E.N.S.S.I.B.
ECOLE **N**ATIONALE **S**UPERIEURE
DES **S**CIENCES DE L'**I**NFORMATION
ET DES **B**IBLIOTHEQUES

UNIVERSITE
CLAUDE BERNARD
LYON I

D.E.S.S. en INFORMATIQUE DOCUMENTAIRE

Rapport de Stage

Inriathèque
Document électronique

KEBIR Hamza

Sous la direction de

Mme TOUZEAU
Responsable du Centre de documentation
INRIA - Rocquencourt
Domaine de Voluceau 78153 Le Chesnay Cedex

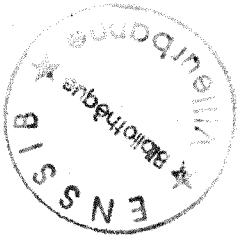
1995

E.N.S.S.I.B.
ECOLE NATIONALE SUPERIEURE
DES SCIENCES DE L'INFORMATION
ET DES BIBLIOTHEQUES

UNIVERSITE
CLAUDE BERNARD
LYON I

D.E.S.S. en INFORMATIQUE DOCUMENTAIRE

Rapport de Stage



Inriathèque
Document électronique

KEBIR Hamza

Sous la direction de

Mme TOUZEAU
Responsable du Centre de documentation
INRIA - Rocquencourt
Domaine de Voluceau 78153 Le Chesnay Cedex

1995

Inriathèque : Document électronique

KEBIR Hamza

Résumé :

L'Inriathèque est un document papier relevant de manière hebdomadaire toutes les nouvelles acquisitions du centre de documentation. Le travail a consisté à la gestion électronique de ce document. Par conséquent, il a fallu étudier les phases d'acquisition, de traitement et de diffusion des données. Les principaux outils, logiciels ou normes standards utilisés pour ce travail sont : HTTP, HTML, WAIS, SCRIPT. Une démarche de GED a été étudiée et mise en application pour gérer ce document désormais accessible sur une interface WEB.

Principaux descripteurs :

- Serveur
- Scanner, numérisation
- Abonnement, sommaire

Descripteurs libres :

- GED, Gestion Electronique de Document, document électronique
- OCR, Reconnaissance Optique de Caractère, digitalisation
- HTML, Hypertext Markup Language
- HTTP, Hypertext Transfer Protocol
- WAIS, Wide Area Information System
- WEB, WWW, World Wide Web

Abstract :

This report describes the method used for the electronic management of a hitherto weekly paper publication, Inriatheque, which lists all the news acquisitions of the documentation center at INRIA. The worse required a preliminary study of the different stages of data acquisition and processing and dissemination of the document. The main tools and standard normes used for the electronic management of this publication were : HTTP, HTML, WAIS, SCRIPT. Inriatheque is henceforth accessible via WEB.

Main keywords :

- Host
- Scanner, numerisation
- Subscription, contents

Free keywords :

- GED, Electronique Document
- OCR, optical recognition character, digitalisation
- HTML, Hypertext Markup Language
- HTTP, Hypertext Transfer Protocol
- WAIS, Wide Area Information System
- WEB, WWW, World Wide Web

REMERCIEMENTS

Dans le cadre du Diplôme d'Etudes Supérieures Spécialisées en Informatique Documentaire, j'ai effectué un stage d'une durée de quatre mois, du 6 Juin au 30 Septembre 1995 au centre de documentation de l'INRIA de Rocquencourt.

Le principal objectif de ce stage était de prendre contact avec le monde de l'entreprise, d'observer son mode de fonctionnement et de participer à son rythme de travail en mettant en pratique les connaissances acquises au cours des années universitaires.

Je tiens à remercier le personnel du centre de documentation, avec lequel il m'a été permis de collaborer, de découvrir ce que recouvraient les expressions telles que "le sens des responsabilités", "le souci de la cohérence", "le besoin d'organisation"...

Je remercie en particulier Madame TOUZEAU responsable du centre de documentation et Monsieur AUBRIE, informaticien, qui m'ont suivi au cours du stage et ont su rester attentifs et disponibles à toutes les questions que je leur ai posées.

Je tiens à exprimer ma reconnaissance à Monsieur MARCHAND Alain, pour son soutien et ses nombreux conseils.

Ainsi, le centre de documentation de l'INRIA a transformé mon stage en une expérience enrichissante.

TABLE DES MATIERES

INTRODUCTION

L'Inriathèque est une publication hebdomadaire du centre de documentation de l'INRIA de Rocquencourt. Elle a pour but de présenter les documents les plus récemment acquis. Sa diffusion a lieu en interne et en externe. Ma mission a consisté à étudier les diverses solutions pour une gestion électronique de ce document qui n'existait jusqu'à présent que sous forme papier.

L'Inriathèque contient notamment des sommaires de périodiques et de conférences, ainsi que la liste des rapports, des thèses et des ouvrages récemment acquis. De nombreux laboratoires, bibliothèques universitaires, organismes de recherche publics et privés sont abonnés à ce document. Cette publication suscite un réel engouement auprès du public extérieur mais également en interne, auprès des chercheurs des différentes unités de recherche de l'INRIA.

Nous présenterons dans une première partie, l'INRIA, son centre de documentation, ainsi que la participation à la vie du service.

Nous aborderons dans une seconde partie, la méthodologie adoptée pour mener à bien ce travail, celle-ci comprend en premier lieu une analyse de la demande et des besoins et d'autre part la démarche GED pour laquelle nous avons opté. Cette dernière repose essentiellement dans l'acquisition, le traitement et la diffusion des données.
L'acquisition de l'information intègre l'abonnement auprès de producteurs d'informations électroniques, la numérisation de documents papiers et enfin, la collecte de données auprès d'éditeurs qui fournissent un service en ligne sur Internet.
Le traitement des données a requis davantage de compétences techniques. La programmation, ainsi que l'indexation ont abouti à la création d'une chaîne de traitements automatiques.
La diffusion des données par le biais de l'Inriathèque, a nécessité d'étudier et d'appliquer des normes et protocoles relatifs au travail en réseau. Les principaux outils utilisés sont HTML, WAIS et HTTP qui conduisent respectivement à présenter, indexer, et communiquer les données. Par conséquent, une interface spécifique a été créée sur le serveur WWW de l'INRIA, et ce, à des fins d'interrogations, consultations et demandes de photocopies d'articles.

1ère PARTIE : L'INRIA ET LA PARTICIPATION A LA VIE DU CENTRE DE DOCUMENTATION

Chapitre 1 : L'INRIA

Section 1 : L'organisation

L'INRIA est un établissement public à caractère scientifique et technologique (EPST) placé sous la tutelle du ministre chargé de l'Enseignement Supérieur et de la Recherche, et du ministre de l'industrie.
L'INRIA, dont le siège est Rocquencourt, est composé de cinq unités de recherche : Lorraine (Nancy-Metz), Rennes, Rhône-Alpes (Grenoble), Rocquencourt et Sophia Antipolis, ainsi qu'une unité de communication et d'information scientifique (UCIS), située elle aussi à Rocquencourt.
L'ensemble des centres de documentation des précédentes unités vient d'être retenu comme pôle associé de la Bibliothèque Nationale de France.

Mille quatre cents personnes travaillent à l'institut, dont un millier sont des scientifiques : chercheurs et ingénieurs permanents, chercheurs extérieurs de l'industrie, chercheurs de laboratoires de recherche publics, chercheurs étrangers invités, étudiants en thèse, stagiaires.

Les activités de l'INRIA comprennent la réalisation de systèmes expérimentaux, la recherche fondamentale et appliquée, le transfert de technologies, l'organisation d'échanges scientifiques internationaux et la diffusion des connaissances et du savoir-faire. Ces activités associent informaticiens, automaticiens, mathématiciens dans le cadre de projets de recherche regroupés dans six programmes.

Section 2 : Les missions

Les missions confiées à l'INRIA sont les suivantes :

⇨ entreprendre des recherches fondamentales et appliquées,
⇨ réaliser des systèmes expérimentaux,
⇨ organiser des échanges scientifiques internationaux,
⇨ assurer le transfert et la diffusion des connaissances et du savoir-faire,
⇨ contribuer à la valorisation des résultats des recherches,
⇨ contribuer, notamment par la formation, à des programmes de coopération avec les pays en voie de développement,
⇨ effectuer des expertises scientifiques,
⇨ contribuer à la normalisation

L'objectif est donc d'effectuer une recherche de haut niveau, et de transmettre les résultats de cette recherche aux étudiants, au monde économique et aux partenaires scientifiques et industriels.

Section 3 : Le site de Rocquencourt

Six cent cinquante personnes travaillent sur le site de Rocquencourt : cinq cents d'entre elles ont une activité de recherche scientifique, tandis que les autres (ingénieurs, techniciens et administratifs) participent directement au soutien et à la valorisation de cette recherche.
Les chercheurs de l'INRIA Rocquencourt dispensent des enseignements dans la plupart des formations doctorales des universités et des grandes écoles de la région parisienne.

Le centre de diffusion et l'UCIS :

L'INRIA est doté d'un service de diffusion scientifique et technique (UCIS) situé à Rocquencourt. Il édite les publications de l'INRIA : rapports de recherche, thèses, Bulletin de liaison de la recherche en informatique et en automatique, INRIA-infos, plaquettes...

L'Unité de Communication et d'Information Scientifique (UCIS) est subdivisée en plusieurs services : l'audiovisuel, la diffusion et l'imprimerie. La diffusion de l'information électronique est à la charge du service de la communication électronique. L'UCIS exploite la base de donnée O2 dans laquelle elle recueille l'information sous forme numérique.

Section 4 : Les moyens informatiques

Le responsable des moyens informatiques de Rocquencourt assure, en collaboration avec les ingénieurs systèmes présents dans les projets, l'achat et l'installation du matériel informatique et des logiciels.
Par ailleurs, le service des prestations informatiques et télématiques participe à la maintenance du parc matériel et effectue la sauvegarde régulière sur disques optiques numériques des informations stockées sur le site.
L'importance de ces tâches peut être mesurée par l'ampleur du matériel actuellement en service :
⇨ 400 stations de travail personnelles sous UNIX (Sun, Bull, Matra, Dec, Silicon Graphics, Pixar ...),
⇨ 200 micro-ordinateurs, principalement utilisés pour le traitement de texte.

A ces postes de travail, s'ajoutent des machines de service (des PYRAMID multiprocesseurs MIS-4 et 9815, un APOLLO et des serveurs SUN sous architecture SPARC) utilisées comme serveurs de données, routeurs de courrier électronique, pour l'impression, l'archivage, la sauvegarde, etc.
L'INRIA Rocquencourt s'est doté en 1992 d'une machine parallèle à mémoire partagée (KSR1), comprenant un anneau de 32 processeurs, qui est utilisée comme machine de calcul et comme machine d'expérimentation d'algorithmes parallèles. Les chercheurs disposent également d'une machine SEQUENT à 10 processeurs et d'un Encore Multimax à 14 processeurs utilisés pour des expérimentations en calcul parallèle. Ils peuvent aussi accéder au CRAY2 du centre de calcul vectoriel pour la recherche à Palaiseau (CCVR).

Chapitre 2 : Le centre de documentation

Le centre de documentation de Rocquencourt a une vocation nationale. Il rassemble des documents variés sur les thèmes d'intérêt de l'institut. Ce centre est ouvert aux personnes extérieures, il accueille près de 4000 visiteurs par an, étudiants, enseignants ou chercheurs des laboratoires de la région.

Section 1 : Les missions du centre de documentation

Il est possible de consulter, emprunter ou photocopier des ouvrages, rapports, thèses ou périodiques. Le catalogue des ouvrages, ainsi que celui des périodiques du centre de documentation sont informatisés, et donc accessibles par les chercheurs depuis leur station de travail, mais également consultable de l'extérieur sur Minitel (3616 INRIA).

Les services offerts par le centre de documentation sont les suivants :

⇨ Consultation des livres et des périodiques
⇨ Prêt de livres
⇨ Recherche bibliographique
⇨ Photocopie
⇨ Abonnement à l'Inriathèque (relevé hebdomadaire des dernières acquisitions)

Une plaquette présentant les activités du centre de documentation est disponible sur le serveur WWW. Celle-ci est soigneusement mise à jour à l'attention des utilisateurs d'Internet et permet d'interroger les catalogues via WAIS.

Section 2 : Le fonds documentaire

Le fonds documentaire recouvre les domaines de l'informatique, automatique et mathématiques appliquées. Ce fonds est particulièrement orienté vers la recherche. La composition des fonds documentaires est présenté dans le tableau suivant :

⇨ Monographies	12090 titres
⇨ Conférences	8230 titres[1]
⇨ Thèses	9920 titres[2]
⇨ Rapports de recherche	40000 titres
⇨ Périodiques	575 titres

⇨ Collections de normes AFNOR, ISO, CCITT...

[1] dont 650 parues dans des périodiques
[2] dont 2360 sous formes de microfiches ANRT

La consultation des catalogues en ligne est possible à partir de TEXTO, WAIS, NETSCAPE ou MOSAIC.

⇨ Au centre de documentation, 6 terminaux sont à la disposition des utilisateurs pour interroger les catalogues. L'accès à ces catalogues a été facilité par l'utilisation de WAIS (en interne et à l'extérieur).

⇨ Deux P.C. sont également à la disposition des utilisateurs pour la consultation sur CD-ROM des 2 principales bibliographies (INSPEC et MATHSCI, Cf. CD-ROM)

Section 3 : Environnement de travail

A - UNIX

L'INRIA a opté pour un matériel UNIX, avec en majorité des SUN 4x fonctionnant sous le SYSTEM 5 d'UNIX. Le monde P.C. est quasiment occulté exception faite pour l'interrogation des CD-ROM au centre de documentation.

Sous le nom d'UNIX est regroupé le système d'exploitation lui même, ainsi que les commandes qui lui sont associées. Cet ensemble de commandes est intitulé shell. Il existe sous UNIX plusieurs types de shell, celui qui est utilisé à l'ENSSIB est le shell Csh plus connu sous le nom de Bourn shell. Ce dernier est reconnaissable à son prompt qui est le caractère $. L'INRIA utilise le shell Tcsh, souvent appelé shell C dont le prompt est le caractère %.
Tcsh est une couche de logiciel situées sur le bourn shell. Quoi que l'on utilise, les commandes sont interprétées pour être traduites et transmises au bourn shell qui les exécutera.

Le système gère l'ensemble des ressources à l'aide d'un système de fichiers hiérarchisés, d'un gestionnaire de processus et de fonctions spécialisées. Les commandes disponibles permettent la manipulation des fichiers, la gestion des données, l'édition de textes, l'assemblage, la compilation et le traitement de texte. Un interprétateur de commandes puissant permet à un utilisateur (ou à un groupe d'utilisateurs) de se fabriquer ses propres commandes et de se définir ainsi, un environnement correspondant à ses besoins. Cette démarche a été menée au centre de documentation : l'administrateur système a ainsi créé des scripts contenant une succession de commandes rendant l'utilisation plus conviviale. Citons l'exemple des commandes spécifiques d'impression (lpr) redirigeant les opérations d'impression vers une même imprimante installée en réseau, ou encore les commandes abrégées "g" et "m" désignant respectivement la commande de recherche textuelle "grep" et celle d'affichage page à page "more".

Pour pallier le manque de convivialité souvent reproché au système UNIX, on utilise le logiciel X-Window permettant de bénéficier sur les terminaux, d'une interface graphique (Cf. le point suivant 2 - X-Window).

Utilitaires sous UNIX :

Les utilitaires suivants ont très souvent été utilisés :

⇨ **xv** : logiciel de traitement de l'image (format texte, image...)
⇨ **gh** : traitement des fichiers au format postscript
⇨ **Emacs** : éditeur de texte intégrant de nombreuses fonctionnalités présentées dans des menus déroulants.

⇨ **Xmh** : logiciel de messagerie interfacé graphiquement. Il présente de nombreuses fonctionnalités avec notamment le mode réponse, la redirection, la sauvegarde régulière et la mise à jour des messages. Son utilisation est conviviale.
⇨ **Crsh** : outil d'aide à l'écriture de shell script.

B - X Window

Ce produit du domaine public écrit par le MIT donne la possibilité à un programme tournant sur une machine distante d'effectuer des affichages graphiques (mode bitmap) sur un écran distant et de recevoir des interactions venant du clavier et de la souris attachés au poste de travail.

Un serveur tourne sur le poste de travail et gère l'écran, le clavier et la souris. Il communique avec des programmes locaux ou distants : les clients X-Window. Ces derniers peuvent faire des affichages graphiques dans des fenêtres ouvertes sur les écrans distants et recevoir des interactions venant du clavier et de la souris attachés au poste de travail. La liaison entre le client et le serveur s'établit en TCP. Un gestionnaire de fenêtres s'exécutant sur le poste de travail ou sur une machine distante va permettre de déplacer, agrandir, occulter, créer et détruire des fenêtres. X-Window a été développé sur la plupart des systèmes d'exploitation. Des serveurs sont notamment disponibles sur des matériels dédiés : des terminaux X-Window, mais aussi sur P.C., Macintosh et machines UNIX. Il existe une multitude de client X-Window dans le domaine public.

Après avoir taper la commande "xinit" pour lancer le Windows Manager (X-Windows), l'environnement de travail devient immédiatement plus conviviale; il est notamment possible de rappeler les commandes précédemment tapées, ouvrir plusieurs fenêtres qui se superposent, appeler des menus permettant l'accès aux différents logiciels ou applications (messagerie, scanner, traitement de texte, WAIS, GOPHER, NETSCAPE, etc. Le travail est sans aucun doute rendu plus agréable grâce à une telle interface.

C - TEXTO

Le logiciel documentaire utilisé jusqu'à présent par les centres de documentation de l'INRIA est TEXTO. Ce logiciel est surtout marqué par son insuffisance au niveau de la convivialité, aucun interfacage graphique n'a lieu. L'impossibilité de travail simultané en écriture sur un même fichier est également un des défauts notoires. Les centres de documentation INRIA ont récemment acquis les logiciels DORIS-LORIS.

D - Les serveurs de données

Pour assurer ses nombreux services et fonctions, le centre de documentation utilise plusieurs serveurs de données :

⇨ **Chausey** est le serveur exclusivement réservé à l'interrogation des bases de données à partir du centre de documentation. Le public dispose de nombreux terminaux pour consulter et interroger le catalogue. Ces connexions et interrogations simultanées justifient à elles seules l'utilisation d'un serveur. Ce serveur aura notamment à sa charge, la maintenance des bases de données qui constituent l'Inriathèque, l'indexation et la diffusion de celles-ci.

⇨ Le serveur **Merengue** sert prioritairement à M. AUBRIE, l'informaticien du centre de documentation, pour tous les travaux d'administration du système, d'indexation et de mise à jour des bases de données.

⇨ **Nuri** est le principal serveur de l'INRIA. Il gère notamment tous les fichiers TEXTO contenant la base de données du centre de documentation, l'ensemble du personnel peut y accéder.

⇨ Le centre de documentation vient récemment d'acquérir un nouveau serveur **Ishi** qui sera utilisé par le nouveau logiciel documentaire Doris-Loris et qui devra à terme remplacer Nuri.

E - Les réseaux

L'INRIA joue un rôle de plus en plus important en matière de réseau dans la mesure où cet institut est fortement impliqué dans l'organisation que constitue Internet et dans le développement du WEB. Ainsi, les discussions portent sur la définition de normes et l'échange de toute information utile à la communication. L'institut est actuellement l'organisme responsable de l'attribution des adresses réseaux au niveau national.
Le serveur WWW de l'INRIA est consultable à l'adresse suivante : http://www.inria.fr/, il permet de collecter de nombreuses informations utiles et présente l'activité générale de l'institut.

