*Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques*

*Université Claude Bernard Lyon I*

***DESS INFORMATIQUE DOCUMENTAIRE***

**- RAPPORT DE STAGE -**

# *Conduite et évaluation d'une étude préalable relative à un projet d'indexation automatique*

**présenté par Karim KAHLAL**

**sous la direction de Monsieur Jean-Paul HOULIER**

*Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques*

*Université Claude Bernard Lyon I*

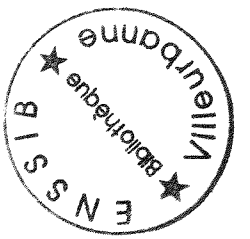
***DESS INFORMATIQUE DOCUMENTAIRE***

**- RAPPORT DE STAGE -**

# *Conduite et évaluation d'une étude préalable relative à un projet d'indexation automatique*

**présenté par Karim KAHLAL**

**sous la direction de Monsieur Jean-Paul HOULIER**



***TITRE*** **:** CONDUITE ET EVALUATION D'UNE ETUDE PREALABLE RELATIVE A UN PROJET D'INDEXATION AUTOMATIQUE

Présenté par KARIM KAHLAL

Stage effectué du 02/03/1992 au 30/06/1992

A

l'ECOLE NATIONALE DE LA SANTE PUBLIQUE
Avenue du Professeur Léon Bernard
35000 RENNES
☎ 99 28 29 30
Monsieur Jean-Paul HOULIER

***RESUME*** **:** L'Ecole Nationale de la Santé Publique a commandé la réalisation d'un système d'indexation automatique ou assistée par ordinateur.
Le système d'indexation (ALETH) proposé par GSI-ERLI, qui comprend entre autres un module syntaxico-sémantique viendra s'interfacer avec l'outil GESTH (gestionnaire de thésaurus) de Cybernetix. L'aspect fonctionnel et les adaptations au contexte informatique y sont analysés.

***DESCRIPTEURS*** **:** Indexation automatique / interrogation base donnée / analyse linguistique / analyse morphologique / analyse syntaxique / analyse sémantique / descripteur / règle indexation / thésaurus / base connaissance / langage naturel / interface utilisateur.

***ABSTRACT*** **:** The National School of Public Health has decided to introduce an automatic or semi-automatic computer-assisted indexing system.

The system proposed by GSI-ERLI : ALETH includes a syntactic and semantic component to analyse user queries.
ALETH wil interface with software GESTH wich manages a thesaurus.
The proposed system is analysed and the adaptations that will be necessary to the existing system are discussed.

***KEYWORDS*** : Automatic indexing / Data base query / linguistic analysis / morphological analysis / syntactic analysis / semantic analysis / Descriptor / indexing rule / Thesaurus / Knowledge base / natural language / user interface.

# Sommaire


***INTRODUCTION*** 7

***CADRE ET OBJECTIF DU STAGE*** 8

***PREMIERE PARTIE : LE CONTEXTE*** 9

I - Présentation de l'ENSP 9
II - Le service Commun de la Documentation 12

***DEUXIEME PARTIE : INDEXATION AUTOMATIQUE*** 23

I - Identification des concepts 24
II - Le cas ENSP 30
III - Le poste d'indexation ENSP 32
IV - Spécifications fonctionnelles 41

***CONCLUSION*** 50

***BIBLIOGRAPHIE*** 53

Table des matières 55

# Introduction

❑

*Le monde de la documentation porte un intérêt particulier aux nouvelles techniques du traitement de l'information. Le traitement automatique de la langue présente cette nouvelle option prise par les spécialistes de l'information. Les enjeux sont importants et répondent parfaitement aux exigences que l'information requiert aujourd'hui.*

*Un nouveau concept emmerge, celui de "l'industrie de la langue", l'union de l'informatique et de la linguistique a engendré un nouveau marché technologique et industriel. Il est vrai que le texte en tant qu'objet de langue pose des problèmes linguistiques, donc tout système doit garder une place privilégiée pour la résolution de ces problèmes. Par ailleurs, le développement fulgurant de la télématisation grand public (surtout en France avec quelques 3 500 services) incite le recours au T.A.L.(*) (langage naturel).*

*Les linguisticiels sont le produit de l'option prise dans la recherche - outils à applications multiples dans le domaine du T.A.L.(*) - ils représentent un enjeu culturel très important (veille technologique) et permettent :*

⇨ *rapidité du traitement de l'information (l'indexation automatique accélère le signalement de l'information)*

⇨ *préservation de la valeur stratégique de l'information*

⇨ *facilités d'accès (langage naturel)*

⇨ *diminution des coûts*

---

*(*) : Traitement Automatique de la Langue*

# Cadre et objectif du stage :

*Mon stage s'inscrit dans la perspective du suivi de l'élaboration d'une étude préalable à la réalisation d'un système d'indexation automatique ou assistée pour l'ENSP.*

*Mon travail consistait donc à élaborer un ensemble de questions et de remarques concernant l'installation du système. Mes interlocuteurs étaient l'ENSP et ERLI :*

*- l'ENSP (le service documentaire) pour l'analyse de l'existant, c'est-à-dire la définition des besoins à travers les "lacunes" du système en place (l'indexation manuelle), la prise en compte du contexte informatique (l'indexation automatique s'inscrit aussi dans cette lancée).*

*Il faut souligner l'apport considérable des remarques des documentalistes qui m'ont éclairé sur l'ensemble des difficultés rencontrées au cours de leur travail.*

*Quant à ERLI, ce sont différentes réunions de travail qui ont permis d'affiner nos propositions en termes de besoins auxquels le système devait répondre.*

*Le responsable du service documentation insistait quant à lui sur l'importance de la tenue des délais pour éviter la même situation que celle connue lors de la réalisation de Biblix/Gesth (interface d'interrogation). Par ailleurs, les études préalables (trois au total) présentées par Cybernetix (avant son rachat par ERLI) et ERLI m'étaient d'une grande utilité pour mieux comprendre la fontionnalité des systèmes dits "Linguisticiels", au même titre que l'ensemble des articles et thèses que j'ai eu à consulter.*

# Première partie : Le contexte

## I - Présentation de l'ENSP

L'Ecole Nationale de la Santé Publique de RENNES a été créée par ordonnance du gouvernement provisoire de la république française le 19 octobre 1945.

La loi du 28 Juillet 1960 la réorganise et l'érige en établissement public national, doté de la personnalité morale et financière. L'école est installée à Rennes depuis le 13 Avril 1962.

### *Missions de l'ENSP*

Les missions de l'ENSP s'articulent autour de quatre grands axes :

- *La formation:*

L'ENSP assure la formation d'une grande partie des gestionnaires du secteur sanitaire et social français (personnel supérieur des services departementaux et nationaux du ministère chargé de la santé, personnel des établissements sanitaires et sociaux publics, ingénieurs et techniciens en génie sanitaire).

- *La recherche:*

L'école est liée par plusieurs conventions de recherche à de grandes écoles et universités ainsi qu'à d'autres institutions spécialisées. En outre L'ENSP est "centre collaborateur" de l'organisation mondiale de la santé en planification sanitaire et formation à la gestion.

Les principaux domaines de recherches concernent notamment les liens entre l'environnement et la santé, la socio-économie de la santé, la gestion hospitalière, l'épidémiologie.

- *Prestations de services :*

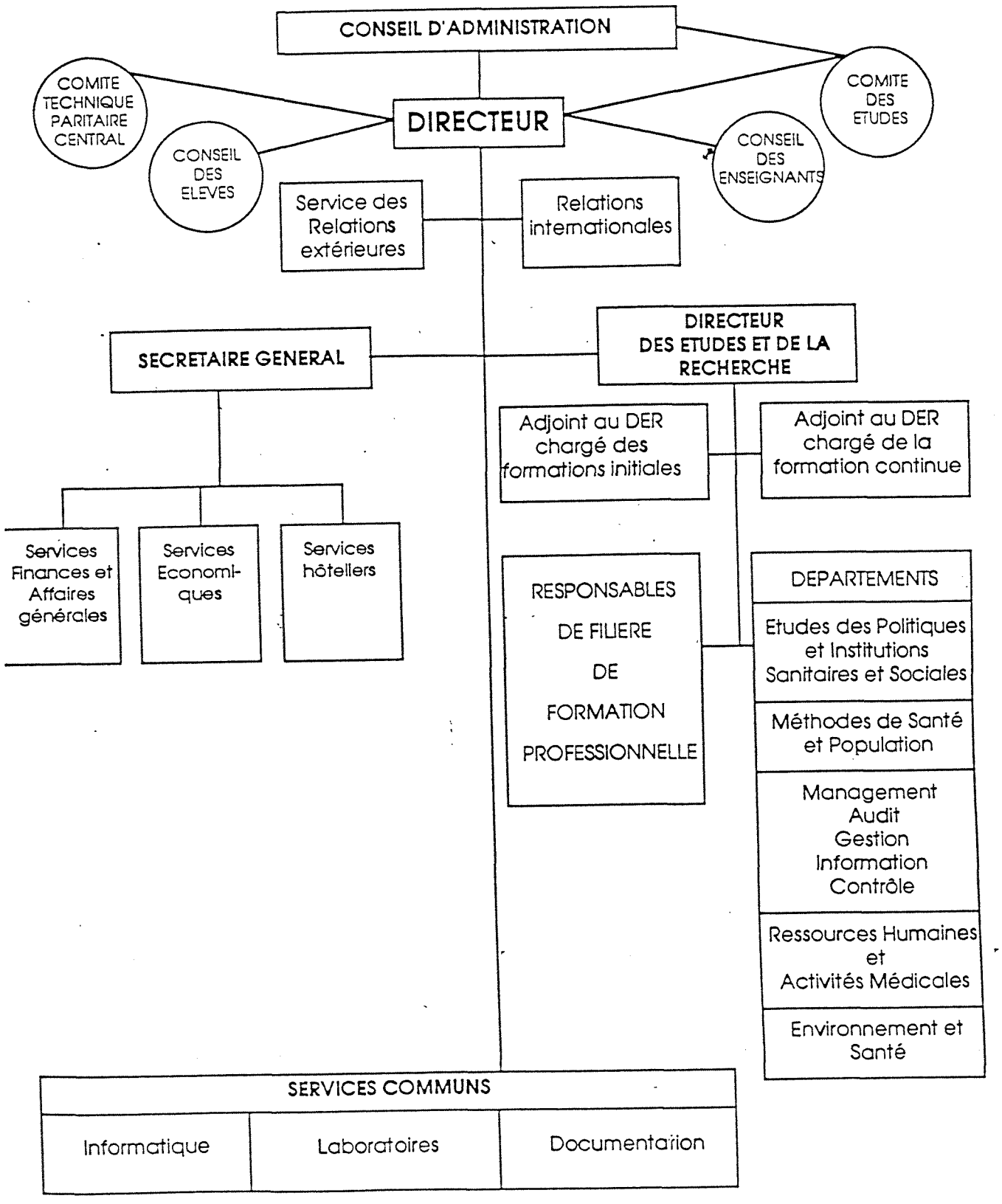
L'ENSP valorise ses moyens en étant prestataire de services, ses laboratoires sont fréquemment utilisés, ainsi que son matériel informatique, pour diverses opérations liées au domaine de la santé publique, mais aussi, en organisant des colloques ou des congrès.

- *L'activité internationale :*

L'école participe activement pour le développement d'une politique de coopération internationale dans de nombreux domaines :

* Accueil d'étudiants étrangers dans ses formations initiales.
* Participation à la formation des professionnels de la santé dans leur pays d'origine.
* Production de matériel pédagogique.
* Contribution à l'enseignement de l'épidémiologie.
* Participation aux associations internationales.

*Organisation*
*de l'Ecole Nationale de la Santé Publique*



# II - Le Service Commun de la Documentation

Le Service Commun de la Documentation est directement rattaché à la Direction de l'Ecole. Il regroupe trois secteur d'activité et mène une politique d'automatisation volontariste.

## *1/ Organisation*

Trois secteurs d'activités sont regroupés dans ce service :

- **L'enseignement par correspondance.**
- **Les éditions ENSP.**
- **La bibliothèque.**

Chaque secteur participe à la production, au traitement et à la diffusion de l'information.

