceci par balayage de ce set.

Il nous a paru intéressant de ce point de vue d'indexer un peu plus en détail, les bases PASCAL et INSPEC.
Ces bases volumineuses comportent des chapitres particulièrement importants sur l'informatique au sens large.

IV Conclusion Générale

Le système de base de données répartie permet l'accès simultanément ou successivement à plusieurs fichiers de la base répartie grâce à un seul langage global. L'utilisateur, ainsi dégagé des soucis de mémorisation des langages d'interrogation, consultera probablement un plus grand nombre de bases de données qu'il ne le faisait en dehors de la base répartie et sa recherche documentaire s'en trouvera améliorée.

La mise au point d'un tel système pose entre autres problèmes ceux de la correspondance entre le langage documentaire utilisé au niveau global et les langages utilisés au niveau de chacune des bases locales.

Des études linguistiques ont déjà été menées qui étudiaient les aspects particuliers de la compatibilité entre les langages documentaires. Certains affinements du système MESSIDOR, en particulier le regroupement sémantique et l'homogénéisation des langues dans le lexique au niveau global ne peuvent être actuellement obtenus. L'avenir permettra probablement de les réaliser.

La description du contenu sémantique des bases à l'aide de groupes de mots-clefs est un outil documentaire simple à réaliser. Cette description n'est probablement pas définitive mais on peut espérer qu'elle sera, après sa mise en conversationnel, une aide pour l'utilisateur occasionnel de la base répartie.

Celui-ci pourra moyennant une assez bonne connaissance du sujet (donc des relations de synonymie entre les mots-clefs) et moyennant la consultation du lexique global par tranche alphabétique (commande LIST de MESSIDOR) effectuer une recherche documentaire de bonne qualité.