Section 4 : Le personnel et ses fonctions

Direction :

⇨ Paule TOUZEAU	- Gestion du personnel - Gestion administrative, budget,... . - Commande des ouvrages - Commande des périodiques

Secrétariat :

⇨ Jacqueline Fortin

Les documentalistes :

⇨ Marie-Francoise Prade	- Catalogage des monographies - Vérification de tous les fichiers TEXTO - Charge des documents demandés à l'extérieur (articles, PEB...)
⇨ Szwarcbaum Nicole	- Rapports de recherche - Contrôle du catalogage
⇨ Lydie Bonnec	- Commandes des périodiques - Catalogage des périodiques - Catalogage des monographies
⇨ Amirshahy Michaneh	- Gestion des emprunts (relations avec les utilisateurs) - Prêt Entre Bibliothèque

⇨ Nicole Denizou
- Maquette Inriathèque
- Cardex, bulletinage non automatisé

⇨ Alain Marchand
- Thèses françaises (contrats et accords avec des laboratoires)
- Catalogage des thèses
- Collecte
- Microfiche de thèses (ANRT : centre à Grenoble - 1000/an)
- Responsable des 2 P.C. + CD-ROM bibliographiques

⇨ Brigitte Briot
- Commande des ouvrages
- Réception
- Catalogage des ouvrages
- Bases de données des manifestations à venir (conférences)

Informaticien :

⇨ Claude Aubrie
- Administration du système, maintenance informatique

Techniciens et Aide documentaliste :

⇨ Nadia Mesrar
- Saisie des rapports
- Couvertures des livres
- Cote

⇨ Micheline Marien et Alfred Zamy
- Rangement dans la bibliothèque
- Photocopie d'articles (service reprographique)

⇨ Anocq Christine et Barbé Michèle
- Gestion des comptes de photocopies
- Expédition et compte utilisateur

Contrat d'Emploi Solidarité :

⇨ Jeanne Labordes
- Saisie des notices de thèses sous forme microfiche
- Rangement en bibliothèque

⇨ Simone Butler
- Saisie des normes
- Rangement en bibliothèque

L'ensemble du personnel documentaliste participe aux permanences du centre de documentation. Celles-ci consistent essentiellement dans l'accueil du public, la gestion du prêt d'ouvrages et l'aide à la recherche bibliographique sur le fonds documentaire.

Chapitre 3 : Participation à la vie du centre de documentation

Tous les membres du personnel ont été rencontrés individuellement au cours de rendez-vous leur permettant d'exposer les rôles et fonctions qu'ils assument au sein du service. Le fonctionnement général du centre de documentation était dès lors, mieux appréhendé.

Un bureau, une ligne téléphonique, une station informatique ont été mis à ma disposition, cela m'a permis de bénéficier d'excellentes conditions de travail. Sur ma demande, l'administrateur système m'a attribué un accès (login et mot de passe) sur plusieurs machines. Il m'a alors été possible et fort utile de travailler à plusieurs endroits.

Section 1 : Domaines d'application du stage

Ce stage a essentiellement requis des compétences dans le domaine des réseaux (protocoles et normes de communication), de la programmation (shell-script et PERL), des langages de description des documents (HTML), des système d'exploitation (UNIX), du document électronique (Scanner, OCR), de la documentation (interrogation de base de données).

Section 2 : Recherche et Interrogation de bases de données

A - Origine de la demande

1 - Chercheurs de l'INRIA

Ceux-ci sont prioritaires par rapport à tous les autres usagers. Il faut exploiter tous les moyens mis à notre disposition pour satisfaire leur demande. Ils disposent de la quasi-totalité du fonds en libre accès. Leur requête se traduit le plus souvent par la recherche d'un document distant et la photocopie d'un article. Les chercheurs ont à leur disposition trois photocopieurs en libre service.

Voici les résultats obtenus en 1994 pour les chercheurs de l'INRIA de Rocquencourt.

⇨ 4 620 prêts d'ouvrages soit 42% du volume total des prêts.
⇨ 145 ouvrages empruntés auprès de bibliothèques extérieures.
⇨ 25 ouvrages empruntés auprès des autres Unités INRIA.
⇨ 2000 articles de périodiques ou de conférences extraits de notre fonds.
⇨ 800 articles auprès de bibliothèques extérieures et 75 articles auprès des bibliothèques des autres Unités INRIA.

2 - Usagers extérieurs

Au cours de l'année 1994, la bibliothèque a reçu 4280 visites de lecteurs extérieurs; et a prêté 6280 ouvrages à l'extérieur, dont 620 en prêt entre bibliothèques (PEB).

1545 cartes de photocopies ont été vendues, celles-ci permettent aux lecteurs d'utiliser les trois photocopieurs.

Les demandeurs d'articles proviennent principalement des universités et établissements publics de recherche, ils représentent 66% du nombre des demandeurs et 73% du volume de demandes. Les

industriels sont plus nombreux en Ile-de-France où ils représentent 55% des demandeurs et 50% des demandes.
Le centre de documentation est un des centres de recours de l'INIST : 223 articles ont été fournis à ce titre en 1994.

La recette correspondant à la fourniture de cartes de photocopies et à la fourniture d'articles s'élève à 395 500 F H.T. (1510 factures ont établies). La recette correspondant à la vente de l'Inriathèque s'élève à 118000 F H.T.

Les usagers extérieurs ne bénéficient que du fonds du centre de documentation de Rocquencourt. Leur demande peut se traduire par une recherche de document, mais aucun rapatriement ou photocopie ne sera effectué à leur intention.

B - Interrogation de bases de données

Recherche préalable

La phase de recherche comprend les étapes suivantes :

⇨ Les usagers doivent soigneusement remplir un bordereau présentant les références exactes de l'article ou plus généralement du document recherché.
Dans le cas où la référence reste incomplète, l'interrogation des bibliographies sur CD-ROM peut être d'une grande utilité pour la reconstituer (Cf. C - CD-ROM). Pour l'obtention d'une référence de monographie, il est intéressant de consulter la Bibliothèque du Congrès aux Etats-Unis avec laquelle il est quasiment certain d'arriver à ses fins.

⇨ Une vérification de la présence du document dans notre fonds est effectuée. Il arrive en effet, que les chercheurs commettent des erreurs lors de leur recherche préalable.

⇨ Ensuite, la recherche a lieu chez nos confrères de l'INRIA, une interface spécifique nous permet de faire une demande par messagerie. Si le document s'avère être en leur possession, une demande de transfert ou de photocopie a lieu.

Quand le document n'existe pas à l'INRIA :

⇨ quand il s'agit d'une revue, comme c'est fréquemment le cas, on peut interroger le CCN (Catalogue Collectif National des publications en série) par Minitel 3617. Ce service correspond au répertoire de périodiques des bibliothèques françaises. Le centre de documentation de Rocquencourt comme les autres unités INRIA participe au CCN.
Une fois le document trouvé, une demande classique de Prêt Entre Bibliothèques (PEB) est effectuée.

⇨ Le centre de documentation interroge les catalogues des grandes bibliothèques scientifiques françaises ou étrangères pour identifier et localiser un document qu'il ne possède pas. Les interrogations ont notamment lieu à l'INIST, université de Paris 6 et 7, Polytechnique, Bibliothèque du Congrès, la British Lybrary... .

1 - Les fonds de l'INRIA sous WAIS

Les interrogations sont faites sur des index WAIS basés sur les sites distants de l'INRIA. Les fonds des unités INRIA (Rocquencourt, Rennes, Nancy, Grenoble, Sophia-Antipolis) sont par défaut compris dans l'interface de recherche réservée aux documentalistes. Ainsi, il est possible pour une recherche de sélectionner un ou plusieurs fonds.
Rappelons que cette possibilité n'est réservée qu'aux chercheurs de Rocquencourt. Les usagers extérieurs ne bénéficient que du fonds local.

2 - Les fonds externes sous GOPHER, NETSCAPE et WEB

Les bases de données en ligne les plus souvent interrogées sont :

Sur le serveur ESA :

⇨ La base du NTIS contient notamment les rapports ou thèses américaines réalisés grâce à des contrats signés avec l'état.
⇨ La base INSPEC se caractérise par un fonds antérieur (1971) à notre CD-ROM (1989) et une mise à jour bimensuelle, alors que la notre n'a lieu que tous les trois mois.

Sur le serveur DIALOG :

⇨ La base Dissertation Abstract On Line qui recense toutes les thèses soutenues aux Etats-Unis, ainsi que quelques autres pays anglophones.
⇨ La base Books In Print qui permet de savoir si un document est disponible. Elle permet aussi de connaître des renseignements sur l'édition ou le prix du document; cette base est l'équivalant de ELECTRE au niveau international.
⇨ La base ULRICH offre les précédents renseignements correspondant aux périodiques.

C - Les CD-ROM

Les CD-ROM disponibles à partir de stations P.C. au centre de documentation sont INSPEC et MATHSCI. Ceux-ci sont installés dans deux tours de lecteurs et l'interrogation est monoposte. Une imprimante est mise à disposition pour les résultats de recherches bibliographiques. Les chercheurs de Rocquencourt peuvent rediriger ces données sur leur station de travail.
Nous avons très souvent recours à l'interrogation des CD-ROM dans la mesure où elle permet entre autre de compléter ou retrouver la notice descriptive d'un document. Rappelons qu'une fois après avoir obtenu des résultats, il faut à nouveau interroger notre fonds pour savoir si nous possédons le document en question.

1 - INSPEC

Ce CD-ROM et la base bibliographique qu'il contient, sont respectivement produits par l'UMI, et l'IEE, Institution of Electrical Engineers. La mise à jour est trimestrielle et recouvre les domaines de l'informatique, de l'électronique, et du "control" (automatique) avec un intérêt pour la littérature anglo-saxonne. INSPEC a une couverture internationale. Les conférences, ouvrages collectifs et périodiques sont particulièrement bien recensés, alors que nous constatons un manque vis-à-vis des monographies. Le mode de recherche peut être de type booléen. La recherche est effectuée sur une liste de champs.

2 - MATHSCI

Ce CD-ROM et sa base bibliographique, sont respectivement produits par la société Silver Plater, et l'AMS, American Mathematical Society. La mise à jour est semestrielle et recouvre le domaine très large des mathématiques.

Section 3 : Accueil des utilisateurs

Le centre de documentation de Rocquencourt compte près de 2000 lecteurs inscrits et accueille plus de 4000 visiteurs par an. Chaque jour, deux documentalistes sont chargés de les accueillir et les renseigner.

Inscription :

↳ Il n'est pas indispensable d'être étudiants ou d'appartenir à l'INRIA pour s'inscrire au centre de documentation, ce dernier est ouvert au public. L'inscription est gratuite, un manuel d'utilisation est fourni lors de celle-ci.
↳ Les nouveaux utilisateurs doivent remplir une fiche de renseignements.

Opération de photocopie

↳ Lorsqu'il s'agit d'une personne appartenant à l'INRIA, les photocopies sont gratuites. En échange de leur carte INRIA, nous leur attribuons un compteur de photocopies récupéré en fin de travail.
↳ Les personnes extérieures à l'INRIA peuvent acheter des cartes de photocopies. Une facture est établie pour toute vente de cartes.

Opération de prêt

Le logiciel de prêt a été développé en interne par M. AUBRIE, l'administrateur système. Il suffit de se connecter sur Nuri pour y avoir accès.
Trois types de prêt sont possibles : pour les documentalistes, les chercheurs et les personnes extérieures à l'INRIA.
↳ les documentalistes ont le droit d'emprunter 6 documents pour une durée de trois semaines.
↳ les chercheurs de Rocquencourt ont le droit d'emprunter 10 documents pour une durée d'un mois.
↳ les personnes extérieures ont le droit d'emprunter 3 documents pour une durée de quinze jours.

Interrogation du fonds de l'INRIA

↳ Les demandes peuvent être effectuées par téléphone, ou directement au centre de documentation. Il faut alors interroger les fonds de l'INRIA. L'interrogation se fait en full text sur les mots de la notice, à partir des termes les plus pertinents.
exemple : "nom auteur" and "nom périodique" and "année publication".
exemple : "nom conférence" and "lieu conférence" and "année".
Système de réservation

↳ Lors de la réservation, une fiche signalétique est automatiquement éditée. La réservation est prioritaire par rapport à l'emprunt ou au réemprunt. Le document réservé est dès son retour mis en attente, c'est à dire qu'il est considéré comme étant emprunté. Automatiquement, un courrier informant de la disponibilité de l'ouvrage est produit et expédié au lecteur. Le document est déposé à l'accueil et remis à l'utilisateur dès sa réclamation.

Consultation de microfiche

↳ Les microfiches de thèses sont classées par ordre alphabétique du nom d'auteur. Elles peuvent être consultées sur des lecteurs réservés à cet effet.

Au cours de ces nombreuses permanences (une par semaine), j'ai pu constaté l'importance que pouvait revêtir le contact auprès du public. Cette expérience a été enrichissante à tous les points de vue.

2ème PARTIE: METHODOLOGIE DU PROJET INRIATHEQUE

Chapitre 1 : Analyse de la demande Inriathèque

Section 1 : L'Inriathèque sous forme papier

L'Inriathèque sous forme papier existe depuis une vingtaine d'années. Elle était à l'origine gratuite et sous une forme moins élaborée mais avait le même principe : présenter les nouvelles acquisitions du centre de documentation.
L'Inriathèque est en partie un bulletin de sommaires, des périodiques et des conférences reçus dans la semaine, classé par ordre alphabétique des titres. On joint à ce document un index permettant aux utilisateurs de repérer les titres qui les intéressent.
L'Inriathèque est également un bulletin bibliographique des ouvrages, thèses et rapports reçus au cours de la semaine. Ce document tel que je l'ai découvert est relativement commode et simple à réaliser mais son utilisation s'accompagne des contraintes traditionnelles liées à la forme papier, c'est ainsi qu'une recherche de documents nécessite de longues manipulations.

Un membre du personnel prépare la maquette qui est ensuite reproduite à 300 exemplaires par l'imprimeur au centre de diffusion.

L'Inriathèque papier est transmise :

⇨ à tous les projets de l'Unité INRIA de Rocquencourt.
⇨ aux centres de documentation des autres Unités INRIA, ainsi qu'aux chercheurs qui en ont fait la demande.
⇨ à 222 abonnés payants extérieurs.

Il n'est pas envisageable pour l'instant d'abandonner l'Inriathèque sous forme papier, dans le mesure où plus de 200 organismes sont abonnés à cette formule et que l'accès à l'Inriathèque électronique n'est encore réservé qu'aux seuls chercheurs du site de Rocquencourt.

Section 2 : Cahier des charges : étude de la problématique

Cette étude correspond à une démarche liée à une communication avec des partenaires internes et externes, donc dépendant de nombreux interlocuteurs. Cette étape consiste à définir les besoins et faire des choix parmi les solutions. Les divers éléments qui ont découlé de cette étude préalable sont les suivants :

A - Opportunité et faisabilité du projet

Le centre de documentation s'est depuis très longtemps engagé dans la voix de l'Inriathèque sous forme électronique, le personnel et ses dirigeants sont conscients d'un tel besoin. L'Inriathèque sous sa forme papier est très largement exploité, un avenir certain est assuré pour sa version électronique.

Les moyens dont dispose le centre de documentation de Rocquencourt sont à la hauteur des exigences que nécessite la mission à accomplir. Citons par exemple le scanner et le logiciel OCR acquis à cet effet et dont l'exploitation n'a pas encore eu lieu.
Le personnel en nombre suffisant peut sans difficulté se mobiliser et s'investir dans une telle démarche. Une personne aura en charge toutes les questions relatives à l'Inriathèque. L'intérêt manifesté par le personnel documentaliste et les questions quotidiennes qui m'ont été posées au long du stage à propos de l'Inriathèque électronique, constituent une preuve de l'adhésion au projet. L'utilité d'un telle méthode de travail n'a été à aucun moment remise en doute.

B - Les objectifs

Les objectifs visés par ce type de projet peuvent, si l'on tente d'en établir une typologie, se répartir en trois grande catégories :

⇨ L'amélioration de la productivité
⇨ L'amélioration de la qualité du service rendu aux utilisateurs
⇨ L'amélioration de la conservation des documents

Il faut réfléchir dans un premier temps sur la définition des améliorations attendues puis sur leur hiérarchisation. Tous ces objectifs vont donc avoir un impact sur l'ensemble des processus de traitement des documents.

Concernant l'Inriathèque électronique, le processus général de fonctionnement visé est schématisé ci-dessous :

C - Les données

Types de documents :

Nous privilégierons dans cette partie deux points de vue, la nature de l'information et les caractéristiques techniques des documents et des supports.

1 - La nature de l'information

Il faut dans un premier temps déterminer les informations qui vont faire l'objet de l'application : seules celles ayant une valeur réelle et avérée doivent être retenues. Les critères de sélection portent essentiellement sur :

- **La nature de l'information** : information textuelle.
- **Les thématiques** : information scientifique (informatique, mathématiques, automatique, physique).
- **le niveau d'élaboration** : information secondaire (références bibliographiques) et information primaire (sommaires de conférences et périodiques).
- **L'origine et la rareté de l'information** : littérature interne, littérature externe et littérature grise
- **Le degré de confidentialité** : faible degré de confidentialité, la diffusion des informations aura lieu en interne et en externe, néanmoins les document se caractérisent par une forte valeur informationnelle.
- **La cible** : en interne, il s'agit de chercheurs experts, étudiants en thèse. En externe, il s'agit d'enseignants, d'industriels, d'étudiants et d'abonnés extérieurs (organismes divers).

2 - Les caractéristiques techniques des documents et des supports

La nature des documents et des supports va déterminer les choix techniques. Plusieurs paramètres doivent être analysés :

- **La forme** : sommaires de conférences et de périodiques, notices bibliographiques de monographie, rapport de recherche, et thèses.
- **Le support** : électronique, analogique
- **Les caractéristiques physiques** : format textuel

Ces typologies permettent de cerner des ensembles homogènes de documents ayant des caractéristiques identiques ou compatibles et pouvant donc faire l'objet du même traitement documentaire et informatique. En début de mission, il n'était pas encore décidé du type de données dont serait constituée l'Inriathèque. Le choix résidait dans le format textuel ou le format image. L'étude a eu lieu sur les deux formats et nous a permis de nous rendre compte que la base de données textuelles s'avérerait la meilleure solution et s'adaptait davantage au contexte du centre de documentation. Cette solution a été dictée par le fait qu'une recherche devait être effectuée sur le contenu des sommaires de conférences et de périodiques.

D - Les besoins

Par rapport aux autres bases de données utilisées au centre de documentation, l'Inriathèque doit offrir :

- ⇨ une sélection de documents,
- ⇨ une mise à jour rapide (à la semaine),
- ⇨ une rapide disponibilité (les documents sont enregistrés dans le fonds dès leur réception),
- ⇨ une interrogation du catalogue Inriathèque (nouveautés et archives),
- ⇨ une demande de photocopie d'article en ligne.

Une des principale source de demande est la photocopie d'articles. La fourniture d'articles est une activité importante, deux personnes au centre de documentation sont affectées à cette tâche. En

1994, sur les 11720 demandes d'articles faites auprès de l'INRIA, près de 2000 ont été assurées par le centre de Rocquencourt. L'Inriathèque pourrait pour son fonds, être un relais au système de demande de photocopies d'articles.

Le contexte de recherche sur lequel repose le centre de documentation, demande de rendre plus rapide le système de diffusion de l'information. En effet, il est nécessaire d'améliorer la productivité de ce service. Les chercheurs doivent disposer d'un outil leur permettant d'acquérir l'information de manière plus vive, d'autant plus que ces informations sont à leur porte, en l'occurrence dans leur centre de documentation. Nous pourrions ainsi, améliorer la qualité du service, les enjeux liés à l'utilisation de l'Inriathèque électronique justifient d'envisager ce type de traitement.

E - Les utilisateurs

Les principaux utilisateurs seront dans un premier temps les chercheurs de Rocquencourt. L'Inriathèque devra selon le responsable du centre de documentation, répondre en priorité à cette demande.
Les utilisateurs externes peuvent être divisés en deux catégories : d'une part les usagers du centre de documentation qui se déplacent lors d'un besoin ponctuel, d'autre part les abonnés à l'Inriathèque.