### a) - L'enseignement par correspondance:

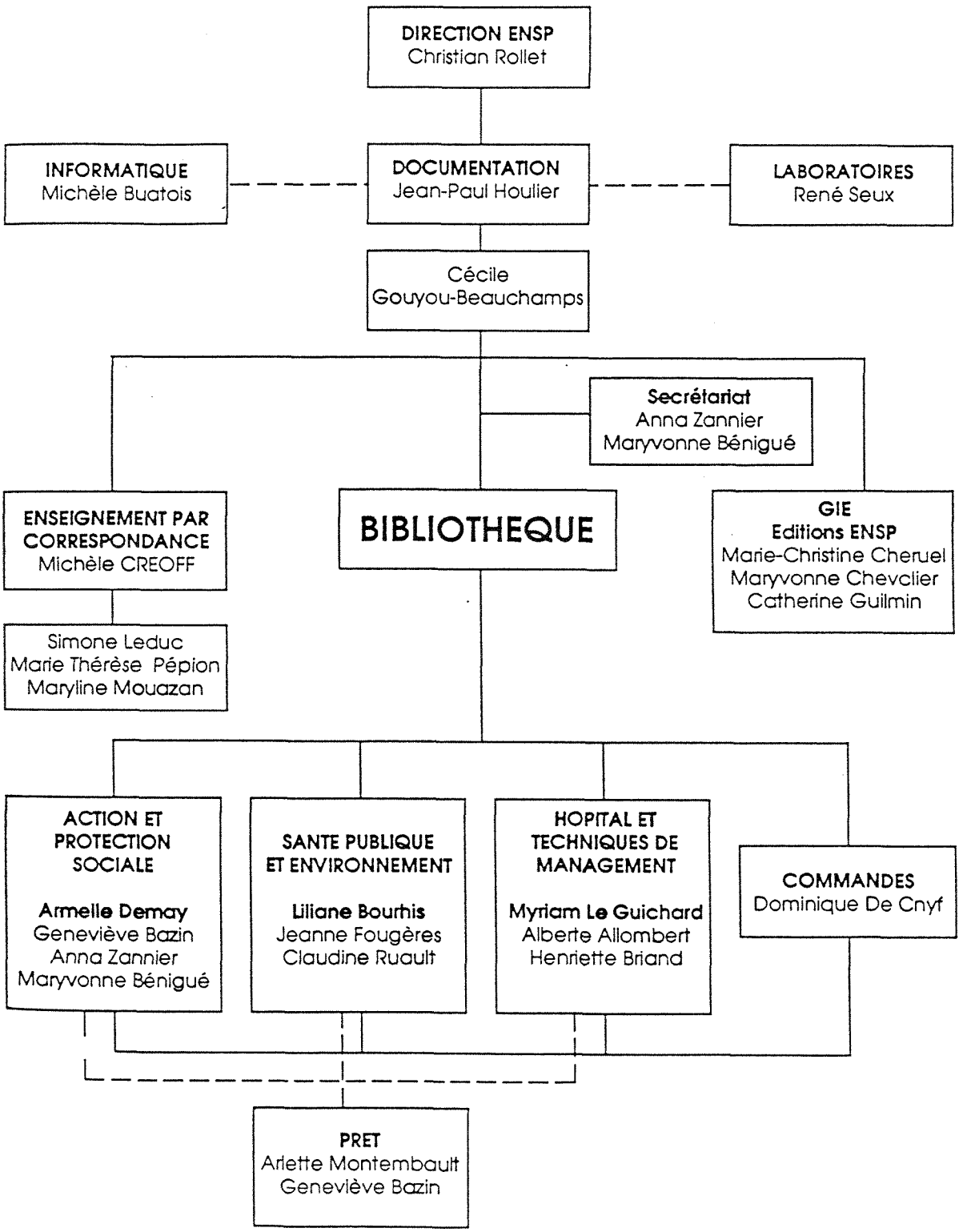
Assure des prestations pédagogiques (cours, dossiers de presse..) pour les candidats aux divers concours organisés par l'Ecole. Par ailleurs des tests de connaissances sont mis à la disposition des candidats et accessibles par MINITEL, avec une mise à jour mensuelle.

Malgré des incertitudes sur la formation (élèves directeurs de 4ème classe) et autres reports de concours (IASS) en 1991, 22 dossiers de presse et 16 cours ont été mis à jour.

### b) - Les Editions ENSP :

Depuis 1989, l'école a constitué avec deux sociétés rennaises, BUROSCOPE et le Groupe Média Calligraphy, un groupement d'intérêt économique (GIE) qui doit permettre à ce secteur de développer son activité dans des conditions plus satisfaisantes.

*Organigramme du service commun de la documentation*



Ce secteur participe à la production et à la diffusion des travaux conduits par les élèves et les professeurs et est de plus en plus sollicité par l'extérieur. Ses principales tâches sont l'édition et la commercialisation de ces travaux.
C'est un secteur productif et rentable (3 millions de CA en 1991). L'effort de distribution vers les librairies lui permet d'envisager une croissance encore plus soutenue.

Enfin, l'ENSP et la société française de santé publique ont créé, en septembre 1988 une revue commune "Santé Publique" qui veut rendre compte des multiples recherches et actions entreprises par les professionnels de la santé publique.

c) - **La bibliothèque:**

Avec plus de 35 000 ouvrages, 6 000 travaux d'élèves, et plus de 500 titres de périodiques,la bibliothèque est devenue le centre de documentation de référence en santé publique. C'est ainsi que l'ENSP a été chargée par 3 ministères de mettre en place une banque de données unique en santé publique.

L'organisation du travail est articulée autour de trois documentalistes spécialisés dans l'un des domaines suivants :

- **la santé publique et l'environnement.**
- **l'action et la protection sociale**
- **l'hopital et les techniques de management.**

Les équipes constituées autour de chaque documentaliste sont en relation permanente avec les enseignants et les usagers, afin de rendre l'information plus pertinente et plus utile.

Le bilan pour l'année 1991, a fait apparaitre une forte augmentation (25%) du nombre de consultations et une évolution positive du nombre d'emprunt (+ de 19%) signe d'une fréquentation de plus en plus importante.

## *2/ Contexte d'informatisation*

a) **Etat des lieux**

L'informatisation de la bibliothèque a débuté en 1987, au niveau du fonds documentaire, par la création de la base de données bibliographiques.

A cette époque l'école participait à deux réseaux documentaires :

Le réseau RESHUS (Réseau en sciences humaines et de la santé) :

| | |
|---|---|
| Producteur : | Institut de l'information scientifique et technique (INIST-CNRS) |
| Domaines : | Santé publique ,sciences humaines et sociales médecine, consommation ,bio-éthique. |
| Nature : | Références bibliographiques. |
| Debut : | 1970 |
| Volume : | 13 700 références + 1 300/an |
| Mise à jour : | Trimestrielle |
| Langue : | Français |
| Accés : | L'européenne de données (RESH) Questel (Francis). |

Le réseau RAMIS (Réseau pour l'amélioration de l'information en santé publique):

| | |
|---|---|
| Producteur : | Association Ramis |
| Domaines : | santé publique,éducation pour la santé, promotion de la santé |
| Nature : | Références bibliographiques |
| Volume : | 17 000 références |
| Mise à jour : | Trimestrielle |
| Langue : | Français |
| Accès : | 3617 code ENSP (depuis le 11 O1 92) |

Les règles d'écritures et le vocabulaire d'indexation devaient donc être définies à partir de l'une de ces deux bases, le choix s'est porté sur RESHUS, avec qui l'ENSP coopérait beaucoup plus.
Le développement de la base a été réalisé en interne à partir du logiciel TEXTO-LOGOTEL, implanté sur un micro-ordinateur Microméga (sous unix) de la société ALCATEL (doté d'un MO de mémoire centrale,d'un disque dur de 45 MO et d'un lecteur de bandes magnetiques (streamer) de 60 MO),qui gérait à la fois la saisie des références et la consultation de la base.
La base de données devait pouvoir être interrogée par tout utilisateur de la bibliothèque, non initié aux techniques de recherche documentaire. Il s'agissait donc avant tout de réaliser un outil de recherche simple et fiable. L'utilisation d'un terminal MINITEL a été privilégié, le MINITEL étant d'une utilisation simple et largement diffusé, un VIDEOCOM et six cartes MODEMS permettaient la connexion du MINITEL à la base de données, l'accès interne se faisait par le 17 et l'accès externe par un numéro a huit chiffres par le biais du réseau téléphonique commuté.
Afin d'améliorer les performances du service, l'école s'est dotée d'un micro-ordinateur (sous unix) de UNISYS (micro-processeur 386 de 16 MO, disque dur de 170 MO),qui est venu décharger Microméga de la fonction de consultation de la base, le transfert des données s'effectue par l'intermédiaire des bandes magnétiques (streamer).

- **Le 3617 code ENSP :**

Très rapidement, la nécessité d'étendre les possibilités d'accés s'est fait ressentir surtout pour un public dispersé géographiquement, en particulier toutes les institutions sanitaires et sociales publiques ou privées, une carte X25 devait donc permettre la connexion au réseau TRANSPAC qui est le mieux adapté à la transmission de données, le 3617 code ENSP devait alors permettre à plusieurs utilisateurs d'interroger simultanément la base de données.

- **Le prêt :**

La gestion du prêt a été réalisée en relation avec le fichier des références bibliographiques, à partir du logiciel TEXTO-LOGOTEL par une société de services (SINORG), il est porté sur UNISYS.

La commande des ouvrages/suivi des abonnements des périodiques :
La gestion de cette fonction est gérée à partir du logiciel 4ème dimension, porté sur Macintosh.

**b) Développements actuels**

- **les limites et les besoins**

Après plusieurs années d'utilisation par le public interne et externe à l'école, il s'est avéré que le système d'informatisation, hormis la simplicité de procédure qu'il proposait, comportait en effet certaines limites :

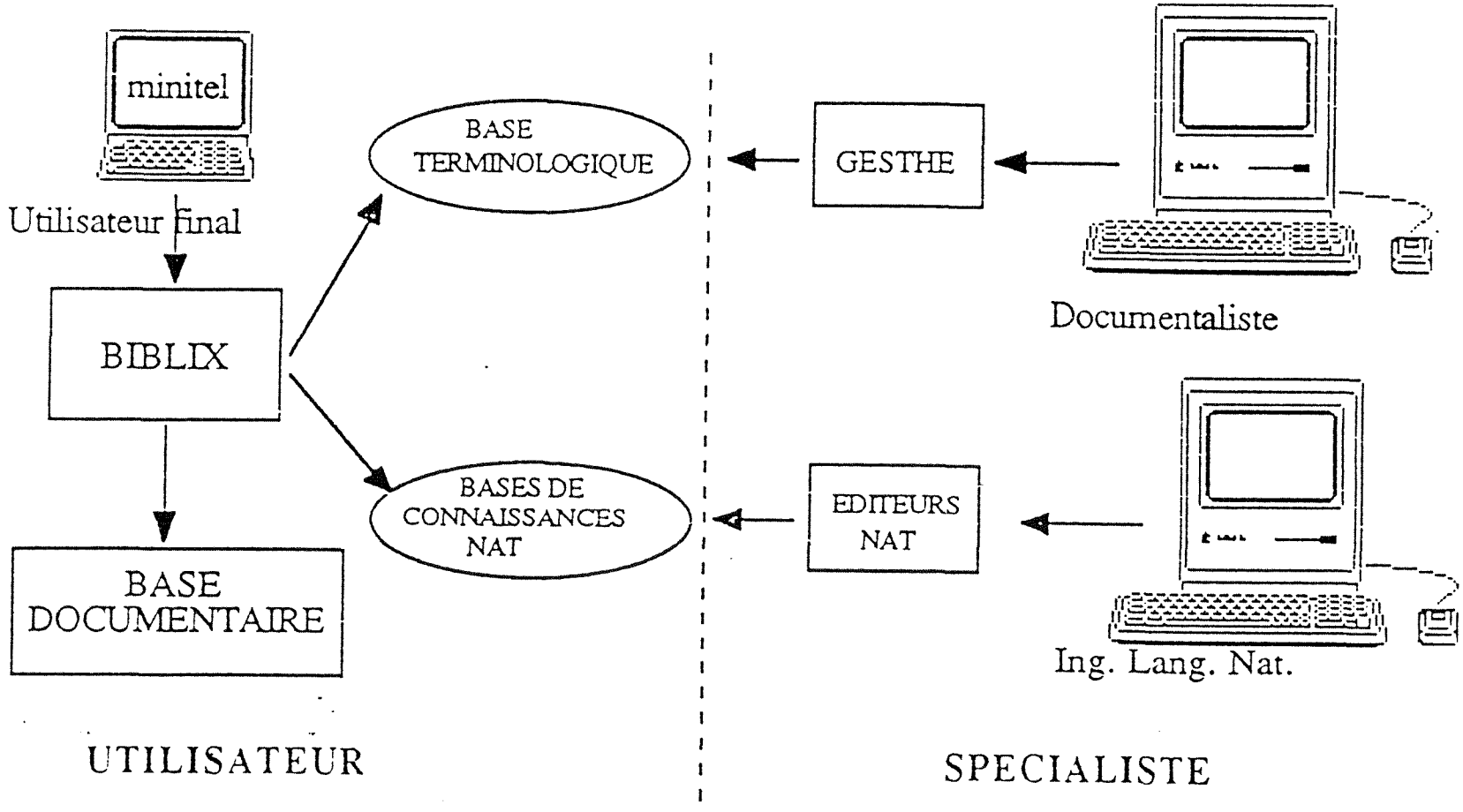
- Le langage d'accès aux informations proposé par TEXTO est très technique, la rigidité de sa syntaxe obligeait les utilisateurs à connaitre les mots de l'index pour interroger.