Chapitre 2 : Démarehe de Gestion Electronique de Document

Définition générale du concept de GED

Les applications de GED, plus que d'autres encore du fait des coûts d'investissement souvent élevés, nécessitent une analyse de l'existant et des besoins pour en déterminer l'opportunité réelle et les bénéfices que l'on peut en attendre (Cf. Section 2 : Cahier des charges : Etude de problématique). la GED implique une conception globale du processus de traitement du document.

Le terme Gestion renvoie à un ensemble de traitements permettant d'acquérir, d'enregistrer, de stocker et de restituer l'information.
La gestion électronique renvoie à un ensemble de traitements effectués avec des moyens électroniques.
Le document est défini par l'ISO comme "l'ensemble d'un support d'information et des données enregistrées sur celui-ci sous une forme en général permanente et lisible par l'homme ou par une machine".
Nous considérerons dans un cadre plus restreint le terme document comme tout écrit, comportant des informations textuelles ou graphiques, en noir et blanc ou en couleur, présenté sur un support papier ou sur un support informatique. La notion de document sera dans notre travail consacrée pour désigner soit un sommaire de périodique, soit un sommaire de conférence. Le terme sommaire sera utilisé pour une table des matières

L'Inriathèque en tant que document électronique ne s'intéressera qu'aux traitements qui permettent de transformer le document de départ, sur support papier ou sur bande magnétique, en un document dont la forme et le support le rendront gérable par des moyens électroniques.

Section 1 : Mode d'acquisition et de stockage des données

L'acquisition et le traitement initial de l'information

Les particularités de la gestion électronique de documents portent pour l'essentiel sur l'acquisition de l'information. Il s'agit en effet de traiter une information brute dans le sens où le plus souvent elle n'a pas été structurée, codée ou codifiée suivant les règles habituelles de l'informatique documentaire. La diversité et la non préparation de l'information à traiter expliquent la complexité de certaines applications.

L'acquisition est la production d'une image électronique de l'objet à traiter. Cette image électronique peut éventuellement faire l'objet de traitements complémentaires divers :
⇨ Soit pour l'amélioration de la présentation. C'est en effet, le cas de notre reformatage en HTML qui permettra d'ajouter aux documents une police de caractère conviviale à la lecture.
⇨ Soit pour transformer en une entité informatique plus facilement gérable. C'est le cas de notre module de reconnaissance optique de caractères qui finalement, permet une recherche sur le contenu du document.
⇨ Soit pour contrôler l'information acquise et associer cette image à d'autres données. C'est le cas cette fois-ci, du rapprochement effectué entre le texte du sommaire et sa notice bibliographique dans le fichier TEXTO correspondant. Cette opération permet ainsi d'identifier clairement chaque document.

Ce traitement initial peut donc dépasser la simple capture d'image et constituer une chaîne de production relativement sophistiquée.

Nous distinguerons les tâches suivantes :

⇨ La capture d'images dont l'objet est l'acquisition d'une image électronique brute de l'objet ou du document à traiter.
⇨ La reconnaissance optique de caractère dont l'objet est la transformation de tout ou partie de l'image électronique brute, en un texte codé suivant les règles habituelles de l'informatique documentaire.
⇨ L'analyse lexicographique, syntaxique et ou sémantique du texte résultant de la tâche précédente, dans la perspective d'un contrôle de l'information ou de l'automatisation d'une partie du travail d'analyse documentaire. Cette opération consistera notamment dans la reconnaissance de certains éléments du texte en vue de traitement.

Suivant la phase de l'application dans laquelle nous nous situons, les résultats obtenus sont très divers :

⇨ Une image électronique, non normalisée, traitable par des moyens informatiques ou par des moyens électroniques. C'est le résultat obtenu lors de la première opération de numérisation.
⇨ Un texte, c'est le résultat acquis par le module OCR.
⇨ Un ensemble de descripteurs et de caractéristiques du document, il s'agit là des notices qui sont grâce à un programme, automatiquement extraites du fichier TEXTO.

A - Acquisition des données

Les données acquises par notre système de GED sont d'une part, les sommaires de conférences et périodiques qui feront probablement l'objet soit d'un abonnement auprès d'un fournisseur, soit de numérisation, soit de récupération sur Internet. D'autre part, les notices bibliographiques de monographies, de rapports de recherche et de thèses qui sont extraites des fichiers TEXTO de la base du centre de documentation.

1 - L'abonnement aux sommaires

1.1 - Etude comparative de plusieurs offres d'abonnement

L'objectif de ce travail est de s'assurer de l'acquisition d'une information fiable et facile à exploiter. Le coût d'un tel investissement est un facteur essentiel à prendre en ligne de compte.

Les problèmes susceptibles d'être rencontrés sont notamment les diverses modalités permettant de récupérer l'information, le risque d'une information incomplète, ou encore des insuffisances au niveau du format des données.

Les sommaires à insérer dans l'Inriathèque doivent nécessairement être accessibles en ligne, et posséder un format à la fois complet mais également cohérent, permettant une consultation compréhensible et une description suffisante du document pour le localiser ultérieurement.

Pour mieux connaître le recouvrement d'une base de données, il existe un moyen qui consiste à interroger les principaux fournisseurs sur les éditeurs de périodiques de l'INRIA, à savoir : Springer (Berlin), Noth-Holland, Elsevier, Pergamon, Academic Press, Wiley, et également les sociétés savantes qui ont également une activité d'édition, à savoir : IEEE, ACM (Association of Computing Machinery), AMS (American Mathematical Society). Notons que les outils qui permettront de produire un taux de recouvrement fiable n'ont pas encore été réalisés. Il est clair que nous n'aurons pas pour l'instant de résultats précis quant au recouvrement, mais nous tenterons d'en avoir une idée plus proche.

Le salon IDT sur l'information électronique du mois de juin 1995, nous a permis d'acquérir des connaissances sur les offres effectuées sur le marché de l'abonnement. Cette visite fut fructueuse, et a contribué à récupérer une abondante documentation.

Cette étude doit mettre en avant la qualité des bases de données sélectionnées auprès des trois fournisseurs suivants :

- ⇨ EUROPERIODIQUES
- ⇨ OCLC
- ⇨ CARL

De manière générale, les questions à se poser pour chaque offre de sommaires sont les suivantes :

- ⇨ Quel est le domaine de couverture ?
- ⇨ Quel est le type de format ?
- ⇨ Quel est le tarif d'abonnement ?
- ⇨ Quelles sont les modalités d'abonnement ?
- ⇨ Quelle est la périodicité de mise à jour?
- ⇨ Quelle est la méthode de récupération des données ?

⇨ Quel est le taux de couverture des offres proposées par rapport au catalogue INRIA ?

1.1.1 - Europeriodiques

Europériodiques a une activité principale de libraire. Cette entreprise est donc susceptible de moins maîtriser le domaine de la documentation que ses concurrents. La demande qui a été faite ne concerne que les périodiques, Europériodiques ne dispose pas de base de données relatives aux conférences.
Europériodiques nous a transmis sur trois disquettes de nombreuses informations relatives au format des données, à leur catalogue de périodiques, au recouvrement de notre fonds, à la tarification, au mode de récupération des données, à la périodicité du transfert de données et enfin à la mise à jour de la base. En supplément, Europériodiques nous a offert une version simplifiée du logiciel documentaire (PROCITESAMPLES) qu'il utilise pour gérer leur fonds.

1 - Les formats obtenus sont de deux types :

⇨ le premier correspond véritablement à un sommaire de périodique, comportant les champs suivants : titre du périodique, date, volume, ISSN, les articles avec leur pagination.
⇨ le second correspond davantage à celui d'une notice résumée d'article, chaque article est référencé par son titre, sa pagination, le titre du périodique, le n°volume et la date.

Les fichiers et formats suivants ont été extraits des disquettes reçues :

Format ASCII :

```
!New!
Information Srategy
ISSN : 0743-8613    1995   Vol : 11        iss : 3

Page : 6-12
Computer-Information literacy for Senior Managment
Kanter, Jerry

Page : 13-17
The CIO's Dilemma : participating in strategics planning
Stephen, Charlotte S

Page : 18-24 ... etc
```

Commentaire :

Aucune mention d'intitulé de champ n'a lieu, et aucun protocole spécifique ne nécessite d'être décrit. L'importation est possible à partir de logiciels tels que WORD ou WRITE sous WINDOWS. Le séparateur délimitant les différentes notices est l'expression !New!.

Format RIS :

```
TY - JOUR
JO - Nerderlandse chemischte industrie
PY - 1995
VL - #3
SP - 4
TI - Wet van ...
ER -
```

Commentaire :

A chaque article est associé une notice comprenant une liste de champs. Il est également décrit comment importer des enregistrements à partir du logiciel Référence Manager. Le séparateur de notice n'apparaît pas pour ce type de format. Une ligne sépare chaque notice.

Format PROCITE (consulté à partir du logiciel PCSAMPLES):

Les données sont présentées sous forme de notice résumant l'article. Chaque article possède donc une notice.

Auth :
Titl :
Jrnl :
Publ :
Date :
IsID :
Page :
Note :
Abst :
Desc :

Commentaire :

La notice est plus complète. Il est notamment possible à partir de ce logiciel et ce type de format, d'exporter ou d'importer des notices, de créer sa base de données en optant pour une structure de champs personnelle incluant tel ou tel descripteurs, mot-clés... etc.

FORMAT TEXTO :

TITRE
Actueel - Bij Hamm is kwaliteit belangrijk
JOURNAL
Bouwmachines
DATE
1995
VOL
30
FASIC
2
PAGE
5-9
ISSN
00068373
//

Commentaire :

La présentation est convenable, une nouvelle fois une notice est associée à chaque article. Pour le format TEXTO, aucun protocole spécifique ne nécessite d'être décrit. L'importation est possible à partir du logiciel SWETSCAN chez Europériodiques. Le caractère qui sépare les notices est //.

Fichier SWETSCAN.JNL :

Ce fichier contient le répertoire des périodiques classés alphabétiquement par nom d'éditeur avec mention d'un code sujet (présentation d'une table de sujet avec Code sujet et nom du sujet). Ce fichier correspond en fait, au catalogue de l'ensemble des périodiques disponibles chez Europériodiques.

exemples de données du fichier SWETSCAN.JNL :

ISSN	TITRE	CODE SUJET
0106-0627	Tidsskrift for Historisk Forskning	960
0340-3386	3R International - Rohre Rohrleitungsbau Rohrleitungstransport	846
0261-6823	AA Files - Architectural Association	843
0065-7158	AACE Transactions - American Association of Cost Engineers	899
0094-6354	AANA Journal - American Association of Nurse Anesthetists	609

exemple de code sujet intéressant l'Inriathèque :

500 Mathematics
835 Computer technology and application
510 Physics

2 - Fonds et Recouvrement :

Le fonds d'EUROPERIODIQUES contient près de 14000 périodiques. Une recherche sur le fonds (fichier SWETSCAN.JNL) a permis d'obtenir plus de 1000 périodiques possédant l'un des trois précédents codes sujet. Une demande leur a été faite pour obtenir davantage d'informations par rapport au recouvrement de notre fonds. Malheureusement, le personnel de cette entreprise n'était pas toujours très disponible, et n'a pu être en mesure de répondre. Un échantillon de données relatif à notre domaine d'application (informatique, automatique...) leur a été demandé, il n'a jamais été possible de l'obtenir. Tenons compte du fait que toutes ces questions leur ont été posées durant une période de vacances. Rappelons également que des utilitaires permettront dans un avenir très proche, d'établir un recouvrement beaucoup plus précis.

3 - Tarification :

L'abonnement à l'intégralité de leur base de sommaire s'élève à 25000 francs H.T.. Cependant, cette offre est modulable, et pourra fluctuer en fonction de leur taux d'activité. Par ailleurs, un abonnement aux 600 périodiques du centre de documentation coûterait le même montant. Dès lors, l'abonnement à la base intégrale s'impose. Cet abonnement pourrait permettre à l'Inriathèque de proposer en supplément, la consultation de périodiques absent du fonds de l'INRIA.

4 - Mode de récupération des données :

Europériodiques met à disposition ses données par disquettes au format P.C.. Ce mode de transfert par disquette, constitue une contrainte et un temps de traitement plus long. En effet, les données récupérées toutes les semaines nécessiteront de nombreuses manipulations : injecter les données par les P.C. grâce à une session FTP. Nous aurions préféré le moyen le plus pratique qui est la transmission électronique des données sur une messagerie par exemple.

5 - Périodicité du transfert de données : hebdomadaire

6 - Mise à jour de la base de données : hebdomadaire

1.1.2 - OCLC Online Computer Library Center, Inc.

Il faut tenir compte du fait qu'OCLC est une entreprise de taille considérable, que le domaine de la documentation semblerait-être mieux appréhendé que le fournisseur précédent. Des relations étroites existent entre cette entreprise et La British Library. OCLC propose un total de 56 bases de données de domaines très divers.
Dès le premier contact téléphonique, nous nous sommes rendus compte que les interlocuteurs montraient énormément d'intérêt à nos questions.
La demande faite auprès d'OCLC concerne cette fois-ci, à la fois les périodiques et les conférences.
Les informations ayant pu être obtenues concernent le fonds et son recouvrement, la tarification, le mode de récupération des données, la mise à jour de la base, et enfin le format des données.

1 - Fonds et Recouvrement :

OCLC propose deux bases relatives aux périodiques, celles-ci s'intitulent ContentsFirst et ArticleFirst. La première contient l'information qui nous intéresse le plus, c'est à dire les sommaires de périodiques. Celle-ci contient 12500 périodiques. La seconde nous attire moins dans la mesure où elle contient le texte intégral des articles.
Cette offre est similaire pour les conférences : une première base ProceedingsFirst propose les sommaires, elle contient près de 19000 références signalées depuis 1993. La base des articles est la même que celle des périodiques : ArticleFisrt.
Les conférences traitées sont les suivantes : Worldwide conference, Professional meeting, Symposia.
Notons que les conférences nécessitent un traitement légèrement plus délicat que celui des périodiques. En effet, il arrive très fréquemment qu'une conférence change de nom, que la périodicité d'apparition varie d'une conférence à l'autre, pour résumer qu'elle ne soit pas aussi régulière qu'un périodique.

Au cours d'une longue conversation téléphonique, notre interlocuteur, conscient de l'intérêt que l'on manifestait à l'égard de leurs bases de données, nous a attribué un accès gratuit pour une durée de 4 semaines, à l'ensemble des bases. Un login et un mot de passe nous ont été attribués à cette occasion. Ce résultat constituait une bonne nouvelle, dans la mesure où nous n'avions encore jamais expérimenté un véritable produit commercial, des données très concrètes pouvaient alors être consultées.

2 - La tarification :

L'abonnement correspond dans un premier temps à un forfait de base équivalant à une somme de 26650 francs H.T., à laquelle s'ajoute un montant d'abonnement de 10200 francs pour chacune des 56 bases disponibles. Dans notre cas, l'abonnement à la base des sommaires de périodiques (ContentsFirst) coûterait 36850 francs (26650 + 10200). Si l'on ajoute la base des sommaires de conférences (Proceedings First), nous obtenons la somme de 47050 francs (26650 + 10200 + 10200).

Notons que cette tarification est plus onéreuse que celle d'Europériodiques, mais qu'elle permet une gestion des conférences identique à celle des périodiques. La numérisation ne concernerait plus que les données non reçues par abonnement.

3 - Mode de récupération des données :

La connexion aurait lieu avec un login et un mot de passe, le téléchargement se ferait en ligne

4 - Mise à jour :

La mise à jour de la base des périodiques est journalière, tandis que celle des conférences est mensuelle.

5 - Format des données :

Next record

JOURNAL : IEEE network.
VOL, ISSUE : Volume 9, Number 2
YEAR : 1995
ISSN : 0890-8044
BL SHELFMARK : 4363.007500
FREQUENCY : Bimonthly,
PLACE : [New York, N.Y. :
PUBLISHER : Institute of Electrical and Electronics Engineers,
J ALT NAME : IEEE Network
SPON. AGENCY : IEEE Communications society. Institute of Electrical and Electronics Engineers.
SUBJECT HDGS : Computer network--Periodicals.
ARTICLE : Establishing a Real-Time Video Link to Antarctica pp. 8
AUTHOR (S) : Leon, M.
ARTICLE : The Use of Artificial Neural Networks for Optimal Message Routing pp. 16
AUTHOR : Wang, C.-J.; Weissler, P.N.
ARTICLE : The Rate-Based Flow Control Framework for the Available Bit Rtae ATM Service pp. 25
AUTHOR : Bonomi, F.; Fendick, K.W.
ARTICLE : Credit-Based Flow Control for ATM Networks pp. 40
AUTHOR : Kung, H.T.; Morris, R.

Next Record

Commentaire :

Il n'existe qu'un seul format à OCLC. De type ASCII, il est clair et complet, les intitulés de champ sont correctement mentionnés. Le séparateur de notice est ici, l'expression Next record. Notons en ce qui concerne la pagination que seule la première page d'article est inscrite, les utilisateurs devront effectuer une soustraction pour obtenir le nombre de pages par article.

1.1.3 - CARL

La société CARL implantée aux Etats-Unis offre plusieurs bases de données en libre accès dont Uncover qui traite des sommaires de périodiques. Il faudrait dans l'hypothèse de l'exploitation de cette base tenir compte des temps d'accès très longs. Les informations ayant retenues notre attention correspondent au fonds documentaire, au format des données et enfin à l'abonnement.

La connexion se fait grâce à une session TELNET. La gratuité d'Uncover laisse supposer que la prestation fournie ne soit pas nécessairement la plus fiable.

1 - Fonds :

La base Uncover contient 17000 périodiques et près de 6 millions d'articles. Après interrogation et consultation de quelques notices dont nous disposions du sommaire à l'INRIA, la qualité de la base CARL a été remise en question. En effet, nous avons décelé quelques erreurs et insuffisances qui nous laisse penser que ces informations sont saisies manuellement augmentant de cette façon le risque de données erronées.

De plus, en consultant le catalogue de la base CARL, il a été découvert de nombreuses erreurs dans la description des périodiques. En effet, de nombreux périodiques ne possédaient pas de numéro d'enregistrement, d'autres n'avaient pas de numéro d'ISSN.

2 - Format :

La recherche peut notamment avoir lieu sur l'auteur ou le titre de périodique. L'interrogation sur un titre permet d'obtenir l'ensemble des numéros et volumes disponibles. Le format du sommaire se présente de la manière suivante :

N° article	AUTHOR	TITLE	PAGE
001	nom auteur	titre numéro 1	numéro de page
002	nom auteur	titre numéro 2	numéro de page

La présentation sous forme de tableau est originale, mais il semblerait que les seuls champs présents soient au nombre de trois (auteur, titre de périodique, pagination). Les informations relatives aux périodiques sont présentées dans la fenêtre précédente, mais n'apparaissent pas dans celle des sommaires.

3 - Abonnement :

En interrogeant la base CARL, nous avons relevé une notion de profil (Profile). En effet, le système nous demande avant et après la connexion, si l'on désire obtenir un profil. Le système du profil permet l'utilisation d'un mot de passe permettant d'être identifié. Cette notion de profil pourrait correspondre à un traitement particulier nous assimilant au statut d'abonné et nous permettant de bénéficier d'une meilleure offre de service. Dès lors, il sera possible de recevoir régulièrement des données. Aucun représentant de la société CARL n'a été trouvé en France, c'est la raison pour laquelle très peu d'informations ont pu être obtenues à propos de leur offre.

1.2 - Tableau récapitulatif

Critères	Europeriodiques	OCLC	Uncover
Contenu	14000 périodiques	12500 périodiques	17500 périodiques
Domaines de couverture	incluant : ⇨ Mathématiques ⇨ Informatique ⇨ Physique	incluant : ⇨ Mathématiques ⇨ Informatique ⇨ Physique	incluant : ⇨ Mathématiques ⇨ Informatique ⇨ Physique
Format	⇨ Notice de sommaire ⇨ Notice d'article ⇨ plusieurs formats disponibles	⇨ Notice de sommaire ⇨ un seul format disponible	⇨ Notice de sommaire ⇨ un seul format disponible
Mode de collecte	Numérisation	Numérisation	Saisie
Transfert de données	⇨ par disquette	⇨ téléchargement en ligne	⇨ téléchargement en ligne
Séparateur de notice	En fonction de notre demande	Next record	non déterminé
Tarif d'abonnement	⇨ Pour la base intégrale : 25000 francs H.T. ⇨ Pour les 600 périodiques de l'INRIA : 25000 francs H.T.	⇨ Pour la base intégrale des périodiques : 36850 francs H.T. ⇨ plus la base intégrale des conférences : 47050 francs H.T.	⇨ Pour la base intégrale : gratuit ⇨ Avec une notion de Profil (abonnement) : non déterminé.
Périodicité de la mise à jour	hebdomadaire	⇨ journalière pour la base périodiques ⇨ mensuelle pour la base conférences	hebdomadaire

Conclusion :

Cette étude apporte des éléments qui permettront à l'avenir de prendre une décision. En effet, le choix d'un fournisseur n'a pas été arrêté à ce jour, et ce notamment, dans la mesure où le budget d'une telle acquisition n'a pas encore été débloqué. Cependant, nous pouvons d'ores et déjà dire qu'en terme de qualité et de complétude de l'offre (périodiques et conférences), OCLC est le mieux placé. D'un point de vue financier, Europériodique propose une offre plus avantageuse en terme de coût et de couverture de périodiques (14000). Uncover semble être un produit complet mais destiné davantage à l'utilisateur final plutôt qu'à une bibliothèque. De plus, nous n'avons pas trouvé de représentant en France de ce produit.
Notons que la société Dawson propose un service d'abonnement aux sommaires de conférences, disponible dans les mois à venir, il serait intéressant d'en étudier les modalités.

Après avoir étudié les propositions d'abonnements, nous allons maintenant traiter de notre seconde méthode d'acquisition des données, il s'agit de la numérisation.

2 - La numérisation des sommaires

Pour numériser l'information, la technique utilisée est l'échantillonnage : au lieu d'enregistrer le signal continu, on enregistre la valeur de ce signal en un nombre limité de points de l'image, en anglais "picture element" ou pixel de l'image. On divise donc l'image selon un quadrillage à l'intérieur

duquel sont prélevés les pixels. La fidélité à l'image d'origine est d'autant plus grande que le quadrillage est serré; le nombre de pixels par unité de longueur constitue la résolution de l'image (elle s'exprime en point/mm ou en dpi : dots per inch).
Les premiers essais de numérisation ont été effectués sur des sommaires de conférences de l'IEEE. Nous pouvons considérer la numérisation comme une méthode permettant d'acquérir les sommaires n'ayant pas pu être obtenus d'une autre manière.

2.1 - Matériel disponible

2.1.1 - Le scanner et l'OCR

Il y a maintenant deux ans que le centre de documentation s'est doté d'un matériel performant pour pouvoir numériser les documents papiers.

Matériel :

➢ Scanner - ScanPartner 10

Caractéristiques :

⇨ lecture de document dans un format A4 ou LTR
⇨ chargement manuel ou automatique des pages
⇨ haute vitesse de lecture
⇨ haute résolution de lecture (max. 300 dpi)
⇨ taille compacte

➢ Logiciel OCR - ScanWorX

Caractéristiques :

Configuration minimum pour un tel type de logiciel :
⇨ architecture SUN 4C ou SUN 4M
⇨ 16 Mb octets de RAM (32 sont recommandées)
⇨ 25 Mb octets d'espace d'échange (swap space) (50 sont recommandées)
⇨ 28 Mb octets d'espace disque pour charger ScanWorX
⇨ SunOS version 4.1 ou antérieure
⇨ un lecteur QIC-24 ou QIC-150 pour installer ScanWorX sur la station de travail

Les fichiers générés par cet OCR sont compatibles avec une grande variété de logiciels de traitement de texte, de traitement de l'image, d'application de bases de données. Les principales fonctions et outils fournis par cet OCR sont explicités dans les points suivants.

2.1.2 - Analyse des fonctionnalités : configuration et paramètrage

Le logiciel d'OCR ScanWorX est doté de fonctionnalités de reconnaissance très puissantes et permet d'obtenir de très bons résultats. Il produit également des images de qualité satisfaisante pour nos besoins. Toutefois, ces résultats ne peuvent être obtenus qu'à condition de bien utiliser les outils mis à disposition par l'OCR, entre autre accorder suffisamment de temps aux opérations d'apprentissage (*verifier*) ou paramètrer minutieusement les diverses configurations.

Le mode d'apprentissage ou vérification :

Lorsqu'il s'agit d'un nouveau document faisant apparaître des polices de caractère qui n'ont pas encore été traitées (grosseur de caractère, épaisseur du trait, graisse, italique ...), il faut lancer une session d'apprentissage de l'OCR en cliquant sur l'option *verify* avant la commande *Go*.(Cf. 2.3.1 - Notice technique d'utilisation)
Ensuite, il suffit dans la fenêtre qui apparaît, de corriger manuellement les probables erreurs dues à l'OCR, grâce aux nombreux outils de vérification. Une fenêtre est notamment disponible pour entrer des caractères spéciaux tels que ceux de l'accentuation (*ScanWorX motif*). Enfin, grâce à la commande "*save verifier set as*" dans le menu document, il faut sauvegarder la session de vérification sous le même nom le fichier de vérification : *verifinria.vfr*.
Il est certain que plus l'OCR est utilisé en session de vérification (*option verify*), plus son travail par la suite sera performant.

Après avoir utilisé l'OCR en session d'apprentissage, nous obtenons un fichier de vérification avec une extension *vfr*. Il a été décidé de ne créer qu'un seul fichier de vérification en vue de simplifier l'utilisation. Ce fichier s'intitule *verifinria.vfr*, il a été conçu à partir d'un large échantillon de documents (ACM, IEEE, Springer...), et tient compte de l'hétérogénéité des divers formats ou polices de caractères.
Notons par exemple que les conférences de l'éditeur Springer sont constituées d'un texte très serré. Le mode vérification de l'OCR montrent que souvent les lettres sont juxtaposées, les mots sont fréquemment coupés et suivis d'un tiret en fin de ligne, et enfin l'utilisation de l'italique est très abondante. Toutes ces spécificités rendent ce type de document plus délicat à traiter, d'où l'intérêt de l'utilisation des sessions d'apprentissage qui enregistreront toutes ces modalités pour les appliquer ultérieurement sur un autre document. Rappelons que les sommaires de conférences contiennent de nombreux noms propres avec une accentuation spécifique au pays d'origine, ceci constitue une difficulté supplémentaire pour le travail de l'OCR.

Le séparateur de page :

Lors d'une opération de reconnaissance de caractère, l'option *use document separator* est extrêmement utile dans la mesure où elle permet d'insérer un signe spécifique, les caractères ^L (ctrl L) séparant ainsi les différentes pages d'un même document.
Notons que cette question de séparateur de documents ne se pose pas en mode image dans la mesure où chaque page constitue un fichier distinct (.tif).

Les photocopie :

Nous pensons qu'il est préférable de faire photocopier les documents à traiter et charger celles-ci sur le scanner. Ce choix s'explique en partie par le fait que la texture et le fond du papier de photocopie sont toujours les mêmes. Nous résolvons de cette façon le problème de luminosité (la valeur 78 est celle qui convient le mieux à la clarté des documents).
L'utilisation des photocopies se justifie également par le fait qu'il était impossible de scanner plusieurs pages successives d'un document relié. En effet, ce logiciel n'était pas étudié pour scanner des livres mais des feuillets séparés. Les pages paires se présentaient dans un sens, les pages impaires étaient à l'envers.

Lors des opérations de photocopies, il faut respecter une consigne importante : le texte doit être cadré. Cela suppose donc que les photocopies soient effectuées avec soin, afin de disposer d'un format uniforme de document rendant le travail de l'OCR plus efficace.

Le gabarit ou modèle :

A partir de la visualisation à l'écran de la page scannée (*preview*), la commande *create template* dans le menu *template* permet de construire un gabarit (*template*), c'est à dire la sélection d'une zone de traitement. Celle-ci peut ensuite être sauvegarder (*save template as*) sous une extension *tpl*. Ce gabarit pourra par la suite être à nouveau chargé pour le même type de document. La régularité et la qualité de la photocopie conditionnent et facilitent l'utilisation du gabarit. Ainsi, si les photocopies sont correctement réalisées, le gabarit pour lequel on a opté reste valable pour toute la session de travail.

Les configurations :

Les modes de numérisation (reconnaissance de caractère, numérisation d'image, ou les deux simultanément) que l'on décrira ci-dessous, devaient être au préalable correctement paramétrés. Les différents paramètres sont les suivant :

-a- customs preset

Il s'agit des paramètres généraux permettant de scanner, de visualiser le document scanné *(preview)*, de reconnaître le texte, de convertir le texte, de zoner automatiquement le document (c'est notamment le cas pour la configuration stiinria.fr) ou de sauvegarder le document. Il a été décidé que le document scanné serait systématiquement visualiser avant le traitement, ceci afin de permettre une sélection des zones spécifiques à traiter.

-b- add pages settings

Il s'agit des paramètres permettant l'utilisation d'un séparateur entre les pages d'un même document, de scanner séparément les deux cotés d'un même document, d'inverser le sens d'une page.

-c- image settings

Ce sont les paramètres propres au traitement d'un document au format image. Plusieurs formats d'image peuvent être choisi (Tiff Uncompressed, Tiff CCITT-3, Tiff CCITT-4, Sun Raster standard, Xerox Interpress ...). Le format que nous utiliserons par défaut est le Tiff CCITT-3, ce dernier est correctement converti en format jpeg et gif par l'intermédiaire de l'utilitaire xv. Il faut également opter pour une résolution de fichier de sortie (par défaut, nous utiliserons la résolution maximum de 1200 dpi), il est enfin possible de faire varier la taille de l'image de 50 à 200%.

-d- text setting

Il s'agit des paramètres concernant les documents traités par reconnaissance de caractères. Ils permettent entre autres d'utiliser un dictionnaire, d'effectuer une vérification du texte scanné, de charger un fichier de vérification préalablement établi, d'opter pour plusieurs types de formats texte (ASCII, asciispaces, ASCII tabs, ASCII intelligent, FrameMaker, Interleaf, WordPerfect 5.x ...). Le format que nous utiliserons par défaut sera ASCII.

-e- scanner settings

Il s'agit enfin des options relatives au paramètrage du scanner. Il est permis de régler le niveau de brillance de 1 à 255 (brightness). Selon la clarté du document, il faudra plus ou moins

augmenter le niveau de brillance. Les valeurs que nous utilisons se situe entre 60 et 80 (78 par défaut), ces choix correspondent au niveau élevé de clarté du papier de photocopie.
Nous devons également choisir le format de page (page size : A4 par défaut), la résolution pour la numérisation du document (valeur recommandée : 300 * 300 dpi). Nous pouvons également utiliser le chargeur automatique de papier (use automatic document feeder).

Les trois modes de numérisation (ou configuration) pour lesquels nous avons optés, sont les suivants :

-1- Configuration Srcinria.set : Scan, Recognize, and Convert

Il s'agit de la configuration la plus complète pour l'obtention de document au format texte. Trois opérations sont successivement réalisées. La numérisation (*Scan*), la reconnaissance (*Recognize*) et la conversion en plusieurs format texte (*Convert*). Le logiciel ScanWorX permet de réaliser séparément ces opérations. Pour obtenir le résultat escompté (document au format texte), les trois opérations sont indispensable, c'est la raison qui nous a poussé à privilégier cette configuration. Dans le cas où un document a déjà été scanné, et dont nous possédons les fichiers TIFF (format image), il est recommandé d'utiliser la commande *Add from Tiff* dans le menu *document*. Il est alors permis grâce à la souris de sélectionner un ensemble de fichiers qui apparaissent automatiquement en bas d'écran. Ainsi, il n'est plus indispensable et moins contraignant de scanner à nouveau les pages du document. L'option de sélection de zone de traitement est en outre toujours possible dans la mesure où l'option *preview* est utilisée par défaut.

-2- Configuration Siinria.set : Scan Image

Cette configuration permet de scanner les documents pour obtenir des fichiers au format image, en l'occurrence .tif. Ce format d'image pourra ensuite être retraité pour être transformé en format jpg, gif, ou bmp grâce à l'utilitaire xv. Au terme de cette étude, les résultats obtenus en format image sont convenables et parfaitement présentables sur une plateforme WWW. Le texte du sommaire est lisible même lorsqu'il s'agit d'une police extrêmement petite.

-3- Configuration Stiinria.set: Scan Text and Image

Cette configuration permet logiquement de produire la somme des résultats des deux précédentes configurations, en l'occurrence des fichiers texte et des fichiers image.

2.2 - Numérisation : reproduction électronique d'un document papier

La reproduction électronique d'un document papier est obtenue à l'aide du scanner. Si le document reproduit est textuel ou comporte des parties textuelles, un programme de reconnaissance optique de caractères appliqué à l'image permettra de transcrire le texte en fichier ASCII pour pouvoir exploiter le contenu textuel du document. Nous allons étudier dans un premier temps les caractéristiques du format image, et ensuite celles du format texte.

2.2.1 - Le format image

Le mode de codage numérique du document ne traite que l'image du document, qu'il s'agisse de texte ou d'illustration. Il permet une reproduction du document avec une fidélité subordonnée à la résolution appliquée lors de l'analyse, à la qualité du scanner utilisé et au réglage des contrastes de la prise de vue.

Les documents noir et blanc ou les documents couleur peuvent être traités par ces techniques. La quasi-totalité des sommaires scannés étaient en noir et blanc, la photocopie de sommaires en couleur devaient être réalisée par un photocopieur spécifique.

Ce mode de codage, s'il s'applique au texte, ne permet plus d'en connaître le contenu mais seulement d'en assurer une reproduction qui pourra être stockée, transmise ou affichée sur un écran ou imprimée. Il permet, à la résolution près, de garantir l'aspect du document d'origine, page par page.
La place occupée par ce type de document électronique est, pour le représentation d'un texte et après compression, environ dix fois plus importante que celle du fichier ASCII du même document pour une résolution moyenne, et vingt fois pour une résolution plus élevée.

Le document image a pour caractéristiques :

⇨ d'être une reproduction exacte de la présentation du document d'origine, utilisable à l'affichage ou à l'impression,
⇨ d'empêcher l'accès au contenu textuel,
⇨ d'être beaucoup plus volumineux qu'un codage caractères pour un document textuel donné.

2.2.1.1 - Domaine d'application

Tout type d'image sur le principe mais la filière image perd certains de ses intérêts si les exigences de rendu, de qualité d'affichage sont extrêmes. Plus la résolution utilisée est élevée, plus le traitement est long et la quantité d'octets pour stocker l'image est importante.

2.2.1.2 - Résultat obtenu

Les fichiers générés par ce logiciel sont compatibles avec une grande variété de logiciels, de traitement de l'image, d'application de bases de données.

2.2.1.3 - Mode d'exploitation du résultat

Les images sont créées avec un format TIFF, nous les avons ensuite formatées grâce à xv en GIF et en JPG, formats permettant la lecture sur les écrans WWW.

Les documents obtenus sont donc stockés sous forme d'image numérique sur un support disque magnétique, alors qu'ils devraient l'être sur un support plus adapté, tel que le disque optique numérique qui permet un stockage plus important des données.

Ces images ont ensuite été rendues accessibles par une ouverture locale de fichiers. Pour cela, nous avons réalisé un fichier HTML grâce auquel il était possible de consulter les premiers documents scannés.

2.2.1.4 - Avantages

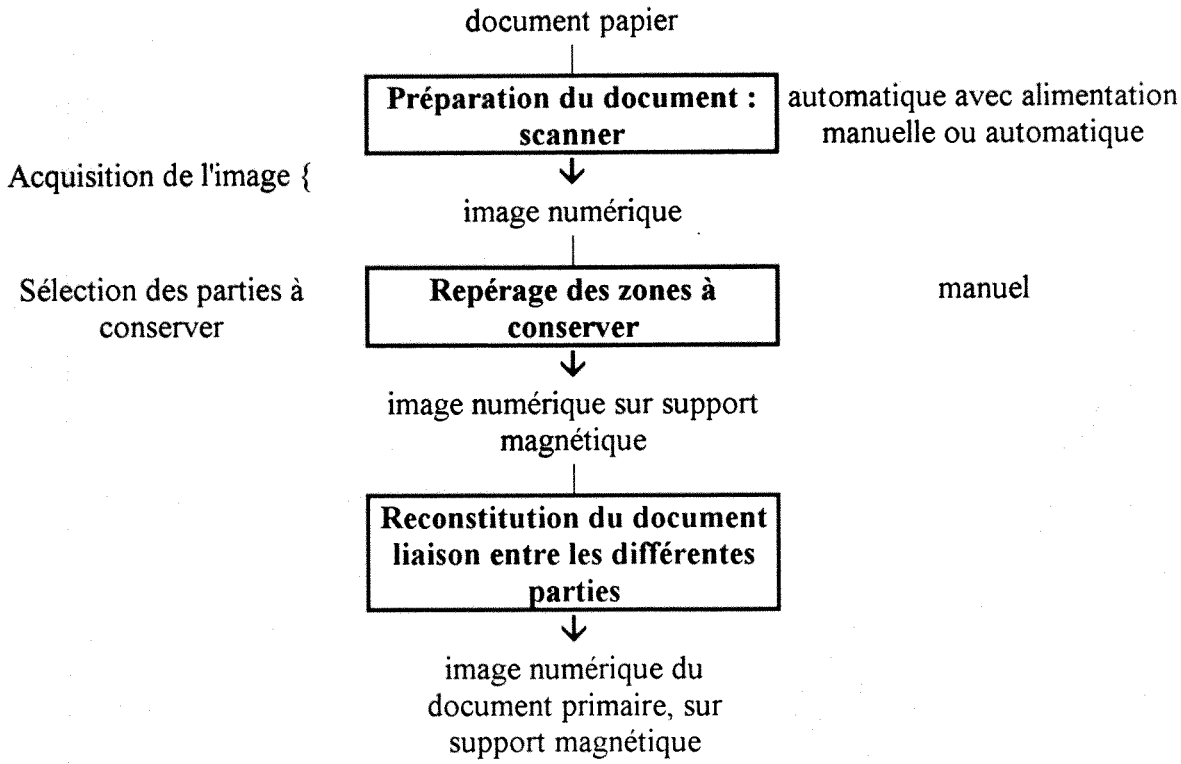
Nos exigences de qualité d'image sont raisonnables. Dés lors, il suffit seulement que les résultats obtenus puissent être lus sans difficulté, pour qu'une telle démarche soit viable. Ces résultats peuvent être repris par une autre application informatique.

2.2.1.5 - Inconvénients, limites, difficultés particulières

La maîtrise des différents éléments de la chaîne de numérisation d'image demande un minimum de connaissances techniques. En effet, toutes les opérations de paramétrage et de configuration ont nécessité un important investissement en temps.

Les temps de réponse sont parfois très décevants et difficiles à maîtriser dans certains contextes d'exploitation, c'est le cas des réseaux locaux par le biais desquels le travail se fait à l'INRIA.

Schéma descriptif de la filière image:



Les principales raisons qui ont poussé à abandonner l'exploitation du format image sont les suivantes :

- ⇨ l'utilisation des formulaires est impossible (Cf. 5 - Formulaire électronique)
- ⇨ la quantité de donnée stockées est trop importante
- ⇨ l'indexation n'a lieu que sur le titre du document
- ⇨ l'impression et la visualisation sur écran prennent trop de temps

Le traitement documentaire n'est pas très bien adapté au format image, il l'est davantage pour le format texte que nous allons maintenant présenter.