- L'index provient d'un thésaurus que seul le public interne pouvait consulter, pour le public externe l'accès en ligne s'avérait inefficace, car le thésaurus est assez lourd à manier et son parcours assez long.

- Il était impossible de réaliser une recherche multicritère, en une seule étape. Avec un seul critère, on obtenait soit trop de documents non pertinents ("bruit"), soit trop peu par rapport au fonds réel, il fallait donc modifier les critères pour obtenir les références désirées et donc ajouter une étape aux étapes déjà réalisées.

Le problème principal qui se posait à l'utilisateur donc, était de trouver les mots connus du système.

- Enfin, un autre inconvénient de taille : l'utilisateur n'avait pas la possibilité de savoir où il en était dans sa recherche.


minitel
Utilisateur final
BIBLIX
BASE DOCUMENTAIRE
BASE TERMINOLOGIQUE
BASES DE CONNAISSANCES NAT
GESTHE
EDITEURS NAT
Documentaliste
Ing. Lang. Nat.
UTILISATEUR
SPECIALISTE

Les besoins détectés étaient de deux types :

- L'interrogation de la base par les utilisateurs :

L'école souhaitait voir réaliser un système qui met en oeuvre de réelles facilités d'accès à l'information. La télématisation grand public de sa base de données engendrait des besoins d'interrogation en un langage informel et compréhensible par le système qui permettrait à l'utilisateur par le biais d'un dialogue évolutif, d'accéder à une qualité de réponse supérieure.

- La construction et la maintenance du thésaurus :

le système devait s'appuyer sur les bases de connaissances constuites à partir des index et du thésaurus de la base documentaire de l'ENSP.
C'est dans ce cadre que l'ENSP a décidé de commander auprès de la société CYBERNETIX (spécialisée à l'époque en ingenierie des systèmes automatiques et robotiques), la réalisation d'une interface en langage naturel conviviale et souple.
Par ailleurs la conception et la maintenance d'un thésaurus était le complément indispensable à la réalisation de l'interface.

**c) La solution Biblix-Gesth (*)**

La solution informatique proposée par Cybernetix est un logiciel composé de deux outils (Biblix/ Gesth) ayant deux fonctions distinctes :

- L'outil GESTH : concerne la fonction "administration de la base" qui permet au service de la documentation de constituer et de maintenir l'ensemble des termes qui seront utilisés par le logiciel dans sa fonction "interrogation", et par les documentalistes pour l'indexation des ouvrages, et qui pourra déboucher dans une deuxième étape sur un système d'aide à l'indexation.

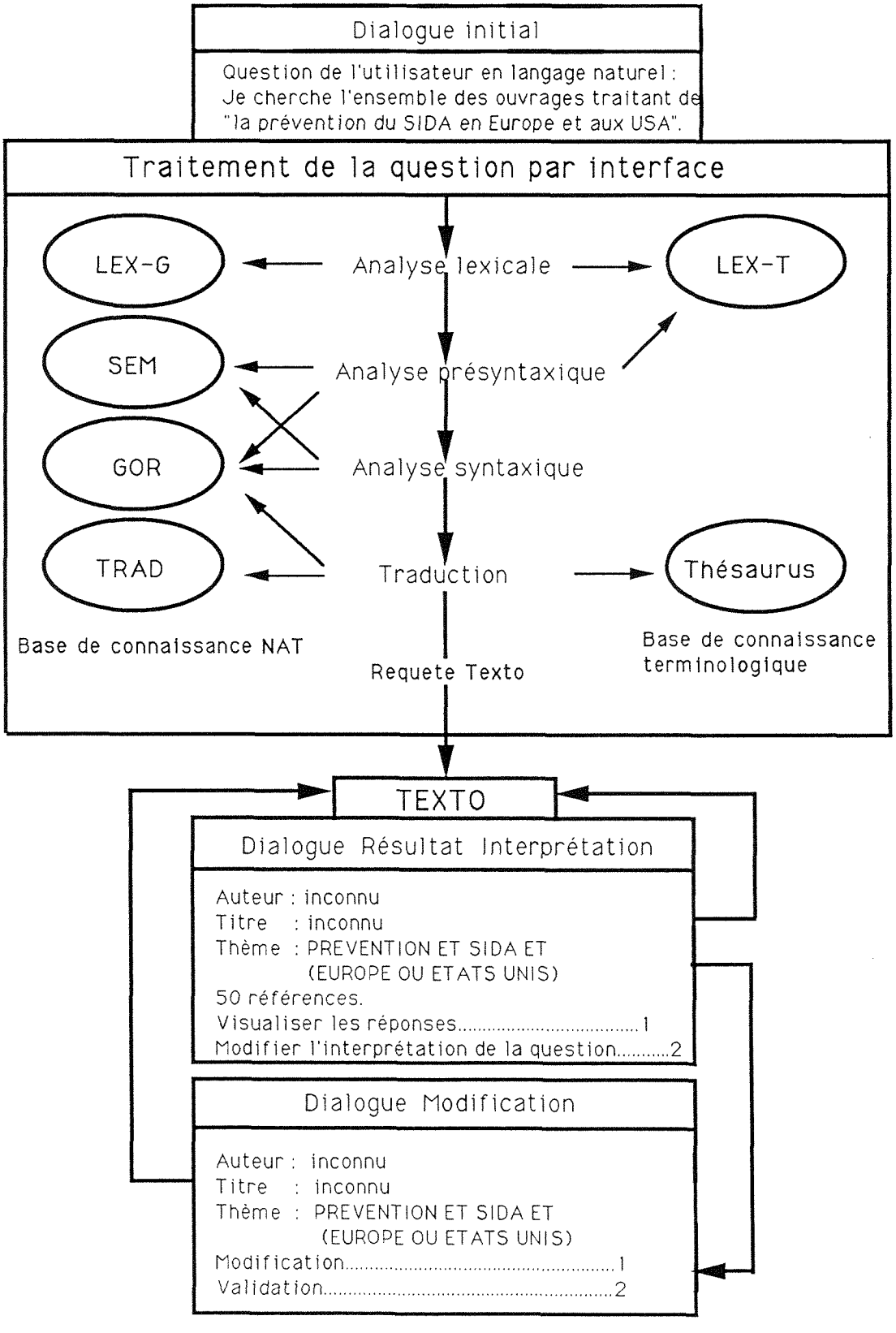
L'Outil BIBLIX - La fonction "interrogation" qui permet à l'utilisateur final d'interroger une première fois la base de données en langage libre et à partir du nombre de réponses obtenues peut lui donner la possibilité de consulter l'ensemble des références séléctionnées, d'affiner ou d'élargir sa recherche en un dialogue "évolutif" ; le système propose à l'utilisateur d'élargir ou de restreindre sa recherche à l'aide d'informations claires sur l'organisation des termes dans le thésaurus et en lui permettant d'ajouter ou d'enlever des critères de recherche.

Ces deux fonctions ont été réalisées par un ensemble de modules utilisant des dictionnaires à bases de connaisances communes. La réalisation du système est basée sur un ensemble d'outils de traitement du langage naturel appelé NAT. (propriété de Cybernetix/Saia).

---

(*) Biblix/GESTH a fait l'objet d'un rapport de stage (fonctionnement, test) présenté par Maurice N'DONG dans le cadre du DESSID (cité en référence).

# 36 17 ENSP
# Interface en Langage Naturel



Ce système proposé par Cybernetix devait permettre :

à l'ENSP de conserver la base bibliographique construite sur un SGBD documentaire conventionnel ;

à l'utilisateur final d'interroger en langage naturel, au départ sur cinq champs (auteur, titre , thème, date, type de document), par minitel ou autres terminaux, une base de données documentaire, la recherche du document étant facilitée par un système connaissant le domaine documentaire grâce à une base terminologique

aux documentalistes de constituer et maintenir la base terminologique grâce à une éditeur convivial et intelligent.

**Evaluation au stade actuel**

Après plusieurs mois de tests, on s'aperçoit que "Biblix" n'est toujours pas opérationnel et ce malgré les estimations "optimistes" de Cybernetix qui porte le taux de réussite du système à 85 %, ce qui semble paradoxal, car les tests effectués font apparaitre un ensemble de lacunes qui remettent en cause la fiabilité du système.

Cette sur-estimation est sans doute liée à la méthodologie des test effectués par cybernetix, en effet, l'interrogation par "thème" ne pose aucun problème (c'est ce qui semble être pris en considération), or il fallait procéder par un échantillon représentatif des diffférents types d'interrogation envisagées, ce qui engendre les remarques suivantes :

- l'interrogation sur le titre ne fonctionne jamais
- l'opérateur "sauf" utilisé entre deux critères fonctionne à "l'envers"
- l'interrogation par auteur est quasiment inopérante, si l'auteur est une personne morale
- la visualisation des références est largement incomplète
- le temps de réponse est relativement long

Cybernetix (GSI ERLI) propose une solution provisoire pour la mise en service interne du système. Cette solution consiste à établir une liste d'instructions destinées à l'utilisateur, éventuellement avec un message d'une part, d'autre par un établissement d'un fichier où seront stockées les interrogations, ce fichier permettra de cerner un ensemble représentatif des différentes anomalies et apporter les solutions adéquates.

Evaluation au stade actuel
(Biblix sous Test interne)

Le logiciel **Biblix** qui est maintenant sous test permet d'interroger la base documentaire de l'ENSP.
Merci de bien vouloir formuler vos questions dans un langage naturel clair et précis. Les traces de vos questions vont nous aider à améliorer ce logiciel.

## Instructions d'utilisation

- tournures courtes
  ~~Je cherche tous les documents de type article qui portent sur~~ ...
  -> articles sur ...

- minuscules en début de question
  ~~Etat de santé~~
  -> etat de sante

- pas d'accents
  ~~état de santé~~
  -> etat de sante

- sigles en minuscules et sans points
  ~~I.N.S.E.R.M~~
  -> inserm

Excuses pour les éventuels problèmes qui pourraient surgir, et bonne interrogation!

# Deuxième partie :
# Indexation automatique

Les spécialistes de l'information, les linguistes, les informaticiens consacrent depuis plus de 30 ans leurs réflexions à l'automatisation de l'indexation.
Ceci se vérifie si l'on considère le nombre croissant de publications dans ce domaine.
S'il s'agissait au départ de recherches purement théoriques celles-ci débouchent maintenant sur des applications industrielles.
Les logiciels ainsi élaborés reflètent les diférentes options prises par la recherche.
Aujourd'hui le marché de l'indexation automatique recouvre une large gamme de produits aux conceptions et performances différentes.
On peut en distinguer deux catégories distinctes :

- les logiciels documentaires ayant un module d'indexation automatique (appelés aussi logiciels classiques compte-tenu de leur ancienneté)

- les logiciels qui reflètent l'option prise par la recherche actuelle basée sur l'analyse linguistique.

Dans cet exposé l'aspect pratique d'un module d'indexation automatique ou assisté sera privilégié. Il s'agira donc de décrire le fonctionnement et les besoins auxquels est censé répondre le logiciel dans une perspective applicative.
Cependant, il parait nécessaire de rappeler certains concepts fondamentaux de l'indexation documentaire qui permettront aux lecteurs de ce mémoire de mieux appréhender le problème.
La représentation de l'évolution des recherches et la typologie des méthodes mises au point ne feront pas l'objet d'une étude détaillée.
En effet, si certains aspects peuvent avoir un intérêt particulier pour étudier le module d'indexation proposé par la Société GSI ERLI à l'ENSP il s'agit dans cet exposé de ne retenir que les références indispensables à la compréhension de la proposition décrite.

# I - Identification des concepts

## *L'indexation*

L'indexation est l'identification et l'enregistrement des unités d'informations minimales, pertinentes qui sont susceptibles de permettre l'accès à l'information par des éléments tirés du contenu thématique d'un texte. L'objectif fondamental de toute indexation est le signalement optimal du contenu des documents. L'indexation repose sur des langages artificiels qu'on appelle langages d'indexation, concus à l'aide d'un ensemble de règles servant à la représentation du contenu d'un texte.