2.2.2 - Le format texte

Le codage du texte communément désigné par le terme de fichier ASCII (American Standard Code for Information Interchange) signifie que toutes les parties textuelles d'un document électronique sont codées en mode caractère plus un certain nombre d'éléments de présentation du texte tels que la ponctuation ou les retours à la ligne qui marquent un paragraphe. C'est ce type de codage qui permet l'indexation de chaînes de caractères du document et la recherche de ces chaînes de caractères par comparaison entre la question et les mots indexés.

Ces techniques s'appliquent aujourd'hui de façon courante dans tous les domaines concernant le document électronique à caractère textuel, dès lors que l'on désire accéder au contenu d'un document, que ce soit à des fins de recherche documentaire, ou plus simplement pour aider à la correction ou à la mise en jour d'un document électronique lors de son élaboration.
Tout traitement de document électronique, dont une des fonctionnalités est l'accès au contenu textuel, a donc pour réponse technique possible le fichier ASCII du document.

Les caractéristiques du fichier ASCII sont les suivantes :

⇨ offrir la possibilité de recherche et de traitement de texte,
⇨ permettre un échange sur une normalisation minimum entre différents systèmes de traitements de texte,
⇨ tenir peu de place en stockage machine : un octet par caractère,
⇨ empêcher de mémoriser une présentation sophistiquée du document électronique, mais conserver seulement les éléments minimum permettant l'intelligibilité du texte.

Le fichier ASCII peut être obtenu par saisie du texte au clavier, c'est notamment ce qui est fait pour les notices lors du catalogage. Il peut également résulter comme nous l'avons fait du traitement par OCR (Optical Character Recognition ou Reconnaissance Optique de Caractères) d'un document électronique textuel codé en mode image.

2.2.2.1 - Domaine d'application

La filière texte s'applique aux textes dactylographiés ou imprimés, à l'exclusion de toute image et de tout graphique. Le choix systématique d'une visualisation préalable avec la possibilité de sélectionner les parties d'un document, a permis d'écarter toutes les images de nos traitements.
Ces textes doivent offrir une bonne qualité d'impression, c'est la raison pour laquelle il a été décidé d'opter pour le traitement des photocopies d'un document, plutôt que les pages originales dont la qualité varient énormément en fonction du type de document.

2.2.2.2 - Résultat obtenu

Nous avons obtenu un document primaire sous forme de texte codé et non pas sous forme d'image. Tout ce qui n'est pas du texte sur l'original n'est pas traité.

2.2.2.3 - Mode d'exploitation du résultat

Ces fichiers textuels ont notamment pu être reformatés, indexés et interrogés.

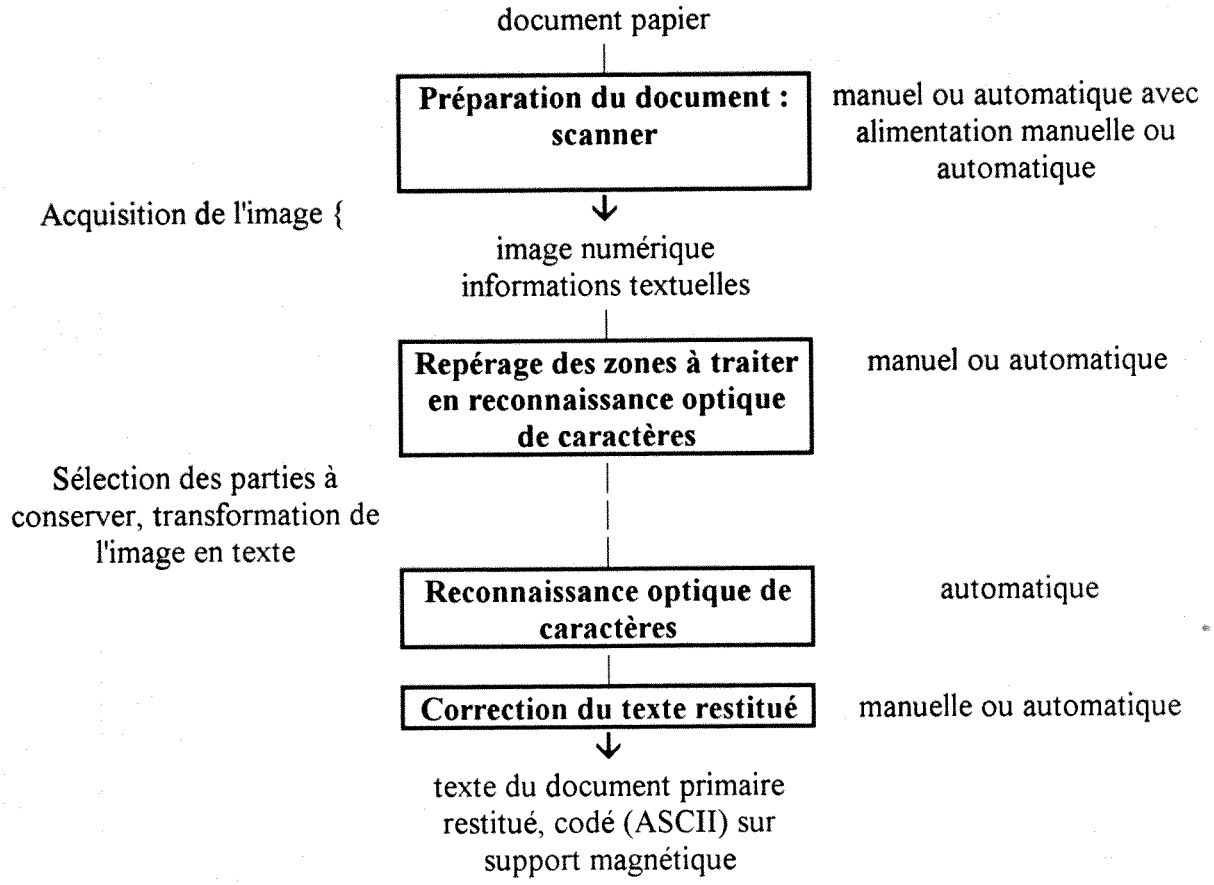
2.2.2.4 - Avantages

Aux avantages de la filière image numérique s'ajoutent les suivants :
Le coût de stockage et d'exploitation est très réduit, dans la mesure où le texte codé est beaucoup moins volumineux que l'image.
Les possibilités de traitement informatique sont diverses : indexation automatique (dans les limites de cette technique), recherche en texte intégral.
Les périphériques de restitution sont des terminaux informatiques classiques; ils n'ont pas besoin d'offrir une qualité d'affichage particulière puisqu'il s'agit de texte qui est traité et non pas de l'image.
Compte tenu des gains en volume, ces textes pourront être stockés sur des supports magnétiques, les temps de réponse seront meilleurs.

2.2.2.5 - Inconvénients, limites, difficultés particulières

Les principales limites sont celles du logiciel de reconnaissance optique de caractères. Les premiers résultats de l'OCR obtenus sans session d'apprentissage étaient médiocres. Après de nombreux apprentissages un taux très faible d'erreurs persiste, mais les résultats sont acceptables.
En l'état actuel de la technique, l'automatisation intégrale de la chaîne d'acquisition est rarement possible. Il faut souvent prévoir un contrôle manuel à l'issue de l'opération de reconnaissance optique de caractères, ce qui pénalise l'application par une augmentation des coûts de fonctionnement. Ce contrôle manuel se traduit dans notre système par l'ensemble des opérations de validation que l'on verra ultérieurement. (Cf. 2.3.3 - Coût de la numérisation : validation et contrôle des résultats)

Schéma descriptif de la filière texte :



Conclusion :

➢ Les résultats obtenus en format texte sont dorénavant honorables. Cependant, un pourcentage d'erreurs persiste. De manière générale, les fichiers textuels produits par l'OCR ScanWorX peuvent être exploités par une recherche en mode texte.
Notons que les fichiers au format image sont largement plus convenables que ceux que nous obtenions en début d'étude. Il est ainsi possible de consulter et lire une image sans difficulté.

Nous allons présenté au cours du point suivant, la démarche concrète à suivre pour numériser et reconnaître les caractères d'une image numérisée.

2.3 - Normalisation de l'utilisation

2.3.1 - Notice technique d'utilisation

Ce mode d'emploi est destiné aux personnes désireuses d'utiliser le scanner. Il est une aide assistée pour pouvoir effectuer une numérisation de document.

Etape n°1 : Lancer le scanner en cliquant sur *scanner* dans le menu des applications ou taper directement dans la fenêtre de travail la commande *scanner -z*.

Etape n°2 : Cliquer sur nouveau document (*new doc*) ou nouveau groupe (*new group*) pour ouvrir un nouveau fichier de travail.

Etape n°3 : Dans le menu Setting, sélectionner la configuration (setting : fichier.set) adéquate en fonction du travail à accomplir. Le chemin d'accès à ces configurations est /scanner/configuration/ :
a) **srcinria.set** : configuration permettant de Scanner, Reconnaître et Convertir en format texte. Cette configuration permet d'obtenir par défaut un fichier dont l'extension est txt.
b) **siinria.set** : configuration permettant de Scanner au format Image. Cette configuration produit un répertoire contenant des fichiers dont l'extension est tif.
c) **stiinria.set** : configuration permettant de Scanner du Texte et de l'Image. Cette configuration produit à la fois des fichiers avec une extension txt et un répertoire contenant des fichiers avec une extension tif.

Etape n°4 : Cliquer sur *Go*. Une fenêtre apparaît dans laquelle vous choisirez le nom et l'emplacement du fichier et/ou du répertoire (*specify output base*) qui va recevoir les documents à traiter (nom de répertoire quand il s'agit d'images, nom de fichier quand il s'agit de texte).

Etape n°5 : Dès que le *preview* (visualisation plein écran du document scanné) apparaît, il faut alors charger ou créer un gabarit (*template* : zone de traitement). Pour charger, cliquer sur *load template* dans le menu Template. Le chemin d'accès aux gabarits existants est /scanner/modèles/.
Pour créer un gabarit, faire glisser la souris pour sélectionner la zone de travail. Utiliser pour cette opération le bouton *T* correspondant à la sélection de zone de texte (*Text zone*) ou le bouton avec *motif* (à droite du T) correspondant cette fois, à la sélection d'une zone image (*Image zone*), la sélection préalable permet d'écarter le traitement des parasites, c'est à dire les zones de bruit.
Qu'il soit créé ou chargé, le gabarit restera actif durant toute la session de travail. Dans le cas où le gabarit ne correspond plus au document scanné (photocopie non uniforme), il suffit de cliquer dans le menu *edit* sur la commande *clear* ou *clear all* si plusieurs zones ont été sélectionnées.
Quel que soit le document scanné, toutes les configurations intègrent une visualisation préalable (*Preview*), cette démarche est nécessaire pour obtenir de meilleurs résultats (sélection préalable des zones de travail).

Etape n°6 : Une fois le gabarit correctement placé, cliquer sur *Résumé*. Dès lors, le logiciel ne prend en compte que les zones sélectionnées et écarte le reste (parasite, bruit ...).

Etape n°7 : Une fenêtre *Continue scanning* apparaît à l'écran, cliquer sur *yes* si d'autres pages sont à scanner.

Etape n°8 : Après avoir cliquer sur *no*, il suffit de cliquer sur *close* dans le menu *File*. Une fenêtre apparaît *This window has been modified, save before closing*, cliquer sur *no* pour ne pas modifier la configuration par défaut.

Etape n°9 : Cliquez sur *Exit* pour quitter l'OCR

2.3.2 - Coût de la numérisation : validation et contrôle des résultats

Comme nous l'avons vu dans la précédente partie, la numérisation demande à la fois des compétences techniques mais également un important investissement en temps. Il est indispensable qu'un membre du personnel qualifié soit affecté à cette tâche qui demandera un temps de travail important.
Il a été calculé un temps moyen de 25 minutes pour traiter un sommaire et le rendre prêt à l'utilisation. Ce temps comprend la connexion au scanner, la numérisation du document, ainsi que son contrôle, opération indispensable, longue et à réaliser avec attention.

La personne chargée de ce travail devra en particulier effectuer une opération d'apprentissage de l'OCR, c'est à dire en mode vérification, lorsqu'il s'agira d'un document inédit n'ayant encore jamais été traité. De manière plus générale, il faudra vérifier la qualité du document obtenu, c'est à dire son taux d'erreur. Il faudra également être très attentif à la structure du sommaire, la modifier lorsqu'elle ne peut être exploitée par nos programmes et enfin porter attention à la présentation des articles qui doivent être impérativement séparés les uns des autres par une ligne vide. Cette dernière contrainte s'explique comme nous le verrons dans la partie indexation, par le fait que l'indexation sous WAIS a lieu avec l'option -para, tous les paragraphes seront considérés comme des articles.
Nous estimons avec une base de 25 minutes par document (de 4 pages en moyenne) et plus de 300 conférences par an (en supposant que les conférences soient systématiquement traitées par OCR) que le temps nécessaire pour réaliser ce travail est 150 heures par an.

3 - Récupération des données sur Internet

De nombreux services en ligne à l'initiative le plus souvent d'éditeurs, constituent notre troisième source d'acquisitions des données. Il suffit de se connecter chez certains grands éditeurs scientifiques pour découvrir un service gratuit de mise à disposition de sommaires de périodiques et de conférences.

Les principaux éditeurs chez lesquels nous nous sommes connectés pour collecter des données sont IEEE, ACM et Springer. Nous les avons choisis dans un premier temps car ils proposaient un service en ligne, mais également dans la mesure où notre fonds contient de nombreuses revues ou conférences éditées par eux.
Etant donné l'importance considérable qu'accordent ces éditeurs à l'information électronique, nous avons constaté une offre sérieuse et rigoureuse d'un point de vue qualité. En effet, les documents mis à disposition sont toujours structurés en champs et une mise à jour régulière a lieu. Les services offerts sont nombreux, notons la recherche d'information, ou la possibilité de commande d'un ouvrage.
Parmi les informations gratuites, deux points d'entrée ont été privilégiés : celui des conférences, ainsi que celui des périodiques. Après avoir accédé à l'information voulue, il suffit sous l'interface WEB utilisée de sauvegarder le document sous un format texte. Dès lors, ce document peur être renommé à notre guise et exploité par nos programmes.

B - Organisation du stockage des données

1 - Les points d'accès à l'Inriathèque

Une fois la connexion réalisée, la page de garde de l'Inriathèque permet de découvrir deux principaux points d'entrée : les nouvelles acquisitions et les archives.

1.1 - Nouvelles acquisitions

Ce point d'entrée permet d'exploiter le fonds de documents les plus récemment acquis. La période durant laquelle, un document est considéré comme nouvelle acquisition est d'un mois. Par la suite, il sera automatiquement transféré dans le répertoire adéquat des archives selon qu'il s'agisse d'un sommaire de périodique ou de conférence, ou bien d'une notice bibliographique. L'interface des nouvelles acquisitions est respectivement la même que celle des archives à la différence que le fond d'écran est différent pour éviter toute confusion dans la recherche.

1.2 - Données archivées

Après être restés disponibles durant un mois dans les répertoires nouveautés, les documents sont archivés. Les utilisateurs voulant consulter l'Inriathèque durant une longue période auront toujours la possibilité d'interroger les archives. Bien sûr, lorsque l'on est connecté sur les archives, les entrées correspondent à d'autres répertoires et d'autres index, le fonds est en effet différent du précédent. A plus long terme, il est envisagé de prévoir sur l'interface archive, des entrées par année.

Trois possibilités d'interrogation sont offertes à partir des deux points d'entrée précédents :

① Sommaire de périodique ou/et de conférence

Ce point d'entrée consiste à interroger sur un titre de périodique ou un nom de conférence. La recherche est faite sur un plus grand nombre de termes, par conséquent le risque de bruit est présent, ceci étant donné que l'indexation a eu lieu sur le contenu textuel du sommaire tout entier.

Nous avons très souvent remarqué que les chercheurs montraient un intérêt pour un titre de revue ou une conférence particulière, cette interrogation sur le sommaire est susceptible de satisfaire à leur demande. Cette solution est profitable aux personnes qui travaillent sur un sujet suffisamment large et qui s'intéressent au domaine relatif à l'ensemble des termes utilisés par leur équation de recherche.

② Article de périodique ou de conférence

L'optique dans ce cas est tout à fait différente. Ce point d'entrée consiste à interroger sur différents éléments pertinents de l'article tels que l'auteur, l'année, ou les mots du titre de l'article.

Le principal avantage de cette entrée est qu'elle permet d'atteindre un degré de précision plus grand. Le phénomène du bruit est alors réduit avec en contrepartie un risque de silence si les termes de la recherche sont trop précis et nombreux.

Le résultat de la recherche est dans cette hypothèse un seul ou une série d'articles répondant à l'équation de recherche. Chacun d'entre eux équivaut à quelques lignes de texte qui sont inclus dans un formulaire permettant de faire une demande de photocopie (Cf. 5 - Formulaire électronique : demande de photocopie).

A partir de chaque article obtenu, il est possible grâce à un bouton (ancre, référence), de remonter au sommaire de conférence ou de périodiques correspondant.

③ Notices bibliographiques de monographies, rapports ou thèses

Ce point d'entrée permet d'obtenir une notice bibliographique, l'interrogation se fera sur des termes pertinents contenus dans le texte intégral de la notice.

Commentaires

La finalité de l'Inriathèque électronique n'est pas seulement une copie d'article, cela peut aussi être une lecture directe au centre de documentation ou une bibliographie personnelle pour un chercheur.

Rappelons que pour l'entrée par sommaire, lorsque plusieurs articles auront été sélectionnés, chaque demande sera transmise par un message électronique (mail). Ceci permettra d'éviter toutes erreurs ou confusions dans les demandes et à plus long terme servir de base aux statistiques qui seront effectuées sur l'Inriathèque.

2 - Hiérarchie et arborescence des répertoires

Rappelons que cette arborescence n'est pas figée et pourra faire l'objet de modifications en fonction des besoins. Dans le répertoire Inriathèque, nous trouvons l'arborescence suivante :

Niveau 1 :

Répertoire Inriathèque

① Répertoire configuration	② Répertoire vérification	③ Répertoire modèle
④ Répertoire Nouveautés	⑤ Répertoire Archives	⑥ Bin
⑦ Admin	⑧ Scanner	

- Les répertoires 1, 2 et 3 contiennent respectivement les fichiers de configuration, de vérification et de gabarits.
- Les répertoires 4 et 5 correspondent au point d'entrée sur les nouveautés et sur les archives.
- Le répertoire 6 contient tous les programmes et scripts servant au traitement des documents de l'Inriathèque.
- Le répertoire 7 contient toutes les documentations relatives à l'Inriathèque.
- Le répertoire 8 contient tous les sommaires scannés qui seront redirigés après traitement dans les répertoires adéquats de conférences ou de périodiques.

Niveau 2 :

Répertoire **Nouveautés**

① Répertoire Conférences	② Répertoire Périodiques	③ Répertoire monographies, rapports de recherche, thèses

- Le répertoire 1 contient tous les nouveaux sommaires de conférences.
- Le répertoire 2 contient tous les nouveaux sommaires de périodiques.
- Le répertoire 3 contient toutes les dernières notices extraites des fichiers TEXTO relatifs aux monographies, rapports de recherche et thèses.

Répertoire **Archives**

① Répertoire Conférences	② Répertoire Périodiques	③ Répertoire monographies, rapports de recherche, thèses

Le contenu de ces répertoires est respectivement le même que ceux qui ont été vus précédemment, à la différence qu'ils contiennent les données archivées, c'est à dire toutes celles qui sont restées au moins un mois dans les répertoires de nouveautés.

Section 2 : Les Procédés de traitement des données

A - Chaîne de traitement du document : les sommaires de périodiques et de conférences

Les sommaires de conférences et de périodiques sont les documents qui ont demandé le plus de traitements. Les sommaires reçus par abonnement ou récupérés sur Internet, sont plus simples à exploiter dans la mesure où ils sont théoriquement dotés d'un format structuré en champs. Les programmes réalisés n'auront donc aucun mal à reconnaître les articles, ainsi que le début du sommaire.
Les sommaires traités par OCR sont plus délicats à traiter dans la mesure ne sont pas structurés.
Qu'il s'agisse d'un sommaire de périodique ou d'un sommaire de conférence, rappelons que le traitement général est le même. En effet, un sommaire prêt à être exploité contient toujours les mêmes données : un titre (identifiant), un terme désignant le début du sommaire et une liste d'articles.