## *Automatiser l'indexation*

Depuis longtemps les chercheurs se sont posés la question de savoir si l'indexation pouvait être sujette à une automatisation.

J. CHAUMIER et M. DEJEAN définissent l'indexation comme un "art".(1)

Cette fonction demande une compétence dans des domaines variés. Ils écrivaient dans les années 70 "aussi peut-on se demander si ce que l'on a appelé un art peut faire l'objet d'un processus de mécanisation".(1)

A cette époque JCL GARDIN parlait de contraintes d'ordre pratique qui poussaient les chercheurs "à contourner l'obstacle".

De par les moyens qu'elle demande et les délais nécessaires à sa réalisation les signes du "Déficit" de l'indexation manuelle se sont fait sentir.

Les documentalistes n'arrivaient plus à soutenir le rythme des parutions de plus en plus important des documents à analyser.

Beaucoup de chercheurs affichaient un optimisme infondé. Ils continuaient à travailler sur le concept d'indexation automatique, alors même que le développement des recherches en intelligence artificielle ne permettait pas d'envisager à court terme une industrialisation des logiciels d'indexation automatique.

Jean-Claude GARDIN définissait l'indexation automatique comme "un ensemble de règles assurant le passage automatique d'un texte écrit dans une langue naturelle à une représentation de ce texte qui soit censée en exprimer le sens, du point de vue largement intuitif où se placent habituellement les documentalistes".(1)

Il mettait déjà le point sur la complexité engendrée par une telle définition "on ne saurait être surpris de la relative complexité des règles en question".(1)

## *Procédures d'indexation (*)*

Au niveau le plus simple un système d'indexation comporte : (2)

- *des flux d'entrées*

ce sont les objets textuels soumis à l'indexation. A ce stade, il s'agit de définir le niveau de structuration des textes : leur structure générale, leur volume et leur domaine de référence.

- leur structure générale :

  - texte sans structure interne particulière
  - texte faiblement structuré (grandes divisions, sous-titres)
  - texte fortement structuré

- leur volume

  - texte condensé
  - texte intégral bref
  - texte intégral long

- leur domaine de référence

  - domaine identifié circonscrit
  - domaine indentifié large ou condensé
  - domaine très vaste
  - organisation concentrique

- *un processeur*

  - nature du processeur

    * humain
    * automatique
    * semi-automatique

Le choix du processeur peut être déterminé par le degré de complexité du problème posé (si le champ est orienté vers une indexation automatique, l'accent sera mis sur le niveau d'analyse requis). Il faut cependant souligner la plus grande adaptabilité de l'indexeur humain dans la mesure où un documentaliste passe aisément d'un degré de complexité à un autre. Or, une machine, aussi perfectionnée soit-elle ne peut que répondre à des éventualités définies par avance. Ainsi en est-il surtout pour les problèmes nécessitant une analyse sémantique.

Jean-Claude GARDIN écrivait "l'analyse sémantique d'un texte scientifique est une opération éminement intelligente, qui exige une double compétence, sur le plan de la langue d'abord, mais aussi sur le plan de la pensée scientifique elle-même (...) La machine doit être instruite de la même manière dans ces deux ordres de compétences".(1)

---

(*) : Il s'agit au fait de la présentation d'une grille d'analyse pour définir une procédure d'indexation.

- *des flux de sortie*

Ce sont les données obtenues en sortie, là aussi, il s'agit de définir la structure ou la nature de ces données (mot ou expression).

*** nature des données obtenues en sortie du système :**

- les données ont obligatoirement une forme simple (unitenues)
- les données peuvent avoir une structure complexe
- les données doivent avoir une structure complexe.

*** contrôle des données :**

- **indexation contrôlée, le contrôle se fait soit :**

  **> sur une liste d'autorité représentant des concepts définis à priori (Thésaurus, listes de vedettes - matières etc...)**
  **> sur un lexique général (permettant de vérifier uniquement formellement l'appartenance des données obtenues à un ensemble).**

- indexation libre (les données obtenues viennent accroître un ensemble non-clos)

*** nature du rapport que ces données entretiennent avec les textes d'entrée :**

- elles représentent de manière exhaustive le texte d'entrée
- elles représentent uniquement de façon sélective le texte d'entrée

*** nature du rapport que ces données entretiennent entre elles :**

- chaque donnée obtenue a un statut identique
- les données obtenues reçoivent une pondération selon leur importance présumée pour la représentation du texte.

B. MENON parle de diversité de procédures possibles au vu des configurations qu'on peut tirer de la grille d'analyse. Ces différents critères enumérés présentent des niveaux de complexité variables, il s'agira selon le cas d'inversion systématique des mots d'un texte bref et non structuré, où chaque chaine de caractère entre deux blancs devra être extraite, ou bien la production d'un index structuré bilingue pour un ouvrage encyclopédique.

En fonction de ce qu'on attend d'un système d'indexation, cette grille d'analyse peut servir de référence.

Dans un article consacré à l'indexation automatique et à l'intelligence artificielle Bruno Menon pose un ensemble de questions qu'il a appelées "questions de stratégie".(2)

Ces questions résument toute la complexité du problème posé par l'indexation aujourd'hui :

- l'indexation manuelle lorsqu'elle est possible est-elle toujours préférable à l'indexation automatique ? (ou vice versa)

- Dans quels cas est-il souhaitable (envisageable) de se doter d'un système d'indexation automatique ?

- l'indexation automatique doit-elle utiliser peu ou presque toutes les techniques de l'intelligence artificielle ? Peut-elle au contraire s'effectuer selon les méthodes algo-rythmiques supposées plus sûres et plus simples à mettre en oeuvre ?

- une indexation automatique "intelligente" doit-elle être fondée sur les techniques de traitement du langage naturel, et/ou sur les techniques de représentation des connaissances et/ou sur la formalisation du raisonnement ?

Ces quatres questions résument toute la complexité de l'automatisation de l'indexation. Néanmoins l'indexation représente une activité et un enjeu économique tels que la réponse à ces quatres questions peut paraître évidente mais le responsable d'un centre de documentation formule d'autres questions :

- quel est le système qui répond le mieux à ses besoins ?

- A quelle architecture matérielle doit-il faire appel ?

- quels sont les coûts de sa réalisation ?

Le choix final dépend donc d'un ensemble de facteurs variés :

besoins identifiés, objectifs fixés au système d'information, moyens diponibles, etc...

## *Evolution des systèmes et enjeux économiques*

La littérature consacrée au domaine de l'indexation automatique n'a cessé de mettre en évidence les freins qui sont devenus aujourd'hui les vrais tremplins.
La diversité des procédures d'indexation, l'évolution des produits et la diminution des coûts sont les véritables atoûts de l'indexation automatique.

- *Systèmes*

D'un point de vue chronologique, les premiers systèmes d'indexation automatique étaient basés sur des techniques statistiques (fréquence des mots), qui se sont enrichis petit à petit par des connaissances linguistiques (flexions - synonymie). Le défaut des systèmes statistiques est d'ignorer les expressions et ne travailler qu'au niveau des mots isolés. C'est pourquoi la méthode de co-occurence a été conçue. Cette méthode permet d'obtenir une table avec la fréquence non seulement des mots isolés mais aussi des paires de mots qui préfigurent les expressions et plusieurs produits utilisent cette méthode
(entre autres : leximapper, le module Passat ...).

Par ailleurs, les systèmes statistiques évoluent vers le modèle linguistique, avec analyse syntaxique et reconnaissance d'expressions (l'exemple de SPIRIT de systex).

Enfin, l'outil ALETH de GSI ERLI comrpend l'analyse linguistique par rapport à un thésaurus en plus des règles d'indexation. Beaucoup d'autres systèmes outre-atlantique se basent sur des règles (machine Aided Indexing de l'Amercian Petroleum Institute, etc...).

Du point de vue produits, on peut en distinguer deux catégories :

*** les logiciels documentaires ayant un module d'indexation automatique :**

Ces logiciels ont une conception plus ancienne que les linguisticiels (ils ne comprotent pas d'analyse linguistique) mais il ont néanmoins un module d'indexation automatiques. Ces logiciels procèdent en général par retrait de mots vides et inversent la totalité des mots restants (on parle beaucoup plus d'inversion automatique que d'indexation automatique).
Il existe sur le marche un certain nombre de logiciels de ce type (BRS, BASIS, AUTO-X (programme de Chemdata) STAIRS, FLECS, etc...). Seuls quelques uns proposent l'analyse "morphologique" du terme (BASIS, AUTO-X).

*** Les linguisticiels :**

Ce sont les logiciels qui se basent sur l'analyse linguistique, les trois systèmes étudiés (ALEH/ERLI, SPIRIT/SYSTEX, DARWIN/CORA) sont assez différents si l'on considère le type d'analyse et le mode de recherche.

Avec ALETH, l'indexation est assistée et en langage contrôlé. La recherche est lancée après l'indexation de la question et la formulation booléenne automatique de celle-ci. Le tratiement d'ALETH est linguistique (analyseur morphologique, syntaxique et sémantique).

SPIRIT procède à l'indexation automatique du texte. La recherche est automatique à partir de l'indexation de la question. Le traitement de SPIRIT est linguistique (analyseur morphologique et syntaxique) et statistique.

DARWIN enfin procède à l'indexation automatique du texte par une analyse linguistique (niveau syntaxique).
Les caractéristiques fonctionnelles donnent beaucoup plus d'avantages à SPIRIT et DARWIN du point de vue commercial, ils ne s'appuient sur aucun thésaurus (contrairement à ALETH) beaucoup de sites se sont orientés vers SPIRIT ou DARWIN pour éviter les coûts relatifs à l'établissement et la maintenance d'un thésaurus. Les choix d'ALETH sont dictés surtout par la mise en valeur et la récupération d'un thésaurus existant (cas de l'ENSP).

- *Enjeux*

L'indexation représente un enjeu économique important pour les producteurs et les distributeurs de l'information. Depuis quelques années, le coût de l'indexation automatique ou assistée est inférieur à celui de l'indexation classique (humaine).

"Nous arrivons à une situation dans laquelle, les coûts de l'indexation humaine sont supérieurs aux coûts d'une indexation automatique ou assistée, ce phénomène a braqué l'attention sur les possibilités de l'indexation automatique". Sur ce plan M. Van Slype a calculé qu'une fois le logiciel mis en place, il pouvait permettre une économie de 1/2 à 1/3 de l'effectif des indexeurs.(1)

Ce phénomène a des chances de s'accentuer de plus en plus, à la fois pour des raisons matérielles (scanners et ordinateurs de plus en plus rapides et puissants et moins chers) et logiciels (OCR et indexation de plus en plus performants).

Cela explique le positionnement des distributeurs et producteurs d'information dans une activité qui a un grand avenir devant elle.

# II - Le cas ENSP

L'ENSP est abonnée à de nombreuses publications périodiques (516 titres) principalement sur les domaines de la santé et de l'administration.

La bibliothèque contient également des ouvrages et des thèses dont le nombre est évalué à 25 OOO.

Une base de données gérée sous TEXTO permet l'accès au fonds.

Un interface en langage naturel réalisé par CYBERNETIX sera prochainement implanté.

Une base terminologique gérée par un logiciel GESTH est également en service.

***Identification des besoins :***

Toutes les revues reçues à l'ENSP sont dépouillées par les documentalistes qui sélectionnent les articles indexés. Cette indexation sélective est opérée grâce à la spécialisation de chacun des documentalistes dans un domaine spécifique.

Les articles qui ne sont pas indexés sont stockés et classés dans des dossiers thématiques que seuls le usagers internes peuvent consulter. L'usager externe est donc confronté à un problème de silence. Si le module Biblix est donc susceptible de résoudre les difficultés d'accès d'ordre formel à l'information, l'absence de l'information recherchée tient plus aux imperfections du mode de stockage et de la sélection de l'information.

L'ENSP désire donc améliorer les performances de l'indexation des documents.

La croissance de la demande conjuguée à un maintien des ressources disponibles justifie la recherche de moyens techniques permettant à l'offre de rejoindre la demande. Le volume et le flux d'informations rendent impossible l'analyse systématique et humaine de l'ensemble des articles.

L'école souhaite donc qu'un système d'indexation automatique, interfacé avec GESTH (gestionnaire de thésaurus) puisse être utilisé pour un certain nombre de revues.

Les articles seraient scannerisés et traités par un logiciel de reconnaissance optique de caractères et stockés sur un support informatique.

La typologie des articles ne peut être définie au préalable (texte simple, texte accompagné de graphique, d'images etc..;) dans la mesure où le choix final sera fonction de la performance scientifique du titre de périodique retenu.

Les contraintes de volume ont été fixées à 100 pages indexées par jour impliquant au maximum 1/2 journée d'aide-documentaliste par jour.

### *La solution Cybernetix*

CYBERNETIX avait réalisé l'interface d'interrogation en langage naturel.

Il s'était engagé à réaliser un système d'indexation automatique en se basant sur les mêmes outils qui avaient permis la réalisation de BIBLIX-GESTH.

Le système devait comprendre

- **un scanner pour éviter la saisie des textes ;**

- **un OCR permettant de constituer un fichier sous forme ASCII.**

Les graphiques et images seraient ignorés. Le fichier ASCII sera alors traité par le module d'indexation qui aura accès à la base terminologique. Elle produira alors une indexation prenant en compte la morphologie des mots, la synonymie et la sémantique.

Le repérage de termes rencontrés souvent mais n'appartenant pas à la base terminologique permettrait de les considérer comme des candidats descripteurs.

Un dialogue s'engagerait avec un documentaliste lui permettant de retenir une des options suivantes :

- **ignorer le terme**

- **le rattacher à un terme existant**

- **le définir comme un nouveau descripteur**

### *La solution GSI ERLI*

GSI ERLI a racheté à CYBERNETIX le département langage naturel. Ce rachat a entrainé les difficultés suivantes :

- les délais de réalisation n'ont pas été respectés. Le caractère opérationnel de BIBLIX GESTH prévu initialement pour fin 1991 ne pourra être vérifié que fin 1992.

- 3 études préalables ont été conduites par CYBERNETIX pour la réalisation du système d'indexation automatique.

- L'apparition d'un nouvel acteur nécessite la révision des contrats en cours pour assurer la pérénité des produits commandés.

Pourtant, la proposition de GSI ERLI diffère peu de celle faite par CYBERNETIX à l'exception du moteur de référence.

# III - Le poste d'indexation ENSP

## *Présentation générale*

La proposition faite par GSI ERLI est donc basée sur ALETH qui contient le modèle linguistique.

Il réalise une analyse syntaxico-sémantique par rapport à un thésaurus. Il comprend en plus des règles expertes de transformation qui permettent le passage des textes vers les descripteurs du thésaurus. Dans le cas de l'ENSP un certain nombre de paramètres seront préalablement tracés pour une meilleure adaptation du système au site.

## *Architecture générale du système*

La proposition de GSI ERLI comprend deux volets :

* les grandes fonctions du système

* l'architecture matérielle

**Les grandes fonctions :**

La chaine d'indexation automatique comprend un ensemble de fonctions :

- La scannerisation du document : permet l'obtention d'un fichier image à partir d'un document physique.

- L'OCR : permet l'obtention d'un fichier texte à partir d'un fichier image.

- L'obtention des descripteurs est l'étape la plus importante du processus, car il s'agit de l'indexation proprement dite. On obtient des descripteurs associés au texte à partir du fichier texte. (cette étape fera l'objet d'un développement particulier).

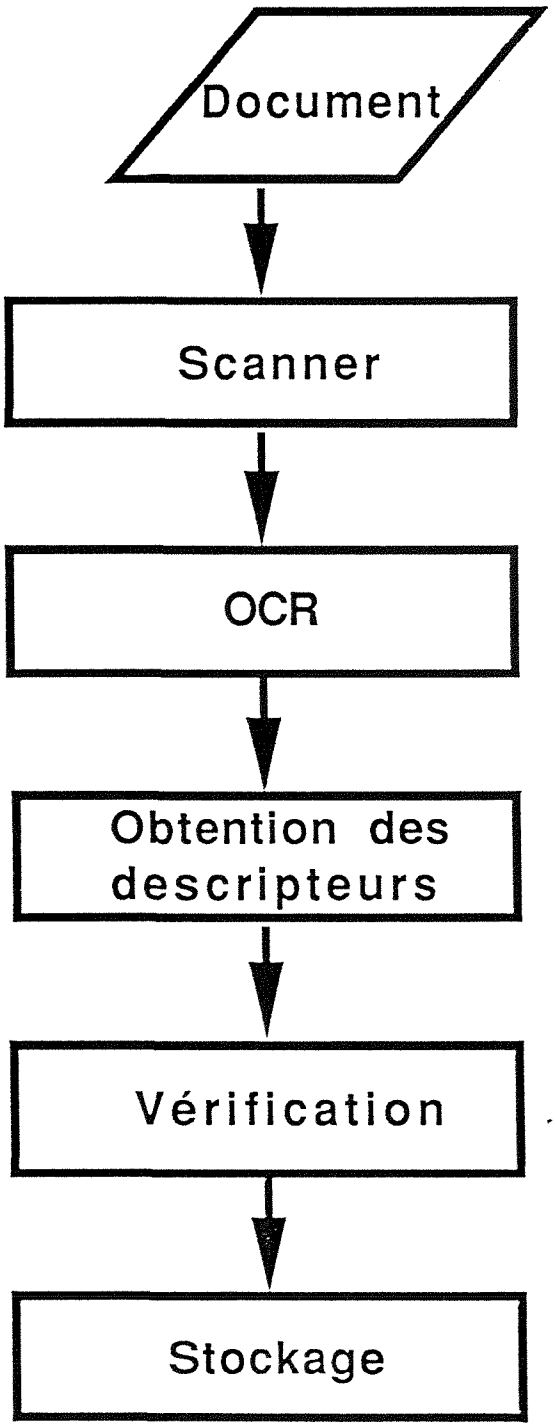
- Vérification : dans cette étape un documentaliste valide le résultat du module précédent ou éventuellement le modifie.

- Stockage : Finalement, les descripteurs sont stockés dans la base de données.

# La chaine d'indexation automatique

# -ENSP-

## Schéma des fonctions



**L'architecture matérielle :**

La chaine d'indexation :

GSI-ERLI suggère l'architecture matérielle suivante :
- Un poste scanner-OCR sur PC ou Macintosh
- Un poste d'indexation proprement dite (station Unix)
- L'Unisys contenant la base de données.

Justification du choix matériel :

Il faut préciser que les données initiales sont les suivantes:
La base de données est sous texto dans une machine Unisys de type Unix.

Le micro :

Il sera utilisé pour le scanner et l'OCR :

**Le scanner :**

- Les scanners sont pilotés par ordinateur relié (plus au moins sophistiqué)
- Le pilotage par un micro de type PC ou Macintosh est la solution la plus courante parce que moins chère et plus souple.

**L'OCR :**

- La situation est comparable à celle des scanners, en effet la solution la plus courante est un logiciel micro standard, PC ou Macintosh .
- La plupart des logiciels OCR offrent en plus la possibilité de piloter le scanner.
Cette dernière possibilité permet donc d'utiliser un seul micro pour le scanner et l'OCR.

La station Unix

La station Unisys étant déjà surchargée, il n'est donc pas souhaitable d'y installer les modules d'obtention des descripteurs et de vérification.

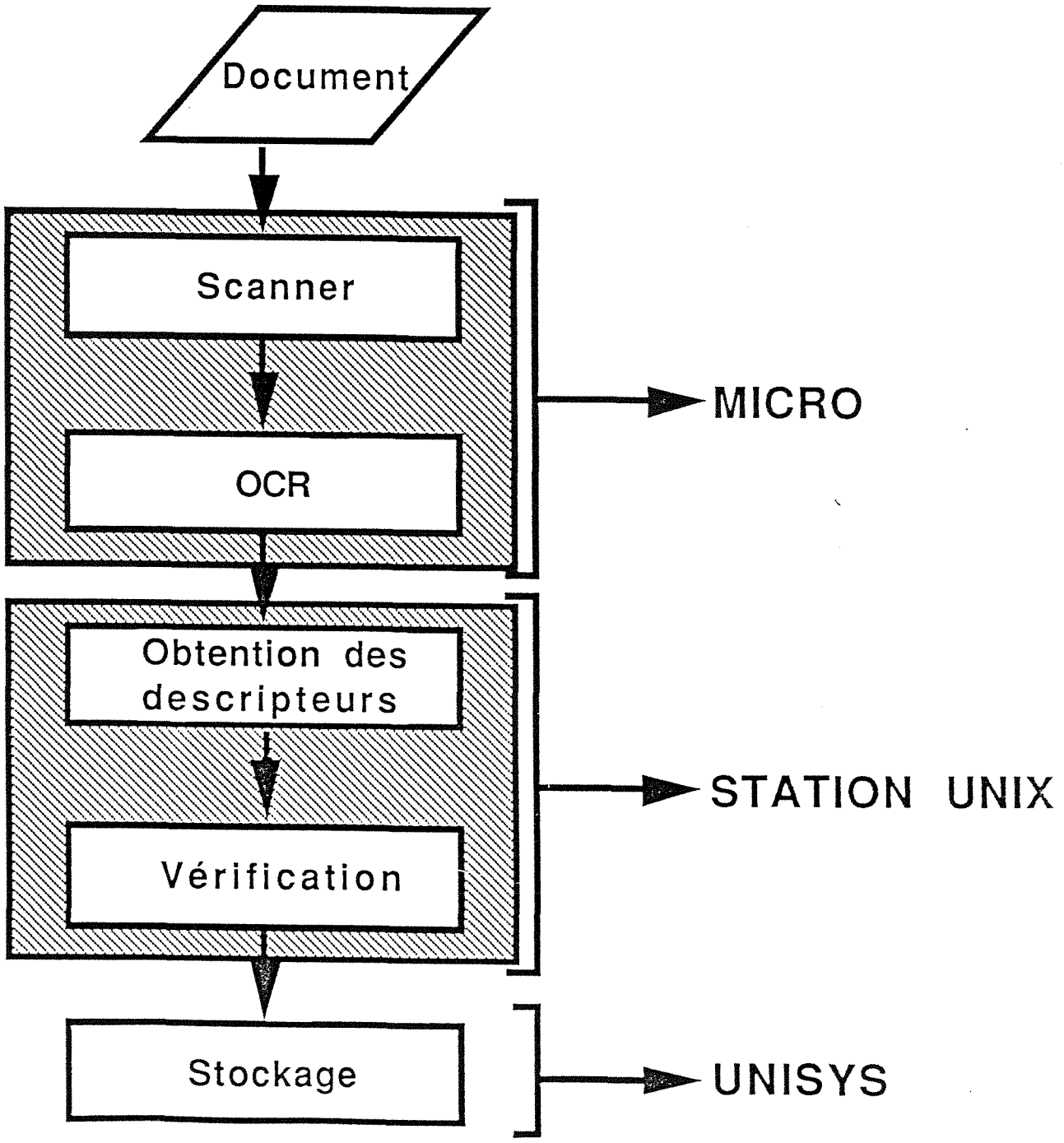
Ces modules sont d'une importance capitale pour la chaine d'indexation, le choix matériel adéquat pour parer à toute eventualité est donc à prendre en considération.
L'option pour une troisième machine (station Unix) s'avère judicieux, car l'installation de ces modules sur la même machine que le scanner et l'OCR, même si cela comporte un avantage économique certain présente néanmoins des inconvénients aux conséquences lourdes :

La chaine d'indexation automatique

-ENSP-

Architecture matérielle

(configuration matérielle)



- L'installation en même temps du pilotage du scanner, du logiciel OCR et de l'environnement d'indexation risquerait de surcharger le micro. En particulier, les logiciels d'indexation ont besoin d'une puissance de calcul importante.
- Les tâches de pilotage de scanner et d'application du logiciel OCR risqueraient d'encombrer le documentaliste, d'autant plus que ces tâches sont très simples et peuvent être effectuées par une personne non specialisée.
En définitif la séparation de ces deux tâches, permettra un travail en parallèle du poste scanner-OCR d'un côté, et de l'indexation automatique proprement dite de l'autre, ceci pour optimiser les performances.

L'OCR :

Pour le choix d'un logiciel de reconnaissance optique de caractère, GSI-ERLI a présenté une étude détaillée sur les produits et leurs caractéristiques techniques. L'étude s'est basée sur des critères de sélection pour l'ENSP.
Un échantillon représentatif de documents à indexer a été fourni par le service documentation pour permettre de bien cerner les besoins.
GSI-ERLI a retenu un ensemble de critères importants quant au choix du logiciel OCR :

- L'OCR doit être installé sur Macintosh et ceci pour des raisons de cohérence interne à l'ENSP et pour des raisons d'ergonomie. Ce critère a permis d'éliminer le choix éventuel d'un logiciel OCR sur PC.