Les principaux outils et logiciels qui nous ont permis de réaliser toute cette partie programmation sont :

⇨ **Les Scripts**

A partir des commandes shell d'UNIX, il est possible de créer des scripts, c'est à dire des programmes contenant des commandes UNIX. Ces scripts sont exécutables directement en tapant leur nom à condition de les avoir conformément rendu exécutables (chmod ...) ou bien en les précédant de la commande qui permet de lancer un programme : sh.
Parmi les commandes UNIX, deux nous été d'une aide précieuse pour le traitement informatique de données textuelles :

① AWK : la commande AWK est un filtre qui permet de manipuler un texte. Ce dernier est entièrement programmable et utilise pour les instructions conditionnelles, les boucles ou les variables une notation semblable à celle du langage C.

② SED : la commande SED est un éditeur non conversationnel utilisé en tant que filtre. Le rôle de SED ressemble beaucoup à celui de AWK. Cette commande permet entre autre une manipulation aisée de grandes quantités de texte.

⇨ **PERL**

Il s'agit d'un langage de programmation structuré qui permet de réaliser les instructions, les affectations, les procédures ou fonctions les plus classiques en programmation.

Le répertoire **CGI** (Common Gateway Interface)-**BIN** regroupe de manière générale tous les fichiers exécutables qui vont apporter des fonctionnalités supplémentaires au serveur WEB. Au niveau de l'Inriathèque électronique, le répertoire qui contiendra les scripts et programmes ayant servi à traiter les documents est **BIN**.

Nous devons au préalable préciser que ces programmes et scripts seront présentés séparément mais qu'ils seront intégrés par la suite dans un seul et même fichier. Les différentes étapes de la chaîne de traitement sont l'extraction des notices bibliographiques ou le complément de celle-ci, le reformatage et l'indexation des sommaires et enfin la transmission d'une demande de photocopie par le biais d'un formulaire.

1 - Extraction des notices bibliographiques

La première étape de la programmation a consisté lorsqu'il s'agissait de sommaires scannés, à récupérer la notice correspondante dans la base TEXTO. Nous faisons ainsi, un lien entre le sommaire et sa notice dans le catalogue correspondant. Pour cela, il a fallu intituler le sommaire en question par un élément unique lui étant propre, en l'occurrence le numéro d'inventaire. Le premier programme réalisé consistait à trouver la notice du sommaire et à renommer le sommaire en fonction des informations trouvées dans la notice.
Par la suite, lorsqu'il s'agira de sommaires reçus par abonnement ou récupérés sur Internet, il suffira d'utiliser les identifiants utilisés par le document lui-même et de récupérer les éléments manquants, tel que la cote du document dans le fichier TEXTO correspondant. En effet, les précédents sommaires contiennent déjà les éléments nécessaires permettant de l'identifier avec exactitude : titre , volume, numéro, date ...

2 - Reformatage : deux méthodes

Un document quel qu'il soit, doit pour être affichable de manière spécifique sur une interface WEB, posséder un format adéquat. Dès lors, un script permet d'ajouter au fichier du sommaire correctement intitulé, des balises HTML produisant ainsi un fichier au format HTML (Cf. Section 3 : Diffusion des données A- HTML). Ainsi, la localisation, la date, le titre du document sont correctement attribués au sommaire et il peut être visualiser sur WWW.
Nous avons également dans ce programme, introduit des instructions retouchant le texte grâce à de nouvelles polices de caractères permettant une lecture plus conviviale, ainsi que des instructions permettant à l'article de faire l'objet d'une demande de photocopie (Cf. ci-dessous 5 - Formulaire électronique et demande de photocopie).

Deux méthodes pour reformater nos données nous étaient offertes, la première avant, la seconde après l'indexation (à la volée) :

2.1 - Reformatage avant indexation

Précisons tout d'abord qu'il s'agit de la méthode la plus fréquemment utilisée. Cette technique de reformatage avant indexation présente l'avantage d'être effectuée une fois pour toute en tout début de travail. Pour les sommaires scannés, le programme aura pour tâche de transformer les éléments de la notice précédemment extraite de TEXTO, en un pavé HTML. Celui-ci est ensuite intégré en en-tête du sommaire qui a déjà été balisé et retouché. Le programme a pour tâche d'insérer des balises HTML. Pour les sommaires obtenus différemment, le travail consistera à ajouter les éléments manquants à l'en-tête.

Lorsque le sommaire provenait du scanner, certains éléments du sommaires devait être repérés et reconnus, c'est notamment le cas du terme désignant le début du sommaire. Il s'agit en l'occurrence des termes tels que Contents, Sommaire, Table des matières etc. auxquels nous devions ajouter des éléments permettant aux utilisateurs de remonter à leur guise au sommaire.

<u>Repérage du terme annoncant le début de sommaire</u>

Comme nous venons de le voir ci-dessus, plusieurs termes peuvent être utilisés pour annoncer le début d'une table des matières. Ces termes sont présentés dans le tableau suivant :

Contents	Table of contents
Sommaire	Sommaire/Conférence
Table des matières	Proceedings

Dans certains cas, aucun terme n'annonce le début du sommaire. Cette liste de mots a été enregistrée dans le script correspondant, elle permettra dès qu'un des termes est rencontré, d'insérer une balise et d'apposer une adresse relative (URL relatif avec la commande NAME et le caractère #) permettant à l'utilisateur de retourner au sommaire lorsqu'il a par exemple obtenu en réponse un article.

Modalités d'apparition et reconnaissance des articles de sommaires scannés

Il faudra que chaque article du sommaire soit isolé et traité pour être convenablement affiché à l'écran. C'est ainsi, que chaque article se verra reformater grâce à un formulaire électronique. Ce formulaire devra entrer en action, lorsque WAIS renverra un article en réponse à une question et permettre de faire une demande de photocopie (Cf. 5 - Demande de photocopie)

Selon les nombreux type de sommaires rencontrés, les articles se présentent de la manière suivante :

Il arrive très fréquemment que le sommaire soit structuré en parties. Il est difficile d'établir une règle générale d'apparition des articles. Ces nombreux cas de figure constitueront un difficulté pour la programmation qui devra tenir compte de cette hétérogénéité pour faire apparaître correctement et indépendamment les articles.
La difficulté tient surtout au fait que le texte numérisé n'est en aucun cas structuré en champs, la seule solution est dans ce cas d'utiliser une indexation par paragraphe (option de WAIS : -para). Les erreurs de reconnaissance d'article et par conséquent de réponses erronées, constituent l'argument le plus fort pour prétendre à un abonnement aux sommaires de conférence plutôt que devoir scanner systématiquement ceux-ci.

Quelques cas de figure :

① Cas de figure n°1

article n°1 : Titre de l'article + série de points + auteur
une ligne de séparation ...

article n°2 : Titre de l'article + série de points + auteur
une ligne de séparation ...

article n°3 : Titre de l'article + série de points + auteur
une ligne de séparation ...

Il s'agit de la structure la plus simple à exploiter. Un article est constitué de deux lignes. La première comprend le titre de l'article, une série de points, le numéro de page. La seconde ligne présente les auteurs de l'article. Chaque article étant séparé par une ligne vide. En effet, l'indexation avec l'option -para n'aura aucun mal à reconnaître ces articles.

② Cas de figure n°2

SOMMAIRE

Part 1 titre partie

Section 1 titre section
numéro de page + Titre de l'article
une ligne de séparation ...
Section 2 titre section
numéro de page + Titre de l'article
une ligne de séparation ...
Part 2 titre partie ...

③ Cas de figure n°3

SOMMAIRE

Part 1 titre partie
Titre de l'article + ligne de points + numéro de page
une ligne de séparation ...
Titre de l'article + ligne de points + numéro de page
une ligne de séparation ...
Part 2 titre partie ...

④ Cas de figure n°4

SOMMAIRE

I partie
1 sous-partie
1.1 sous-sous-partie
Titre article + ligne de points + numéro de page
aucune séparation entre les article

⑤ Cas de figure n°5

SOMMAIRE

Le sommaire est dans ce cas constitué de deux colonnes. Des sous-titres sont comprises dans le sommaire (exemples : Invited papers, Communications...)
Une ligne de séparation entre les articles d'une colonne.
Titre de l'article + ligne de points + numéro de page

Ce cas de figure est le plus délicat à traiter, dans la mesure où l'OCR ne peut quand il s'agit de deux colonnes reconstituer le sommaire. Puisque les sommaires étaient préalablement photocopiés, la solution pouvait se traduire par le découpage de la page en deux pour ensuite scanner séparément chaque colonne.

Nous avons également relevé le cas de sommaires dont les articles ne contenaient pas de numérotation. Le document était de petite taille. Il arrive également que la numérotation ne soit pas chiffrée mais en lettres.

Commentaires :

Lorsque le sommaire est scanné, nous rencontrons donc une grande variétés de structures de sommaire. En effet, cet ensemble très hétérogène de données oblige à prendre en compte tous les cas de figure imaginables.

2.2 - Reformatage à la volée

Cette démarche demande davantage de programmation car elle consiste à utiliser un CGI-BIN s'intercalant entre la demande de l'utilisateur et la réponse renvoyée à la recherche. Cette technique consiste à demander à l'indexeur WAIS d'effectuer une recherche et de passer la main à un programme qui renvoie les réponses correctement reformatées. C'est notamment, la méthode utilisée pour les notices de monographies, de thèses et de rapports qui sont reformatées dès que WAIS trouve une réponse à une question posée. Cette façon de faire n'a pas été utilisée pour les sommaires, dans la mesure où le reformatage avant indexation est selon M. AUBRIE, la technique la plus simple et la plus adaptée pour la période de temps qui nous était impartie. Elle nous permettait ainsi, d'obtenir des résultats concrets et exploitables. Néanmoins, rappelons que l'indexation à la volée pour les sommaires est actuellement en cours d'examen.

3 - Indexation des documents

Il faut dans un premier temps préciser qu'au centre de documentation, les indexations n'ont pas lieu sur les champs. En effet, celles-ci se font toujours sur le texte intégral de la notice. Cette méthode engendre quelques difficultés avec notamment un risque de bruit plus élevé. Cependant, c'est en connaissance de cause que ce choix a été fait. Le fonds du centre de documentation est suffisamment restreint pour pouvoir indexer de la sorte et ne pas poser trop de problèmes. Pour le fonds d'une bibliothèque universitaire, le choix d'une indexation sur champ se justifie davantage. Néanmoins, le personnel documentaliste est conscient des avantages que pourrait procurer l'indexation sur champs, il nous a été confié que l'interrogation notamment sur le champ auteur pourrait être d'un grand intérêt. A l'avenir, la question sera étudiée plus en détail, d'autant plus que les outils et versions de WAIS permettant une telle indexation sont ceux que nous avons utilisés pour les bases de l'Inriathèque.

3.1 - Présentation de WAIS

Les principales fonctions qu'apporte l'indexeur WAIS sont les suivantes (pour de plus amples informations sur WAIS Cf. Section 3 B - Les bases WAIS) :

⇨ Constitution d'index à partir de mots contenus dans des fichiers de formats divers (texte ASCII, Postscript, DVI, MIME, SGML, GIF, TIFF...);
⇨ Interrogation d'index à distance en fournissant une liste de mots, puis en élargissant au besoin la recherche aux documents voisins par un mécanisme dit de "relevance feedback"; ce mécanisme consiste à inclure dans la question des parties de documents déjà trouvées;
⇨ Rapatriement des documents sélectionnés et possibilité de les visualiser sur son poste de travail avec les filtres de son choix.

3.2 - Les commandes d'indexation

C'est à partir du résultat des sommaires reformatés que l'indexation pouvait avoir lieu. Le sommaire est indexé avec l'option -T HTML, c'est à dire une indexation du texte brut des fichiers HTML. L'inconvénient d'une telle démarche est que le résultat des recherches sous WAIS est frappé d'un phénomène de bruit plus important. En effet, si l'on indexe les pages HTML, de nombreux termes deviennent complètement insignifiants. C'est notamment le cas des instructions contenues dans les balises HTML : input, hidden, action, get, method, submit ... . Rappelons que le reformatage avant indexation a eu lieu pour transformer chaque article du sommaire en un formulaire de demande de photocopie.

Option d'indexation :

Les options suivantes sont à utiliser à la suite de la commande d'indexation waisindex :

⇨ -a : ajoute à un index déjà existant
⇨ -m 10 : limite le nombre de réponses retournées (exemple 10)
⇨ -stop fichier : indexe en tenant compte d'un fichier contenant une liste de termes ou expression à ne pas prendre en compte dans l'indexation
⇨ -export : permet de rendre l'index interrogeable à partir de l'extérieur par défaut sur le port 210 du serveur.
⇨ -t format : indexe un format donné de fichier
⇨ -T type : indexe en attribuant un type à un document
⇨ -M type type : indexe en attribuant plusieurs types à un document
⇨ -d nom : attribue le nom d'index
⇨ -para : indexe en mode paragraphe
⇨ -nocat : ne créé de catalogue suite à la question posée. Le catalogue avec une extension cat est ensuite créé manuellement et sert à une aide en ligne dans le cas où WAIS ne donne aucun résultat à la question posée.
⇨ -r : indexe récursivement tous les répertoire de la hiérarchie indiquée.

Les intitulés d'index :

Le nombre important d'index et leurs fréquentes utilisations nous ont poussé à porter attention aux intitulés qu'on allait leur attribuer. Les noms doivent être pertinents et facilement compréhensibles par les personnes qui les manipulent. La liste suivante présente l'ensemble de nos index :

▫ Inriath-Nouv-Conf-Somr : cet index correspond au point d'entrée sur le fonds des nouveaux sommaires de conférence.

▫ Inriath-Nouv-Perio-Somr : cet index correspond au point d'entrée sur le fonds des nouveaux sommaires de périodique.

▫ Inriath-Nouv-Conf-Arti : cet index correspond au point d'entrée sur le fonds des articles des nouveaux sommaires de conférence.

▫ Inriath-Nouv-Perio-Arti : cet index correspond au point d'entrée sur le fonds des articles des nouveaux sommaires de périodique.

▫ Inriath-Archi-Conf-Somr : cet index correspond au point d'entrée sur le fonds des sommaires de conférence archivés.

▫ Inriath-Archi-Perio-Somr : cet index correspond au point d'entrée sur le fonds des sommaires de périodique archivés.

▫ Inriath-Archi-Conf-Arti : cet index correspond au point d'entrée sur le fonds des articles des sommaires de conférence archivés.

▫ Inriath-Archi-Perio-Arti : cet index correspond au point d'entrée sur le fonds des articles des sommaires de périodique archivés.

▫ Inriath-Perio-Conf-nouv : cet index correspond au point d'entrée sur le fonds des nouveaux sommaires de conférence et périodique

▫ Inriath-Perio-Conf-archi : cet index correspond au point d'entrée sur le fonds des sommaires de conférence et périodique archivés.

▫ Inriath-Nouv-Mono-Rap-These : cet index correspond au point d'entrée sur le fonds des nouvelles notices de monographies, rapports et thèses.

▫ Inriath-Archi-Mono-Rap-These : cet index correspond au point d'entrée sur le fonds des notices de monographies, rapports et thèses archivées.

Quelques Commandes d'indexation :

1 - (Inriath-Nouv-Conf-Somr)
waisindex -d /usr/local/wais/wais-sources/ Inriath-Nouv-Conf-Somr -export -nocat -mem 5 -t filename -T HTML -r *

2 - (Inriath-Nouv-Perio-Somr)
waisindex -d /usr/local/wais/wais-sources/ Inriath-Nouv-Perio-Somr -export -nocat -mem 5 -t filename -T HTML -r *

3 - (Inriath-Nouv-Mono-Rap-These)
waisindex -d /usr/local/wais/wais-sources/ Inriath-Nouv-Mono-Rap-These -export -nocat -mem 5 -t para *

4 - (Inriath-Archi-Mono-Rap-These)
waisindex -d /usr/local/wais/wais-sources/ Inriath-Archi-Mono-Rap-These -export -nocat -mem 5 -t para *

5 - (Inriath-Perio-Conf-nouv)
waisindex -d /usr/local/wais/wais-sources/ Inriath-Perio-Conf-nouv -export -nocat -mem 5 -t para -T HTML -r *

6 - (Inriath-archi-Perio-Conf)
waisindex -d /usr/local/wais/wais-sources/ Inriath-archi-Perio-Conf -export -nocat -mem 5 -t para -T HTML -r *

Ces commandes d'indexation sont comprises dans les scripts les concernants.

4 - Formulaire électronique et demande de photocopie

Lorsque la recherche sur l'index WAIS engendre un résultat, selon la demande qui a été faite un article isolé ou un sommaire composé d'articles apparaît à l'écran. Chaque article comme nous l'avons vu précédemment est reconnu et doit faire l'objet d'un traitement pour pouvoir être demandé en photocopie. Les instructions d'un formulaire ont été introduites lors de l'opération de reformatage vue ci-dessus, permettant ainsi à chaque article d'être enrobé par un masque de saisie. C'est notamment l'instruction <FORM> d'HTML qui permet le traitement des formulaires. Le masque de saisie peut ainsi recevoir des informations (coordonnées de l'utilisateur) et transmettre au service adéquat une demande de photocopie.

Un formulaire apparaît à l'écran sous forme d'une fiche à remplir par l'utilisateur. La soumission d'un formulaire entraîne l'envoi d'ordre GET au serveur WEB, accompagné des paramètres entrés par l'utilisateur. Le formulaire est alors exploité par le serveur à l'aide d'un programme CGI.

La méthode permettant de sélectionner les articles d'un sommaire n'est pas encore définitivement choisie. Il pourra s'agir d'une liste sur laquelle on pourra cliquer pour sélectionner les articles, plutôt que la méthode utilisée qui consiste à enrober chaque article par un formulaire.

En ce qui concerne les demandes de photocopies d'article, les nom et prénom constituent les éléments de base à l'identification de l'usager et au contrôle de ses droits d'accès. Ces éléments sont traduits pour obtenir l'adresse électronique, le lieu de travail, la machine d'où provient la demande, etc. Dans le cas d'une personne extérieure à l'INRIA, il sera nécessaire de vérifier son compte de photocopie. Dans l'hypothèse d'une insuffisance de compte ou même d'absence de compte, les photocopies ne pourront avoir lieu et nous le communiquerons à la personne en question.

En cliquant sur le bouton Submit du formulaire de demande de photocopie, un programme permet la réalisation des tâches suivantes :

- Affichage de la demande
- Accusé de réception dans la messagerie du demandeur
- Contrôle des informations reçues
- Transmission par un message électronique de la demande de photocopie au service correspondant.
- Une trace de la demande est enregistrée dans un fichier. Ce dernier servira notamment à toutes les questions de statistiques.
- Un message est transmis à l'administrateur

Depuis l'émission d'un formulaire électronique, jusqu'à l'accusé de réception du traitement adéquat par un serveur HTTP, les protagonistes sont les suivants :

⇨ HTML qui propose les éléments <FORM>, <IMG> et <ISINDEX> respectivement pour le traitement des formulaires, des images réactives et la communication d'un mot-clef à un moteur de recherche;
⇨ HTTP (et plus particulièrement la version HTTP/1.0) qui procure les méthodes POST et GET pour transmettre des paramètres à un serveur WWW;
⇨ Les serveurs HTTP qui permettent l'appel de programmes externes pour traiter les informations transmises par les clients. Une interface standardisée appelée CGI (Common Gateway Interface) précise les règles d'écriture et d'exécution de procédures (les CGI scripts). Les principaux serveurs HTTP respectent la convention CGI.