- Il etait défini que la vitesse d'indexation visée est estimée dans un premier temps à 100 pages par jour,ce qui implique le choix d'un logiciel qui répond au critère de la "vitesse".
- Les documents à indexer par l'ENSP possèdent des graphiques, des images et des tables et ont plusieurs colonnes ou figures dans des pages à côté d'autres articles.

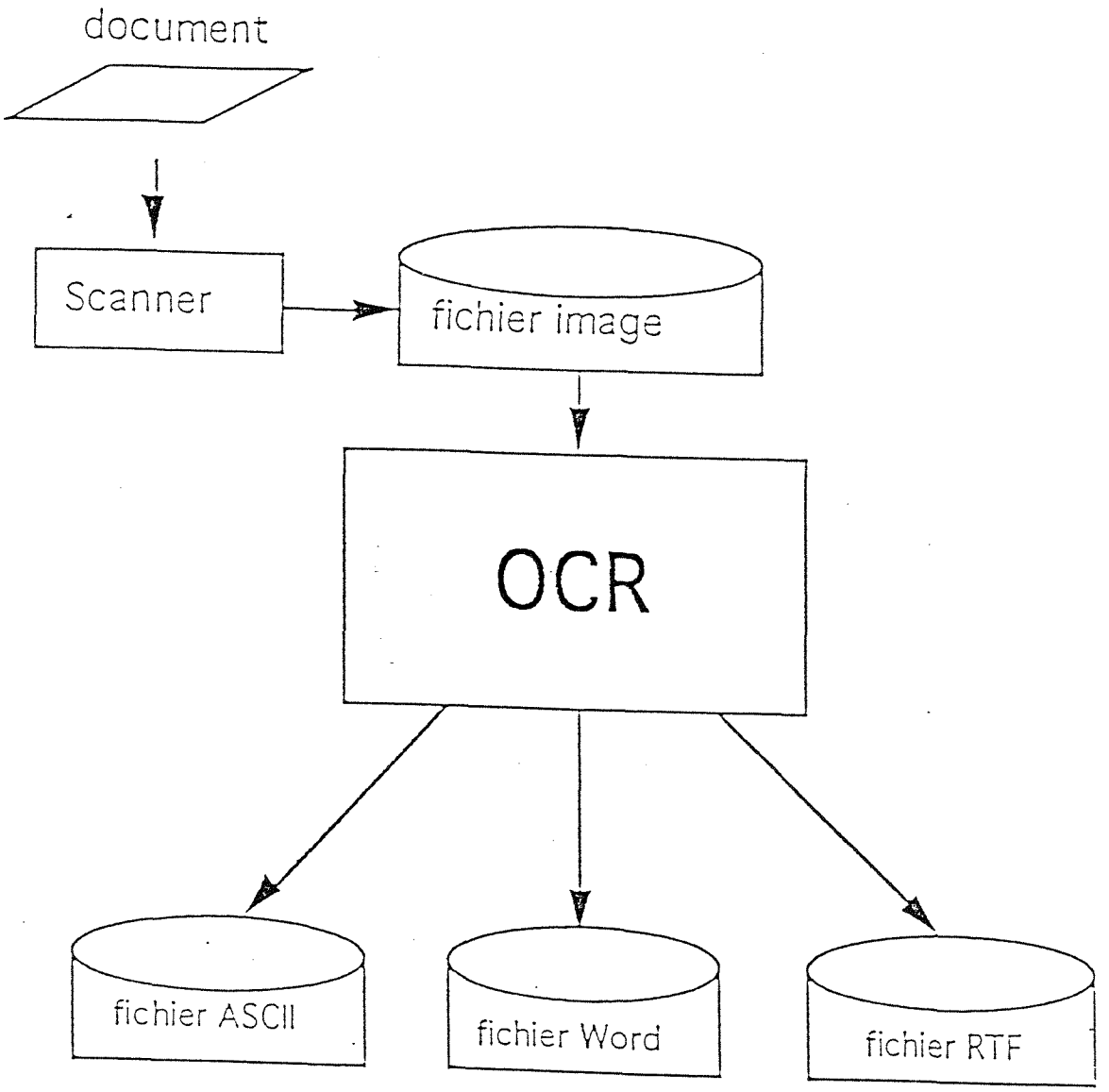
Le besoin est donc un logiciel qui ne soit pas perturbé par la présence de ces formes là.

A ce niveau GSI-ERLI propose deux solutions possibles : soit le logiciel permet de decouper à l'aide de la souris la zone de la page faisant partie de l'article, soit il le reconnait automatiquement (surtout dans le cas d'images).

- Par ailleurs GSI-ERLI suggère, pour les besoins d'indexation automatique, de garder une trace de l'importance de certaines phrases dans le texte qui seront interpretées comme des titres ou des sous-titres, des mots soulignés en gras, etc.. Ceci n'est possible qu'en ayant une sortie différente de la sortie ASCII.

GSI recommande une sortie RTF, SGML ou équivalente (standards non associés au traitement de textes).

- Enfin, les performances de reconnaissance dépendent du scanner utilisé. Curieusement, comme le note l'étude,ce n'est pas la peine d'utiliser des scanners haut de gamme : un scanner monochrome 300 dpi suffit. Par contre, avant reconnaissance il est utile de bien paramétrer la luminosité et le contraste du scanner. Les performances sont évidemment meilleures pour les scanners à plat de type photocopieuse (pas à la main ).


document
Scanner
fichier image
OCR
fichier ASCII
fichier Word
fichier RTF

**Conclusion de l'étude :**

L'étude effectuée a présenté cinq logiciels d'OCR pour Macintosh et scanner à plat (Accutext, Omnipage , Read-it, Readstar, Wordscan). Trois logiciels pré-selectionnés répondent effectivement aux besoins précités (Omnipage ,Accutext ,et Wordscan).GSI-ERLI suggére Omnipage, car il offre les meilleures performances en vitesse (400 à 1000 mots par minute),et en taux de reconnaissance (99%,voire 99,5% dans certains cas), ainsi que pour la taille de caractères acceptés (entre 6 et 8 pour les plus petits) et (entre 24 et 72 pour les plus grands ).

Le seul inconvénient qu'il semble présenter c'est l'impossibilité de traiter des documents en différé. Cet inconvénient est en partie résolu grâce à l'emploi (de toutes façon nécessaire) de deux ordinateurs différents : le Macintosh pour l'OCR et le SparcStation pour l'indexation proprement dite. La documentaliste peut en effet piloter l'indexation d'un document pendant que le Macintosh traite la reconnaissance de caractères du document suivant.

## *L'environnement d'indexation*

L'environnement de l'indexation comprend plusieurs opérations :

*La sélection du document à indexer :*

Les documents sont envoyés à la station Unix depuis le micro par l'intermédiaire d'un logiciel standard de type telnet.
Ils sont ensuite stockés dans un répertoire "à indexer", avec des éléments permettant de les identifier de façon unique pour éviter des problèmes éventuels au moment du stockage dans la base de données après indexation.
Cet identifiant peut être un code ou un ensemble de champs.
La fonction de sélection du document doit permettre de choisir un document parmi ceux qui sont stockés dans ce répertoire.
Cette sélection se fait dans une fenêtre appelée "indexation".

*Obtention des descripteurs :*

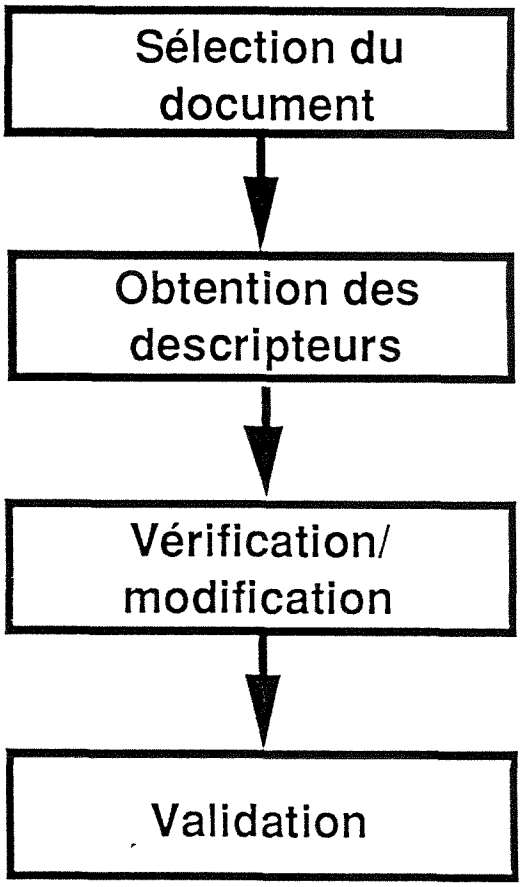
Dans cette étape, le travail d'indexation automatique proprement dite est fait. Le résultat est un fichier contenant des descripteurs, éventuellement commentés, toujours associé aux éléments identifiant le document.
Du point de vue de l'interface utilisateur, une fois cette étape achevée, une fenêtre contenant les descripteurs est affichée.

*Vérification/Modification :*

Ce bloc contient au fait plusieurs fonctions :

- La visualisation : implique le parcours (scrolling) dans une fenêtre ("description du texte indexé") de tous les descripteurs sélectionnés et d'en sélectionner un, pour point de départ des fonctions suivantes :

# L'environnement d'indexation



- La navigation : permet de situer un descripteur sélectionné dans le thésaurus. Cette navigation se fait dans une fenêtre différente de celle de visualisation.
En fait, la navigation se fait dans une fenêtre Gesth ou Alethgraf.
Le menu de la fenêtre "description du texte à indexer" offre un choix ("consulter") qui permet le lancement de Gesth et/ou Alethgraf focalisé sur le descripteur sélectionné.

- La modification : Elle peut avoir plusieurs formes :

  - Suppression pure et simple d'un descripteur, ou
  - Ajout d'un autre descripteur séléctionné dans la fenêtre Gesth ou Alethgraf.

*Validation :*

C'est une opération incluse dans le menu de la fenêtre principale "indexation",qui permet de valider le résultat, qui est alors stocké soit dans un répértoire "vers BD", soit directement dans la base de données.

## *Alethgraf*

C'est un outil de consultation graphique avec une approche hypertexte, conçu par GSI-ERLI.
C'est un outil interactif sous SunView, permettant de consulter des bases Aleth, en particulier en représentant graphiquement les grappes relationnelles des lexies. On peut également consulter la base "à plat" (sous forme textuelle), lister les extensions de ces lexies et enfin sélectionner ces lexies pour "relancer" un graphe.
La principale fonctionnalité d'Alethgraf est de visualiser de manière graphique une partie du réseau lexico-sémantique de la base.
La construction de ce graphe se fait comme suit : à partir d'une lexie de départ, on explore une liste de relations (définie par l'utilisateur).On obtient ainsi un ensemble de léxies auxquelles on applique le même traitement. Le nombre d'itérations de ce processus est dit profondeur de recherche dans le graphe. Une fois cette profondeur atteinte,on "boucle" le graphe en parcourant les relations du dernier ensemble de lexies atteintes qui ont pour image des lexies déja obtenues. L'ensemble des lexies et des relations parcourues sont alors presentés sous forme d'un graphe qui a pour noeuds les lexies obtenues et pour arcs les relations joignant ces lexies.

# IV - Spécifications fonctionnelles

L'architecture générale du système d'indexation présente -comme on l'a vu - les grandes opérations appelées "grandes fonctions", les spécificités du fonctionnement présentent quant à elles l'aspect concret de la demarche effectuée par le système.
Par ailleurs il semble important d'accorder un intérêt particulier à l'étape d'indexation proprement dite (obtention des descripteurs), cette étape est le noyau d'application basé sur l'outil Aleth.

### *Les blocs fonctionnels :*

Comme précisé ailleurs, GSI-ERLI souhaite regrouper la scannerisation et l'OCR dans un même bloc fonctionnel, car ces deux tâches bien que distinctes sont pilotées par le même logiciel.
GSI-ERLI propose donc les blocs fonctionnels suivants:

- *Scanner/OCR :*

| | |
|---|---|
| Entrée : | Document papier en français dans le domaine du thésaurus de l'ENSP , A4 ,bonne qualité d'impression. |
| Interface humaine : | Pilotage du scanner,découpage du document, correction eventuelle de caractères non prévus,sauvegarde du fichier de sortie. |
| Sortie : | fichier texte sous format RTF dans un répértoire "à transmettre" dans le micro,contenant dans son nom des éléments suffisants pour l'identifier de façon unique (code ou champs). |

- *Transmission :*

| | |
|---|---|
| Entrée : | fichier texte sous format RTF dans un répértoire "à transmettre" dans le micro, contenant dans son nom contenant des éléments suffisants pour l'identifier de façon unique. |
| Interface humaine : | Lancement de la commande de transmission. |
| Sortie : | fichier texte sous format RTF dans un répertoire "à indexer" dans la station Unix, contenant dans son nom des éléments suffisants pour l'identifier de façon unique |

- *Sélection du document :*

Entrée : fichier texte sous format RTF dans un répertoire "à indexer" dans la station Unix.

Interface humaine : sélection d'un document dans une liste.

Sortie : un fichier texte sous format RTF dans le répertoire à indexer est sélectionné (son nom est stocké).

- *Obtention des descripteurs :*

Entrée : le fichier texte sous format RTF sélectionné et son nom.

Interface humaine : lancement.

Sortie : un fichier contenant une liste de descripteurs, éventuellement avec des renseignements, entre autres sur leur fréquence dans le document. Le nom de ce fichier est associé de façon unique au nom du fichier RTF.

- *Vérification/Modification :*

Entrée : le fichier contenant une liste de descripteurs, éventuellement avec des renseignements, entre autres sur leur fréquence dans le document.

Interface humaine : manipulation de la liste et navigation dans le thésaurus, suppression et ajout de descripteurs.

Sortie : un fichier contenant une liste de descripteurs, éventuellement avec des renseignements, entre autres sur leur fréquence dans le document.Le nom de ce fichier et le même que celui de la sortie précédente.

- *Validation :*

Entrée : le fichier contenant une liste de descripteurs, éventuellement avec des renseignements, entre autres sur leur fréquence dans le document.

Interface humaine : lancement

Sortie : soit: a) un fichier contenant une liste de descripteurs,dans le répértoire "vers la BD" de la station Unix, et dont le nom identifie de façon unique le document.

soit : b) l'insertion automatique des descripteurs dans le champ thème du document pré-catalogué dans la BD.(cas du précatalogage).

Par ailleurs, et parallèlement à ces tâches essentielles, d'autres tâches sont effectuées : un historique de l'activité d'indexation, une liste de candidats descripteurs avec les identifiants des documents où ils apparaissent, des traces, des programmes, etc...

### *L'obtention des descripteurs*

Cette étape constitue comme nous l'avons dit le noyau de l'application. C'est l'indexation proprement dite, l'entrée de cette étape est un fichier contenant le texte de l'article à indexer sous format RTF,le nom de ce fichier contient des éléments l'identifiant.

Le contenu de ce fichier est traité par un module basé sur l'outil ALETH de GSI-ERLI.

### *ALETH*

A cette étape il est intéressant de savoir les contraintes auxquelles ce genre de conception (Aleth) doit faire face, car on est loin d'avoir une solution "clé en main", la mise en place d'un système de ce type nécessite un effort considérable
de part et d'autre (GSI-ERLI et l'ENSP) en terme d'application. Le concepteur qui a conçu ce système basé sur l'analyse linguistique (pièce centrale du dispositif d'indexation Aleth) est soumis à la contrainte du paramétrage, Aleth est une "boite à outil" qui est capable de réaliser diverses applications : (Pas spécialement l'indexation automatique), chacune de ces applications fait appel à une étude approfondie des besoins auxquels elle doit répondre,ces besoins peuvent être originaux et susceptibles d'évolution. Il serait aléatoire -par exemple- d'établir une comparaison entre deux systèmes d'indexation de deux sites différents (EDF et l'ENSP) même s'ils utilisent le même outil (Aleth), car les besoins et les objectifs sont différents.

#### *Présentation d'ALETH*

GSI-ERLI utilise une "boite à outil" appelée ALETH pour réaliser des applications de linguistique automatique et des applications documentaires. C'est un atelier de développement de modules langage naturel, qui comprend un module linguistique en réalisant une analyse syntaxico-sémantique par rapport à un thésaurus, en plus des régles expertes de transformation qui permettent le passage des textes vers les descripteurs du thésaurus (règles d'indexation).

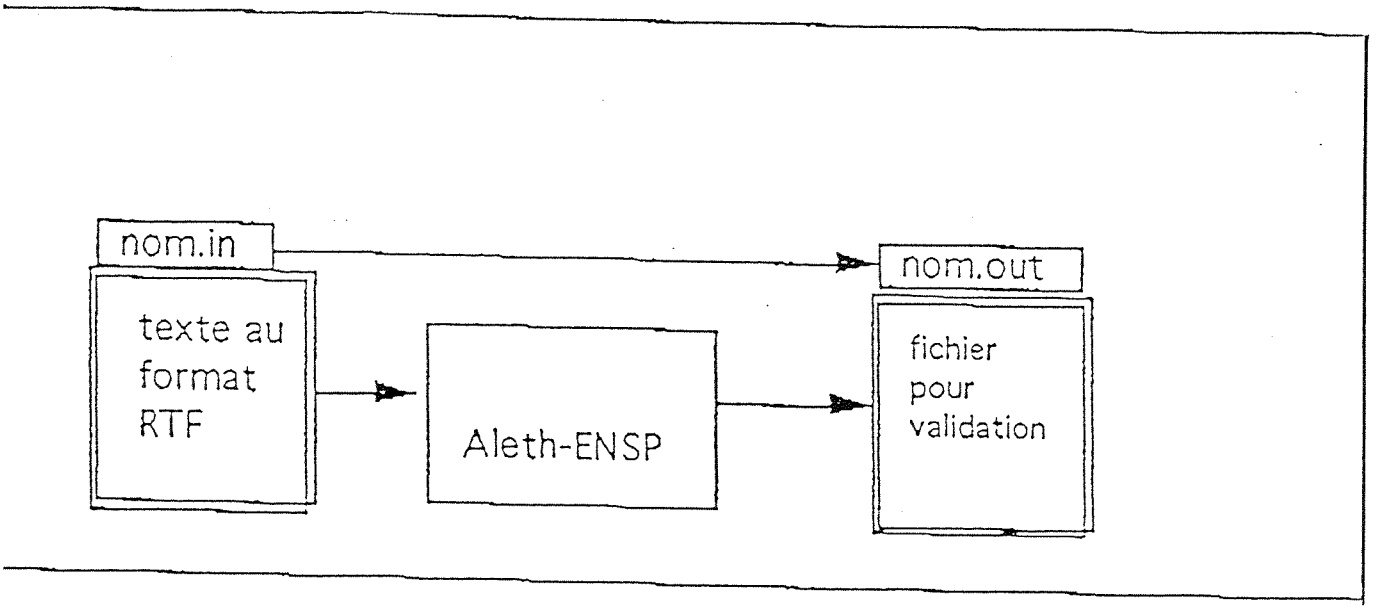
Cet atelier regroupe des connaissances, des fonctions , des procédures :

- Les connaissances sont entre autre sous forme de dictionnaires et de grammaires.

Les fonctions sont de deux types :

- fonctions d'accès aux connaissances.
- fonctions d'analyse d'énoncés linguistiques.

# Obtention des descripteurs



Les dernières reposent sur les premières et donnent en général en sortie une représentation sémantique en rapport avec le référentiel applicatif (cas typique le thésaurus)

- Les procédures quant à elles, concernent le mode de fonctionnement (batch ou interactif) et les entrées-sorties.

En particulier, ALETH peut effectuer les opérations suivantes sur les phrases :

- Lemmatisation
- Consultation d'un dictionnaire
- Analyse syntaxique
- Extraction des constituants suivant des critères
- Transformation d'un enoncé suivant des régles
- Recherche des équivalences avec un réferentiel

Par contre, les opérations qu'il ne peut effectuer sont les suivantes :

- Génération de texte
- Traduction
- Gestion d'informations multiphrastiques.

GSI-ERLI dispose pour le moment de trois dictionnaires : Français-Anglais-Italien

Ces trois langues sont décrites de façon générale à l'aide de grammaires, par ailleurs chaque application développe des traitements spécialisés.

Pour améliorer ces performances, GSI-ERLI participe à deux programmes :

-Le programme Eureka-Genelex qui vise à la normalisation et à l'enrichissement des dictionnaires.
- Le programme Eureka-Graal qui permettra la normalisation et l'enrichissement des grammaires.

*Réferences d'applications ALETH :*

Aleth a été utilisé dans plusieurs applications opérationnelles :

Application Aleth/EDF:

L'expérience de la direction des études et des recherches basée sur Aleth. Le niveau applicatif a été conduit et developpé par l'EDF.
Le système est basé sur le thésaurus français-anglais EDF.DOC qui contient plus de 20000 descripteurs.
Les conclusions principales de cette expérience sont les suivantes :

- Le taux de recouvrement moyen avec l'indexation manuelle est de 50% .
- Le taux de pertinence de l'indexation automatique est jugé de 70% .
- L'homogénéité des indexations est très grande
- Un des problèmes rencontrés est la polysémie des termes
- La qualité de l'indexation est directement proportionnelle à celle de la base terminologique.

*Application Aleth/Le monde :*

Un autre exemple d'application d'Aleth est le système télématique d'annonces du journal Le monde, réalisé par GSI-ERLI. Le but de ce service est de fournir une sélection pertinente d'annonces d'offres d'emploi vis-à-vis d'un candidat et d'un CV donné. Les annonces et les CV sont indexés automatiquement ,leur appariement étant assuré par un système expert.

*Application Aleth/L'annuaire électronique italien :*

Pour l'équivalent italien du service 11 du minitel et plus particulièrement les rubriques ("PGE",pages jaunes électroniques),un système d'indexation automatique de fiches a été mis en place.
Cette indexation est faite par Aleth à partir d'un thésaurus à rôle décrivant justement les rubriques. Ces développements ont été faits par la société SARITEL.

*ALETH/ENSP :*

Comme on a vu plus haut Aleth nécessite d'être paramétré et adapté pour chaque application. C'est le cas pour l'ENSP, ces adaptations concernent deux grandes étapes du travail :
La première étape, qui représente le gros du travail, car c'est le développement de l'indexation proprement dite.

## *l'intégration du module Aleth à biblix*

Mise au point des règles :

La grammaire générale du français et les règles de base d'indexation ne sont pas suffisantes pour obtenir une indexation cohérente. Le paramétrage des règles s'avère donc nécessaire pour obtenir les descripteurs du thésaurus ENSP, et surtout pour exploiter de façon utile les connaissances qui s'y trouvent (relations). D'où l'interêt porté par GSI-ERLI pour l'étude initiale des corpus de textes à indexer à l'ENSP, et la comparaison manuelle avec le thésaurus.
L'étude préalable présentée par GSI-ERLI stipule que l'étape definitive et la mise au point ne peut se réaliser qu'après test grandeur nature.

Adaptation Aleth et analyse RTF:

L'analyse du format RTF servira à distinguer les titres et les sous-titres du document, d'un poids supérieur à celui d'une phrase normale du corps du document
Un fichier sous format RTF (rich text format) est un fichier ASCII qui contient non seulement le texte d'un document, mais aussi des informations sur la taille des caractères,les paragraphes, etc...
L'option prise par GSI-ERLI pour le format RTF exprime en quelque sorte l'importance accordée aux titres et sous-titres, comme étant un principe d'indexation, qui ne peut se réaliser autrement. GSI-ERLI justifie ce choix par les arguments suivants :
Dans un article, en général, les titres et les sous-titres décrivent en une seule expression le plus important du paragraphe ou de la section qui suit. Ainsi "Analyse du format RTF" décrit en gros le contenu de cette section. Bien entendu, il ne faut pas réduire l'indexation aux titres, cela donnerait une liste très pauvre et imparfaite de descripteurs. Mais il est important, dans l'évaluation des expressions pouvant décrire le document, de donner un poids particulier aux titres et aux sous-titres, ce qui est possible grâce à l'analyse du format RTF.

Adaptation de Gesth format Aleth :

La représentation des connaissances dans Gesth, répondent à un modèle conceptuel légèrement différent de celui des connaissances d'Aleth, ces différences se situent aussi bien au niveau des fichiers sortie de Gesth (.U et exportations ASCII), qu'au niveau du traitement interne du programme.
L'évaluation de GSI-ERLI précise qu'il aurait été possible de développer un programme prenant en entrée la base terminologique de Gesth et obtenant en sortie une base de connaissances au format Aleth. L'inconvénient de cette approche est la nécessité d'appliquer ce programme à toute la base terminologique après chaque modification dans Gesth, si minime soit elle.
C'est pourquoi une modification légère de Gesth est nécessaire de façon à ce qu'il puisse produire directement en sortie une base de connaissances au format Aleth.