La description des formulaires fait partie de HTML niveau 2, elle est bornée par les environnements (commandes ou tags) <FORM> et </FORM>. L'attribut ACTION donne ici l'URL

basé sur ce principe. Il nous a suffit d'emprunter une chaîne de traitement déjà établie pour le fonds des notices de l'Inriathèque.
Cette chaîne de traitement repose sur la multi-interrogation qui consiste dans l'accès à plusieurs bases de données avec une seule formulation de question. Dès lors, le reformatage est fait à la volée. Un CGI script permet de s'intercaler entre la question posée et la réponse renvoyée par WAIS.

Section 3 : Diffusion des données

Les outils ayant contribué à la diffusion des données de l'Inriathèque électronique sont HTML : un langage de description des documents, l'indexeur WAIS, HTTP : un protocole de communication, et le serveur WEB qui combine l'utilisation de tous ces outils et offre la possibilité de mettre à disposition du public, l'information au sens général du terme.

A - HyperText Markup Language (HTML)

La plupart des documents accessibles sur le WEB sont programmés en HTML. Cet outil est indispensable si l'on désire développer un serveur d'informations multimédia, c'est à dire la technique de l'hypertext intégrant les formats tels que l'image, le son etc.. La plaquette de l'Inriathèque et l'ensemble des fichiers qui la composent ont été développés en HTML.

1 - Présentation d'HTML

HTML est une classe de document SGML. Il permet de présenter des documents structurés comprenant un titre, du texte sous diverses tailles et formes, des listes, des points d'ancrage sur lesquels pointeront les liens internes au document. Le texte HTML est un format clair et lisible et peut être fabriqué manuellement à partir de n'importe quel éditeur de textes.

HyperText Markup Language est le langage utilisé pour la diffusion de documents par les serveurs W3. Il s'agit d'un ensemble simple de commandes de formatage de documents, on entre dans un environnement en le citant borné par les caractères < et >, on le quitte en le citant borné par les caractères </ et >.

Comme dans tout hypertexte, on va pouvoir poser des liens sur des mots. En HTML, ces liens vont pointer vers d'autres parties du document, des fichiers ou objets situés éventuellement sur des machines distantes. A la lecture du document, les mots sur lesquels des liens ont été placés seront signalés, en vidéo inversée par exemple, et le fait d'en choisir un en le sélectionnant avec la souris provoquera le rapatriement et l'affichage de l'objet pointé.
Les liens peuvent aussi être placés dans des images et le fait de sélectionner telle ou telle partie de l'image déclenchera telle ou telle action.

Les fichiers pointés dans des liens peuvent contenir du texte, des images fixes ou animées, du son, des séquences audio-vidéo, et des programmes exécutables. Lorsque de tels objets sont sélectionnés par une application cliente, ils lui seront communiqués, à charge pour elle de lui appliquer le traitement informatique approprié.
Les objets distants pointés peuvent aussi être des fichiers accessibles en FTP, des menus pris sur des serveurs GOPHER, des index WAIS, des "news groups", des documents pris sur un autre serveur WWW, des sessions TELNET.

Les URLs :

Les pointeurs utilisés dans les liens sont des URL (Uniform Ressource Locator). Ce sont des chaînes de caractères de la forme "protocole - serveur - port - chemin - d'accès". Citons pour exemple, l'URL correspondant à la connexion sur l'Inriathèque :

HTTP://	merengue:	8005	/net/inriath/ maquette.html
⇩	⇩	⇩	⇩
méthode	machine serveur	port de connexion	chemin d'accès

Le champ méthode indique le protocole à utiliser. Les protocoles pouvant être utilisés sont les suivants : FILE, FTP, HTTP, TELNET, GOPHER, WAIS, NEWS.

L'information contenue dans un lien doit indiquer de manière non ambiguë où et comment atteindre la ressource référencée. l'URL est l'extension au niveau de l'Internet de la notion de nom de fichier sur une machine. Il permet d'adresser de manière précise toute ressource accessible sur l'Internet.

Nous pouvons distinguer entre l'URL absolu et l'URL relatif. Le premier correspond au nom complet d'un fichier par rapport au répertoire root de la machine serveur, tandis que le deuxième consiste soit en un nom complet d'un fichier, soit en un nom relatif au document où l'on se trouve. Il faut entendre par le nom complet d'un fichier, son nom par rapport au répertoire root du serveur W3.

On peut utiliser un URL relatif dans un document pour référencer un autre document localisé sur le même serveur et accessible par le même protocole.

L'utilisation d'URLs relatifs permet de simplifier l'écriture de document HTML. Il est conseillé d'utiliser le nom complet du document (relativement au répertoire root du serveur), car celui-ci restera valide même si le document ou il est utilisé change de place.

L'utilisation du caractère # suivi d'une chaîne de caractères à la fin du nom d'un document permet de référencer un endroit (une ancre) dans celui-ci. ex : http: // web.urec.fr / docs / www / web.1.html#HDR 2 8 permet d'accéder directement à un endroit précis du document. Il faut bien sûr pour cela que l'adresse correspondante existe dans le document en question, la création d'une telle ancre est décrite ci-dessous.

L'utilisation du caractère ? suivi d'une chaîne de caractères à la fin du nom d'un document correspond soit à l'interrogation d'un document indexé, soit à une liste de paramètres pour l'exécution d'un programme. Dans ce cas, les caractères spéciaux (blanc, caractères accentués...) sont changés, ex : wais://quake.think.com:210/directory-of-servers?inria correspond à l'interrogation de la base WAIS directory-of-severs sur le serveur quake.think.com, INRIA étant le paramètre de l'interrogation.

De façon générale, le caractère ? permet de passer des paramètres à un programme qui sera exécuté par le serveur WWW.

2 - Les commandes usuelles

Voici quelques unes des balises les plus couramment utilisés pour les fichiers HTML de l'Inriathèque, pour plus d'informations sur la syntaxe HTML, il suffit de se reporter aux nombreux écrits en la matière, notamment sur le réseau Internet.

⇨ Les amarres de documents = <A HREF> </A>
⇨ Les entêtes de document = <HEADER> </HEADER>

⇨ Les paragraphes = <B> </B>
⇨ Le retour à la ligne =
⇨ Le saut de ligne ou paragraphe = <P>
⇨ Le trait horizontal = <HR>
⇨ Les caractères accentués sont traduits dans une syntaxe HTML, le e grave devient par exemple è
⇨ La mise en forme de textes : le langage HTML définit six niveaux pour reformater le texte. Le niveau 1 dont les balises sont <H1> et </H1> est le niveau de mise en valeur le plus important.
⇨ La mise en forme de mot : I pour l'italique, B pour le gras
⇨ Les liens vers d'autres documents
⇨ Les images dans le texte <IMG SRC>
⇨ Les listes = <UL> </UL>, <LI> </LI>, <OL> </OL>
⇨ Les parties de texte pré-formaté = <PRE> </PRE>
⇨ Les formulaires <FORM> </FORM>
⇨ Les boutons radio ...

3 - La page d'accueil Inriathèque

File Edit View Go Bookmarks Options Directory Help

Back Forward Home Reload Images Open Print Find Stop

Location: http //merengue 8005/kebir/inriatheque/maquette/maquette html

– Rocquencourt –
Centre de Documentation de l'INRIA

Bienvenue sur l'INRIATHEQUE ELECTRONIQUE

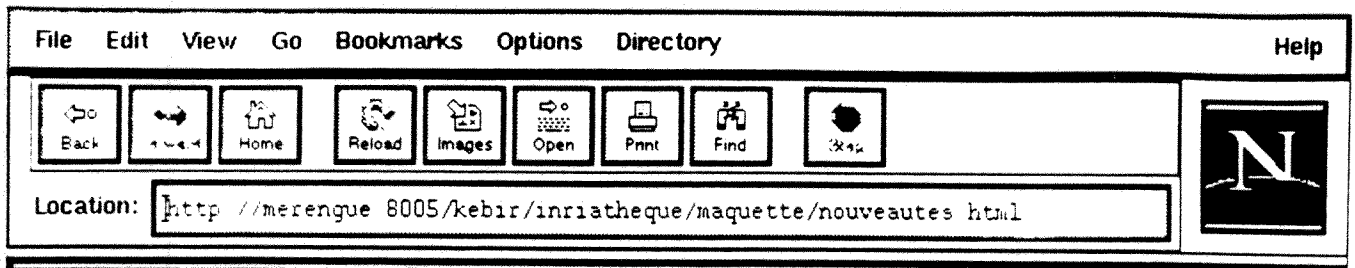
Le contenu de l'Inriathèque est désormais accessible par réseau.
Si vous êtes chercheur,
vous pouvez interroger, consulter, et demander un photocopie d'article.
Pour plus d'Informations sur cette nouvelle version de l'Inriathèque.

Vous avez accès au :

- Catalogue des NOUVEAUTES
- Catalogue des ARCHIVES

Ce service est réalisé et maintenu par le personnel du centre de documentation.
En cas de suggestions, vous pouvez transmettre un message ...

Document Done



CATALOGUE DES NOUVEAUTES

Les informations ...

Pour obtenir une *table des matières*, vous pouvez interroger à partir du :

- Titre de la CONFERENCE
- Titre du PERIODIQUE
- Titres de CONFERENCES et de PERIODIQUES

Pour obtenir un *article*, vous pouvez interroger à partir :

- Des TERMES PERTINENTS de l'article :

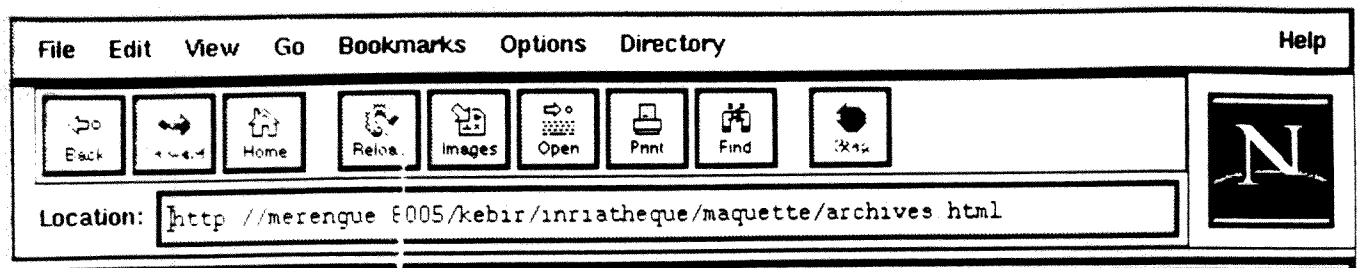
  – Auteur, Date, Mots du titre...

Pour obtenir une *notice bibliographique*, vous pouvez interroger à partir :

- Des TERMES PERTINENTS des ouvrages, des rapports et des thèses :

  – Auteur, Date, Mots du titre...

Retour à la *page de garde* de l'Inriathèque.



CATALOGUE DES ARCHIVES

Les informations ...

Pour obtenir une *table des matières*, vous pouvez interroger à partir du :

- Titre de la CONFERENCE
- Titre du PERIODIQUE
- Titres de CONFERENCES et de PERIODIQUES

Pour obtenir un *article*, vous pouvez interroger à partir :

- Des TERMES PERTINENTS de l'article :

  – Auteur, Date, Mots du titre...

Pour obtenir une *notice bibliographique*, vous pouvez interroger à partir :

- Des TERMES PERTINENTS des ouvrages, des rapports et des thèses :

  – Auteur, Date, Mots du titre...

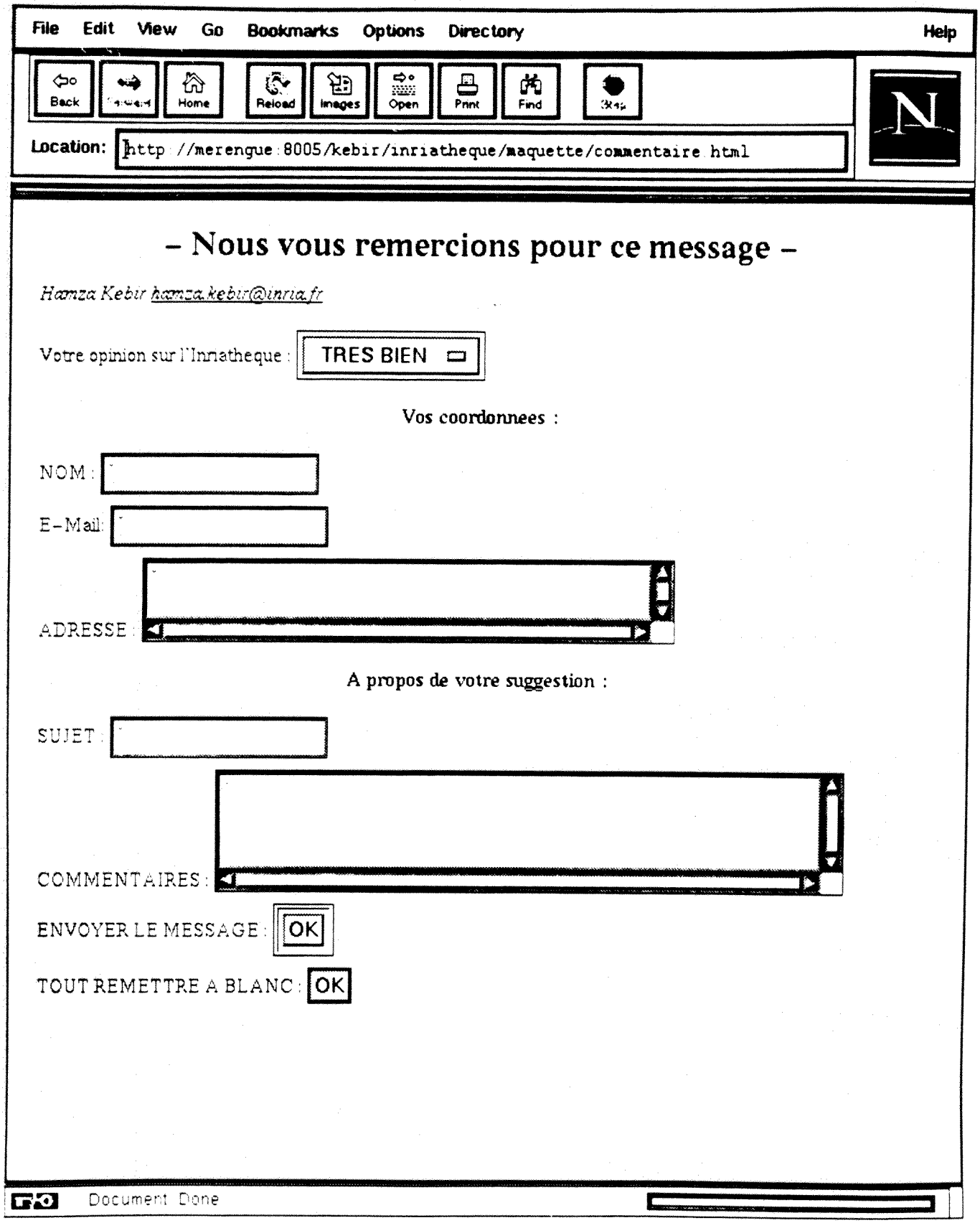
Retour à la *page de garde* de l'Inriathèque.

4 - Quelques fichiers HTML

```
<HTML>
<HEAD>
<A NAME=DEBUT>
<TITLE>inriatheque.html</TITLE>
</HEAD>
<BODY BACKGROUND="cj.jpg">
<FONT size=5>
<B>
<CENTER>
<IMG SRC="Logo-57_Rocq._quadri.gif"><BR><BR>
<IMG SRC="Unite-Rocquencourt03.gif"><BR><BR>
- Rocquencourt -<BR>
<A HREF=HTTP://www.inria.fr:80/Services/Rocquencourt/Doc/plaquette/
centre-de-documentation-fra.html>Centre de Documentation</A>
de l'<A HREF=HTTP://www.inria.fr>INRIA</A><BR>
<BR><BR>
<EM>Bienvenue sur l'INRIATHEQUE ELECTRONIQUE<BR><BR>
<H3>Le contenu de l'Inriath&egrave;que est d&eacute;sormais
accessible par r&eacute;seau.<BR>Si vous &ecirc;tes chercheur,<BR>
vous pouvez interroger, consulter, et demander un photocopie d'article.
<BR>Pour plus d'<A HREF="information.html">Informations</A> sur cette
nouvelle version de l'Inriath&egrave;que.</H3>
</CENTER>
</FONT><BR>
<IMG SRC=http://www-doc-rocq.inria.fr:8004/bm/images/sep/ligne.gif><BR>
<H3>
Vous avez acc&egrave;s au :
<UL>
<LI>Catalogue des <A HREF="nouveautes.html">NOUVEAUTES</A><BR><BR>
<LI>Catalogue des <A HREF="archives.html">ARCHIVES</A>
</UL>
</H3>
<IMG SRC=http://www-doc-rocq.inria.fr:8004/bm/images/sep/ligne.gif><BR>
<H3><EM>Ce service est r&eacute;alis&eacute et maintenu par
<A HREF="personnel.html">le personnel</A> du centre de documentation.<BR><BR>
En cas de suggestions, vous pouvez transmettre <A HREF="commentaire.html">
un message ...</EM></A>
</H3>
</BODY>
</HTML>
```

```
<HTML>
<HEAD>
<A NAME=DEBUT>
<TITLE>archives.html</TITLE>
</HEAD>
<BODY BACKGROUND="cj.jpg">
<FONT size=5>
<B>
<CENTER>
<EM>CATALOGUE DES ARCHIVES<BR><BR>
</CENTER>
<H4>Les <A HREF="information.html">informations</A> ...</H4>
</FONT><BR>
<H3>
<IMG SRC=http://www-doc-rocq.inria.fr:8004/bm/images/sep/ligne.gif><BR><BR>
Pour obtenir une <EM>table des mati&egrave;res</EM>, vous pouvez
interroger &agrave partir du :
<UL>
<LI>Titre de la <A HREF="rechconference.html">CONFERENCE</A>
<LI>Titre du <A HREF="rechperiodique.html">PERIODIQUE</A>
<LI>Titres de <A HREF="wais://merengue.inria.fr:5280/archi-titre-conf-et-perio">
CONFERENCES et de PERIODIQUES</A>
</UL><BR>
<IMG SRC=http://www-doc-rocq.inria.fr:8004/bm/images/sep/ligne.gif><BR><BR>
Pour obtenir un <EM>article</EM>, vous pouvez interroger &agrave partir  :
<UL>
<LI>Des <A HREF="wais://merengue.inria.fr:5280/archi-sommaires-conferences-et-
periodiques">TERMES PERTINENTS</A> de l'article :
<BR><BR>- Auteur, Date, Mots du titre...</A>
</UL>
<IMG SRC=http://www-doc-rocq.inria.fr:8004/bm/images/sep/ligne.gif><BR><BR>
Pour obtenir une <EM>notice bibliographique</EM>, vous pouvez interroger &agrave
 partir  :
<UL>
<LI>Des <A HREF="http://merengue:8005/cgi-bin/formulaire-nouveautes-inriatheque.pl">
TERMES PERTINENTS</A> des ouvrages, des rapports et des th&egrave;ses :
<BR><BR>- Auteur, Date, Mots du titre...</A>
</UL>
<IMG SRC=http://www-doc-rocq.inria.fr:8004/bm/images/sep/ligne.gif><BR><BR>
<H3>
Retour &agrave la <A HREF="maquette.html"><EM>page de garde</EM></A> de
l'Inriath&egrave;que.
</H3>
</BODY>
</HTML>
```

```
<HTML>
<HEAD>
<A NAME=DEBUT>
<TITLE>nouveautes.html</TITLE>
</HEAD>
<BODY BACKGROUND="cj.jpg">
<FONT size=5>
<B>
<CENTER>
<EM>CATALOGUE DES NOUVEAUTES<BR><BR>
</CENTER>
<H4>Les <A HREF="information.html">informations</A> ...</H4>
</FONT><BR>
<H3>
<IMG SRC=http://www-doc-rocq.inria.fr:8004/bm/images/sep/ligne.gif><BR><BR>
Pour obtenir une <EM>table des mati&egrave;res</EM>, vous pouvez
interroger &agrave partir du :
<UL>
<LI>Titre de la <A HREF="rechconference.html">CONFERENCE</A>
<LI>Titre du <A HREF="rechperiodique.html">PERIODIQUE</A>
<LI>Titres de <A HREF="wais://merengue.inria.fr:5280/new-titre-conf-et-perio">
CONFERENCES et de PERIODIQUES</A>
</UL><BR>
<IMG SRC=http://www-doc-rocq.inria.fr:8004/bm/images/sep/ligne.gif><BR><BR>
Pour obtenir un <EM>article</EM>, vous pouvez interroger &agrave partir  :
<UL>
<LI>Des <A HREF="wais://merengue.inria.fr:5280/new-sommaires-conferences-et-periodiques">
TERMES PERTINENTS</A> de l'article :
<BR><BR>- Auteur, Date, Mots du titre...</A>
</UL>
<IMG SRC=http://www-doc-rocq.inria.fr:8004/bm/images/sep/ligne.gif><BR><BR>
Pour obtenir une <EM>notice bibliographique</EM>, vous pouvez interroger &agrave partir  :
<UL>
<LI>Des <A HREF="http://merengue:8005/cgi-bin/formulaire-nouveautes-inriatheque.pl">
TERMES PERTINENTS</A> des ouvrages, des rapports et des th&egrave;ses :
<BR><BR>- Auteur, Date, Mots du titre...</A>
</UL>
<IMG SRC=http://www-doc-rocq.inria.fr:8004/bm/images/sep/ligne.gif><BR><BR>
<H3>
Retour &agrave la <A HREF="maquette.html"><EM>page de garde</EM></A> de l'Inriath&egrave;que
</H3>
</BODY>
</HTML>
```


File Edit View Go Bookmarks Options Directory Help
Back Home Reload Images Open Print Find
Location: http://merengue:8005/kebir/inriatheque/maquette/commentaire.html
– Nous vous remercions pour ce message –
Hamza Kebir hamza.kebir@inria.fr
Votre opinion sur l'Inriatheque : TRES BIEN
Vos coordonnees :
NOM :
E-Mail :
ADRESSE :
A propos de votre suggestion :
SUJET :
COMMENTAIRES :
ENVOYER LE MESSAGE : OK
TOUT REMETTRE A BLANC : OK
Document Done

File Edit View Go Bookmarks Options Directory Help

Back Forward Home Reload Images Open Print Find Stop

Location: http://merengue:8005/cgi-bin/formulaire-nouveautes-inriatheque.pl

Vous pouvez aussi interroger le fonds documentaire global Inria (sauf périodiques), ou le fonds des notices de périodiques.