Intégration du module Aleth dans Biblix :
(Le module d'interrogation)

La détection de descripteurs dans les questions se fait actuellement dans Biblix par l'intermédiaire d'un module appelé "Biblix-Nthes". Ce module reçoit en entrée de mots avec des informations de la base terminologique, et produit en sortie une liste de combinaisons de descripteurs pondérées. La combinaison de poids le plus fort est celle qui sera utilisée par Biblix pour le dialogue avec la base de données.
Il est établi que le module d'interrogation Biblix-Nthes malgré sa performance, est trop lent, car la recherche de descripteurs dans les phrases ne tient pas compte des règles et se fait en quelque sorte "tous azimuts".GSI-ERLI propose donc de le remplacer par son équivalent d'Aleth, qui lui a l'avantage d'être rapide et plus robuste. En plus,il sera de toutes façon utilisé pour l'indexation. Ce qui permet d'avoir un seul module de sélection de descripteurs pour l'indexation et l'interrogation, qui est par ailleurs plus performant.

- *L'interface utilisateur*

L'interface du poste d'indexation comprend :

-La sélection du document.
-Le lancement de l'obtention des descripteurs.
-Le lancement de la vérification/modification.
-La validation.

Ceci est organisé de la façon suivante :

*Fenêtre "Indexation" (bloc principal)

Choix:

- sélection du document
(ouvre une boite de dialogue de sélection de fichiers, qui se se positionne par défaut sur le premièr fichier du répertoire "à indexer")

- obtention des descripteurs
(déclenche le module de même nom et ouvre la fenêtre "description du texte indexé")

- validation
(la verification/modification est finie,l'indexation courante est gardée,le fichier source est retiré du répertoire "à indexer",les autres fenêtres sont fermées)

- quitter
(dialogue de confirmation si indexation sans validation)

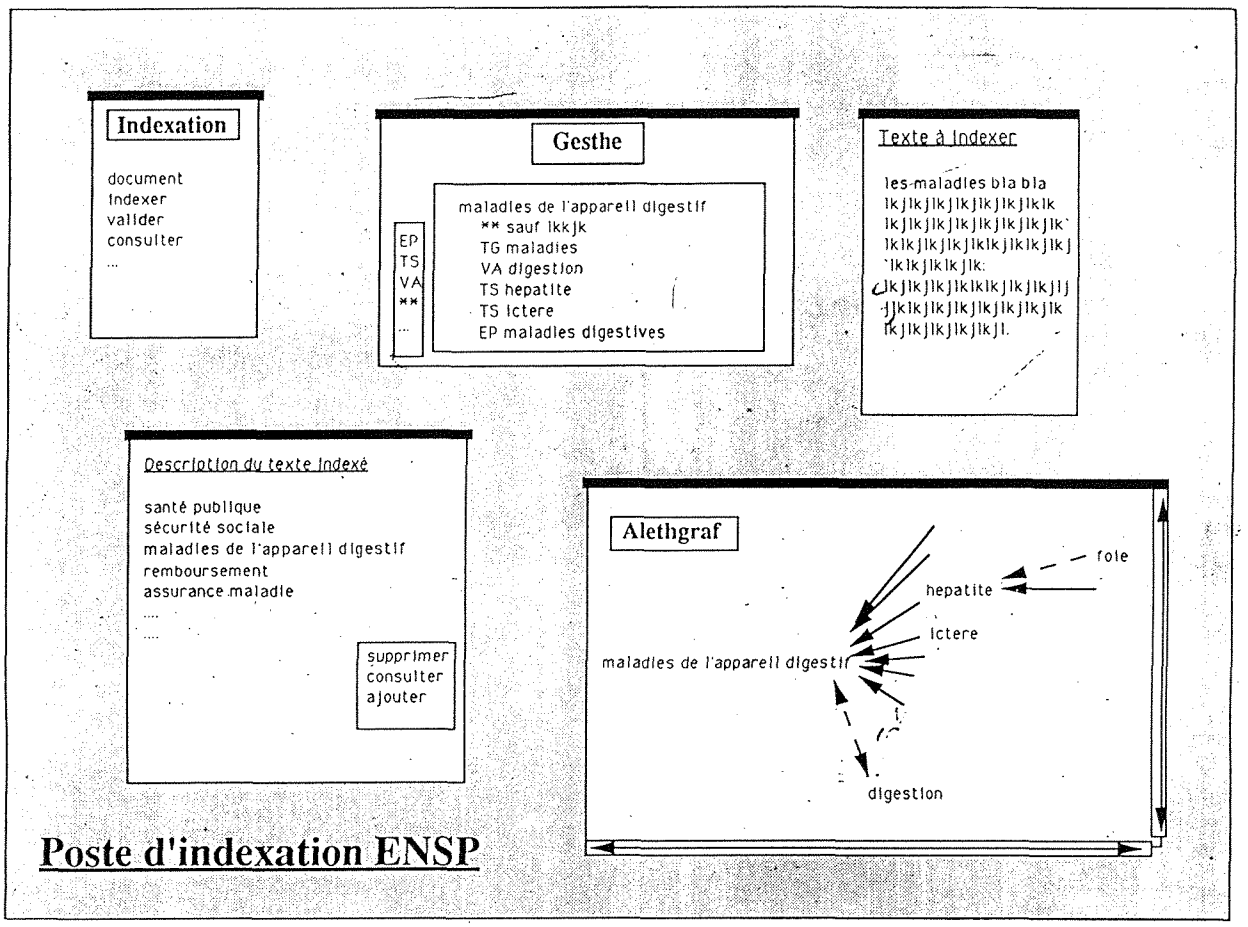
*Fenêtre "texte à indexer"
(à cette étape il s'agit uniquement de visualisation)

*Fenêtre "description du texte indexé".
affiche la liste de descripteurs obtenus par l'indexation automatique et comporte un menu :

- supprimer (suppression du descripteur courant)
(le descripteur est tout simplement supprimé)

- consulter (vérification/modification à partir du descripteur)
(lance Gesthe et/ou Alethgraf positionné(s) sur le descripteur courant.L'utilisateur est libre de naviguer dans le thésaurus,et sélectionner un autre descripteur)

- ajouter (ajout d'un descripteur séléctionné)
(ajout soit le descripteur sélectionné dans Gesthe ou Alethgraf, soit le dernier descripteur supprimé).




nêtre Indexation:
ioix "document" : dialogue sur le document à indexer, lançant éventuellement le pilotage du scanner et du logiciel OCR
ioix "indexer" : part du document sélectionner, et en obtient une description, autrement dit une liste de descripteurs,
uis une fenêtre est ouverte avec cette liste : voir Fenêtre Description du texte indexé
ioix valider : une fois la liste de descripteurs visualisée et éventuellement modifiée, envoie l'indexation à la base
ioix consulter : ouvre les fenêtres Gesthe et Alethgraf, qui permettent de visualiser le contexte de la description

enêtre Gesthe :
enêtre Gesthe en visualisation, communiquant avec les autres fenêtres

enêtre Texte à indexer
ontient le texte du document sélectionné dans le choix "document" de la fenêtre indexation

enêtre Description du texte indexé
ontient la liste de descripteurs produite par le choix "indexer" de la fenêtre Indexation. Cette liste peut être modifiée
l'aide d'un menu propre à cette fenêtre.
hoix "supprimer" : supprime le descripteur sélectionné
hoix "consulter" : affiche dans Gesthe et dans Alethgraf les informations associées au descripteur sélectionné
hoix "ajouter" : ajoute un descripteur sélectionné dans Gesthe ou Alethgraf

enêtre Alethgraf
l s'agit du logiciel de Gsi-Erli qui permet la consultation graphique d'une base terminologique.

# Conclusion

# Conclusion

Il était souhaitable de voir évoluer le système en grandeur nature pour évaluer ses performances, car ce genre de système même s'il parait fiable du point de vue théorique, réserve souvent des surprises quant à leur application réelle. L'expérience BIBLIX/GESTHE est là pour en témoigner.

Les systèmes d'indexation automatique ou assistée réalisés à partir d'analyseurs linguistiques n'offrent pas une solution "clé en main". Bien au contraire, à la lecture des contraintes liées à l'installation du système ALETH/ENSP, on se rend compte de l'effort à consentir pour la définition des dimensions d'application à chaque site (paramétrage de l'application). En effet, si l'analyse morpho-lexicale représente une phase constante de l'application, le traitement syntaxico-sémantique nécessite quant à lui le développement de grammaires adéquates relatives au domaine traité par le fonds documentaire de l'ENSP. Les "règles" doivent dépendre du contenu de la base de connaissance. Cette phase représente le gros du travail qui attend l'ENSP pour fournir à ERLI un ensemble exhaustif et applicatif des connaissances du domaine de la santé publique et des sciences annexes sur lequel se construira le module syntaxico-sémantique.

Les tests grandeur nature permettront enfin de tester les réelles capacités du système pour l'interrogation et la pertinence de l'indexation.

# BIBLIOGRAPHIE

## *I - REFERENCES CITEES*

(1) - CHAUMIER (J), DEJEAN (M). "L'indexation documentaire de l'analyse concepturelle humaine à l'analyse automatique morphosyntaxique". Documentaliste, nov.-dec. 1990, 27 (6),p 275-279

(2) - MENON (B). "L'indexation automatique et l'intelligence artificielle : quelques questions de stratégies".
Image et indexation automatique dans l'information scientifique et technique. Cours INRIA, 1988.
Ingénierie linguistique et documentaire : Recueil d'articles, mise à jour, mai 1991, P. 105-137.

## *II - REFERENCES UTILISEES*

- BONNET (A). "L'informatisation des Langues naturelles".
L'informatique professionnelle, Dec. 1989, p. 39-46

- BERTRAND (R), (COLL). "Micro-ordinateur et traitement de l'information"
Paris NTD, mise à jour 1992

- BERTRAND (R) (COLL) "Micro-informatique et documentation" PARIS :
La documentation Française, 1987

- CHAUMIER (J), DEJEAN (M). "l'indexation assistée par ordinateur. Principes et méthodes". Documentaliste 1992, VOL 29, n° 1, P. 3-6.

- CHAUMIER (J). "Le traitement linguistique de l'information documentaire. L'analyse documentaire", - PARIS : Entreprise moderne d'édition, 1977.

- CHARTRON (G), DALBIN (S), MONTEIL (MG), VERILLON (M). Indexation manuelle et indexation automatique : dépasser les oppositions". Documentaliste, oct. 1989, 26 (4-5) : p. 181-186.

- N'DONG (M), "Une interface utilisateur pour l'interrogation en langage naturel pour la base bibliographique de l'ENSP". Mémoire DESSID. ENSSIB LYON, 1991.

- RIVIER (A), "Construction des langages d'indexation - Aspects théoriques". Documentaliste, nov-dec. 1990, 27 (n°6), P. 263-270.

- TARAVELLA (J.P.). "L'indexation automatique en France. Etat de la recherche, problèmes rencontrés et analyse de produits disponibles". Mémoire DESS. IEP PARIS, 1990.

- VAN SLYPE (G). "Les langages d'indexation : conception, construction et utilisation dans les systèmes documentaires". PARIS : Les éditions d'organisation, 1987.

- VIDAL (M). "L'indexation assistée par ordinateur : Les clés, les ouvertures. BASES 1990, n° 48, P. 5-6.

# Table des matières

SOMMAIRE ..... 5

INTRODUCTION ..... 7

CADRE ET OBJECTIF DU STAGE ..... 8

PREMIERE PARTIE ..... 9

I - Présentation ..... 9

Missions de l'ENSP ..... 9

II - Le service commun de la Documentation ..... 12

1) organisation ..... 12

a) l'E.P.C. ..... 12
b) les Editions ..... 12
c) la Bibliothèque ..... 14

2) contexte d'informatisation ..... 14

a) état des lieux ..... 14
b) développements actuels ..... 16
c) la solution Biblix Gesth ..... 18

DEUXIEME PARTIE ..... 23

I - Identification des concepts ..... 24

l'indexation ..... 24
automatiser l'indexation ..... 24
procédures d'indexation ..... 25
Evolution des systèmes et enjeux ..... 28

II - Le cas ENSP ..... 30

Identification des besoins ..... 30
La solution Cybernetix ..... 31
La solution GSI ERLI ..... 31

III - Le poste d'indexation ENSP ..... 32

Présentation générale ..... 32
Architecture générale du système ..... 32
l'environnement d'indexation ..... 38
Alethgraf ..... 40

IV - Spécifications fonctionnelles ..... 41

Les blocs fonctionnels ..... 41
l'obtention des descripteurs ..... 43
Aleth ..... 43
Intégration du module Aleth à Biblix ..... 46

CONCLUSION ..... 51

BIBLIOGRAPHIE ..... 53


*9596429*