Interrogation de l'Inriathèque numérisée

Id : E-Mail:

Nb de réponses retournées (par base interrogée) : ◆ 20 ◇ 40 ◇ 60 ◇ 100

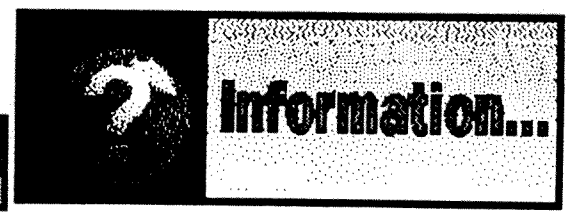
Recherche :

Recherche Reset

Sous Mosaic :
Pour désélectionner une base appuyez sur la touche Ctrl en cliquant.
Pour ne sélectionner qu'une seule base, cliquez sur son nom.

Sous Netscape :
Pour sélectionner/désélectionner une base, cliquez sur son nom.

Nouveautes-Articles-Inriatheque
Nouveautes-Notices-Inriatheque
Archives-Notices-Inriatheque



B - Bases WAIS

Avec la prolifération des sources d'informations, on a vu fleurir au cours de ces dernières années des outils permettant de découvrir, rechercher, rapatrier et visualiser des informations localisées sur le réseau Internet, WAIS est celui que nous avons utiliser au cours de ce stage pour indexer les données de l'Inriathèque électronique.

WAIS (Wide Area Information System) a été développé par Thinking Machines Corporation. La distribution et les versions futures ont maintenant été reprises par le CNIDR (Clearinghouse for Networked Information Discovery and Retrieval). WAIS est un logiciel freeware (gratuit) disponible sur le réseau Internet.

Le logiciel d'interrogation WAIS est bâti sur le modèle client/serveur et s'appuie sur une variante du protocole ANSI Z39.50 avec un transport TCP.

Protocole Z39.50 :

Ce protocole, conçu pour uniformiser les dialogues entre un client et un serveur dans le cadre d'applications documentaires, s'inscrit dans la philosophie du modèle OSI de l'ISO, comme un protocole d'application. Il s'appuie sur des dialogues en mode connecté et fournit des services pour :

⇨ l'initialisation avec contrôle de la communication client/serveur INIT;
⇨ le déclenchement de recherches (Search)
⇨ la récupération de résultats (Present)
⇨ la terminaison de la communication

Z39.50 voit une base de données comme un ensemble de fichiers contenant des documents. Chaque document peut être structuré en champs. Z39.50 permet le codage de question de divers types. Certains types autorisent l'emploi d'opérateurs booléens et de proximité.
Le protocole actuellement utilisé admet des questions de Type 3 non prévu dans la norme. Ces questions sont des suites de mots cités explicitement ou contenus dans des portions de documents. C'est un "ou" logique qui est pris comme opérateur implicite. L'ensemble des résultats est effacé à chaque question.

Le logiciel WAIS :

Les logiciels WAIS se composent d'une partie client qui tourne sur la machine émettant des requêtes, d'une partie serveur qui tourne sur la machine abritant la base de données et d'une partie indexation qui permet de constituer la base d'informations.

Indexation :

Le logiciel d'indexation Wais-Index est un programme C qui va parcourir une hiérarchie de fichiers contenant des documents à indexer et constituer un index avec tous les mots de deux caractères et plus. Lors de l'indexation, on va préciser un type qui va déterminer :

⇨ les règles de découpage des fichiers en documents; ainsi un document sera, par exemple, une ligne pour le type "one-line", ou un paragraphe pour le type "para", ou un message pour le type "mail digest";

⇨ la façon de constituer la ligne d'en-tête (header line); cette ligne est rendu comme première réponse aux questions; par exemple pour le type "mail", la ligne d'en-tête contiendra les champs "from" et "subject".
⇨ le format du document : texte, GIF, TIFF, etc.; ce type sera fourni au client lors de la visualisation du document; ceci lui permettra d'appliquer le filtre de visualisation de son choix.

Il existe un nombre important de types prévus en standard et il est facile de rajouter de nouveaux types en modifiant deux modules du programme d'indexation. La version officielle indexe les mots contenus dans le documents pour tous les types contenant du texte (mail, news, para, one-line...) et se contente d'indexer les mots contenus dans le nom du fichier pour les autres types (postscript, GIF...).

L'index comporte une entrée par mot. Chaque entrée contient le poids et les positions du mot dans les divers documents. Le poids d'un mot est calculé de la façon suivante : 5 points pour la première occurrence, plus 1 point à chaque nouvelle occurrence du mot dans le document, ou 10 points dans chaque occurrence du mot dans la ligne d'en-tête.
Il est également possible de maîtriser complètement l'indexation en utilisant des extensions qui consistent à utiliser un filtre externe pour l'indexation. Lors de l'indexation, on précise à Wais-Index (argument-filter) le nom d'un filtre qui va être un programme ou un script lancé par Wais-Index. Wais-Index va communiquer au filtre les noms des fichiers à indexer et va attendre du filtre, pour chaque document, la ligne d'en-tête, la date du document et les mots clés avec leur poids et leur position. Avec cette technique, on va pouvoir maîtriser totalement l'indexation des fichiers texte et réaliser des indexations de fichiers qui sont dans d'autres formats (Postscript, images...) en tirant les mots clés de fichiers explicatifs associés ou bien en lançant dans le filtre des convertisseurs en ASCII des fichiers à indexer (tels post-Ascii ou un module OCR).

Les tailles des index dépendent essentiellement de la nature des fichiers indexés. Cette taille peut varier de 0,03 fois (cas des fichiers postscript avec images) à 3 fois (cas de références bibliographiques) la taille des fichiers à indexer.
La vitesse d'indexation dépend de facteurs multiples tels que la puissance du CPU, la taille mémoire et l'architecture de la machine. Sur la base de références bibliographiques de l'INRIA, on indexe 4000 références de 0,5 ko par minute sur un Sun SS10. Ces performances d'indexation élevées permettent d'envisager des indexations totales et de pallier les déficiences du mode mise à jour qui ne permet que de rajouter des documents nouveaux.

Client :

Les index WAIS vont être consultés à distance par un logiciel client. Celui-ci fonctionne sur le principe suivant; l'utilisateur dispose de trois répertoires dans son environnement :

⇨ *Wais-Sources* contient les documents .src donnant les coordonnées de sources intéressant l'utilisateur;
⇨ *Wais-Questions* contient le texte de questions déjà posées et prêtes à l'emploi;
⇨ *Wais-Documents* contient les documents sauvegardés lors d'interrogations précédentes;

L'utilisateur va interroger une ou plusieurs de ses sources en donnant une liste de mots. Une fois les sources sélectionnées et les mots de la questions tapés, le client va établir la connexion TCP avec le serveur, ouvrir la communication Z39.50 et soumettre la question. Le serveur va consulter l'index et déterminer les documents répondant à la question. Pour chaque document, le serveur va calculer une note en additionnant les poids des divers mots de la question et en ramenant les chiffres obtenus à une échelle de 1 à 1000. Dans Free-Wais, le nombre d'occurrences du mot dans la base et la taille du document vont intervenir dans le calcul du poids de façon à ne pas défavoriser les

documents courts. Le serveur va présenter au client une liste triée et limitée (40 par défaut) des documents les plus pertinents.

Pour chaque document, c'est la ligne d'en-tête, la note et la taille du document qui seront indiquées. Le client va pouvoir visualiser tout ou partie des documents en les sélectionnant. Le type (format) du document est communiqué au client qui pourra s'en servir pour appeler l'outil de visualisation le plus approprié. Dans la version X-Window du client, l'association des filtres aux divers types de documents se fait par une ressource X.

L'ensemble des index créés grâce à WAIS ont été présentés précédemment (Cf. Section 2 : Les procédés de traitements des données, A - Chaîne de traitement du document, 3 - Indexation)

C - HyperText Transfer Protocole (HTTP)

L'Inriathèque est amenée à être interrogée à l'extérieur, dans un premier temps, il s'agira des chercheurs de l'INRIA à partir de leur bureau mais à plus long terme, la consultation se fera par un autre serveur qui sera considéré comme un client. La connexion se fera dans ce cas là, en mode client/serveur et le protocole permettant de mettre l'Inriathèque à disposition du client sera HTTP.

Le WEB fonctionnant en client/serveur, il est nécessaire qu'un protocole commun permette les échanges entre le client (NETSCAPE, MOSAIC ...) et le serveur WEB. Le protocole HTTP fonctionne selon le principe suivant :

⇨Le client demande une connexion vers le serveur à l'aide d'une requête GET avec pour paramètre le document demandé, en l'occurrence l'Inriathèque électronique.

⇨ Le serveur accepte la connexion et fournit au client le document dont l'adresse est l'URL demandé.

⇨ Le serveur coupe la connexion (Il n'y a pas de session permanente entre un client W3 et un serveur, le client et le serveur n'occupent la ligne que durant le transfert des données)

D - Expérimentation de l'Inriathèque

1 - Echantillon de chercheurs

Les permanences au centre de documentation m'ont permis de rencontrer de nombreux chercheurs, je n'ai eu aucun mal à constituer un échantillon de personnes volontaires pour expérimenter de manière concrète l'Inriathèque. Ceux-ci appréciaient volontiers l'offre proposée et estimaient d'ores et déjà qu'il s'agissait d'un projet d'une grande utilité.

2 - Modifications

Après s'être connectés à l'Inriathèque, ils ont commencé à explorer les fonctionnalités de l'outil. Très vite, ils ont transmis leurs premiers avis et suggestions. Certains termes n'étaient pas suffisamment explicités. De nombreuses modifications ont été apportées. Ils se sont montrés impatients de pouvoir utiliser cette interface à partir de leur bureau et faire une demande de photocopie d'articles par ce biais.

E - Serveur WWW

L'Inriathèque électronique devra à l'avenir être insérée sur le serveur W3 de l'INRIA. Pour cela, de nombreuses questions, notamment celles des droits de copie devront être étudiées et solutionnées. C'est à partir de la plaquette du centre de documentation que l'on pourra accéder celle de l'Inriathèque électronique.
Le WEB utilise la technique de l'HyperText, c'est à dire qu'il utilise des liens entre les documents pouvant être n'importe où sur l'Internet. De plus, les documents ainsi référencés peuvent être accessibles par différents protocoles (HTTP, FTP, GOPHER...). Ces documents peuvent non seulement être des fichiers mais aussi, l'exécution d'un programme, ou encore comme le feront les utilisateurs de l'Inriathèque électronique, l'interrogation d'une base de données.

Le WEB utilise le modèle classique client/serveur (comme FTP et TELNET, mais les rôles respectivement du client et du serveur sont moins évidents). Un serveur WWW est un programme qui tourne sur un ordinateur dans le but de répondre aux requêtes de logiciels clients WWW fonctionnant sur d'autres ordinateurs.

L'une des propriétés importantes du fait de l'exécution de programme sur le serveur, est la possibilité d'interfaçage de WWW sur n'importe quel logiciel.

Hypertext et hypermédia :

Un document tel que l'Inriathèque électronique est un hypertext c'est à dire un fichier de texte normal avec une différence importante : il comporte dans son texte des liens soit vers d'autres parties du document lui même, soit vers d'autres documents appartenant à un autre ordinateur du réseau.

Un lien hypertext, ou hyper-lien est formé par une ancre et par l'adresse du document ciblé. Une ancre peut être un mot (ou groupe de mots) ou une image mis en évidence (caractère gras, encadrement...) dans le document. Dans un document hypertext, une ancre peut servir soit comme origine d'un lien, soit comme destination d'un lien.
Un document référencé par un lien hypertext peut être sur tout type de serveur, son adresse devra donc, entre autre, indiquer la méthode d'accès (méthode d'adressage)

Un document hypermédia est un hypertext avec la différence que les liens peuvent référencer également des fichiers sons, images ou vidéo. Des images peuvent être incluses dans un fichier hypermédia et peuvent elles-mêmes servir pour référencer d'autres documents.
Le langage standard utilisé par W3 pour créer et reconnaître les documents hypermédia s'appelle HyperText Markup Language (HTML).

Il n'y a pas de session permanente entre un client W3 et un serveur (W3, FTP...). Le langage utilisé entre un client W3 et un serveur W3 s'appelle HyperText Transmission Protocol (HTTP).

F - Diffusion et reproduction de documents à partir de l'Inriathèque

La diffusion de documents dont nous ne sommes pas les auteurs soulève une question de droit. En effet, il n'y aucune entorse à la loi à partir du moment où un document est exploité personnellement, mais dès qu'il s'agit d'en faire une exploitation au titre d'un organisme tel que l'INRIA, nous devons penser à toutes les modalités d'application dans ce domaine, obtenir les autorisations adéquates et peut être payer des droits.

Le CFC (Convention sur le Droit de la photocopie) :
Les photocopies effectuées et vendues par le centre de documentation à des utilisateurs extérieurs font l'objet d'une redevance annuelle de 75000 francs H.T., versée au Centre français d'exploitation du droit de copie. De plus, régulièrement la liste des documents qui ont fait l'objet de photocopies d'articles payantes lui est transmise.

Le travail réalisé au cours de ce stage n'est actuellement exploité que par le site de Rocquencourt. L'expérience d'une telle démarche pourra conduire à en étendre l'utilisation aux autres sites l'INRIA.
De nombreux questions restent en suspens, notamment auprès de quel centre de documentation seront faites les demandes de photocopies. Il est légitime que le centre de documentation de Rocquencourt ne veuille pas assumer seul la charge de ce travail. En effet, étant donné que notre centre de documentation possède le fonds documentaire le plus important, la conséquence logique en serait une quantité accrue de demandes de photocopies.

Il faudra tirer les conclusions de la première expérience au niveau interne de l'INRIA pour envisager l'implémentation d'un tel service pour les abonnés extérieurs.
Pour l'ensemble des abonnés extérieurs, l'utilisation de l'Inriathèque fait apparaître de nouvelles contraintes. Ces questions devrons notamment être étudiées avec les fournisseurs d'abonnement. Notons par exemple, que les abonnés extérieurs devront posséder un compte de photocopie auprès du centre de documentation. Il serait impossible d'accepter une demande sans identification ou vérification préalable d'identité et de compte de photocopies.

CONCLUSION

Le travail a consisté entre autres à collecter, sélectionner et contacter des producteurs d'information pour bénéficier d'un abonnement aux sommaires de périodiques et de conférences. Il s'est également agit d'étudier et exploiter au mieux les fonctionnalités du scanner et du logiciel de Reconnaissance Optique de Caractères pour pouvoir numériser les sommaires ne pouvant être obtenus par abonnement. Il a enfin fallu mettre sur pieds un système de Gestion Electronique de Documents complet et cohérent mettant à disposition du public l'intégralité de ces données.

Le but était de construire une plateforme de travail, celle-ci a été réalisée mais n'est pas complètement finie. A présent, le système mis en place va logiquement tendre vers plus de précision, il faudra également que des choix soient faits notamment en ce qui concerne l'abonnement aux sommaires et les accès à l'Inriathèque électronique à partir de l'extérieur.

Les enseignements théoriques reçus au cours de l'année universitaire, tels que le système d'exploitation UNIX ou la numérisation pour ne citer qu'eux, ont pu être approfondis et maîtrisés.

Le bilan de ces quatre mois de stage est largement positif. En effet, ce stage constituait ma première véritable expérience professionnelle, j'ai eu une véritable approche du monde de l'entreprise avec notamment une demande concrète, établie sur un besoin réel que ne peut pas offrir le milieu universitaire.

J'ai notamment appris à organiser mon temps de travail. En effet, cette tâche était pour le moins problématique en début de stage. Gérer son emploi du temps devint très vite une tâche primordiale.

Malheureusement, le travail et l'approche du projet Inriathèque auraient pu être mené réellement à bout, si la période de stage convenue n'était pas si courte. Cette période de stage en entreprise a répondu à mon attente et m'a permis d'évaluer mes capacités à gérer un projet.

BIBLIOGRAPHIE

⇨ GOLDWASER, Daniel, LENART, Michèle. Applications documentaires de la GED dans les bibliothèques et centres de documentation. Paris : A jour, 1993. ISBN 2-903685-50-9

⇨ BERTRAND, Roland. Micro-ordinateur et traitement de l'information. Paris : A jour, 1993. ISBN 2-903685-36-3

⇨ INRIA. Le traitement électronique du document, cours INRIA, 3-7 octobre 1994, Aix en Provence. Paris : ADBS, 1994. ISBN 2-901046-76-2

⇨ INRIA. Le document électronique, cours INRIA dirigé par Christian Bornes, 11-15 juin 1990, Châtelaillon. Rocquencourt : INRIA, 1990. ISBN 2-7261-0619-6

⇨ LEMAY, Laura. Teach yourself WEB Publishing with HTML in a week. Indianapolis : SAMS, 1995. ISBN 0-672-30667-0

⇨ CNRS & Universités. Internet professionnel, témoignages, expériences, conseils pratiques de la communauté enseignement et recherche. Paris : CNRS. ISBN 2-271-0525-6

⇨ BOURNE, Steeve. Le système unix. Paris : InterEditions, 1991. ISBN 0-7296-0014-0

⇨ SCHWARTZ, Randal. Learning Perl. Sebastopol, CA : O'Reilly, 1993. ISBN 1-56592-042-2

⇨ Cahier Gutenberg. Diffusion des documents électroniques : de Latex à WWW, HTML et Acrobat, n° 19, janvier 1995. ISSN1140-9304

⇨ PELIKS, Gerard. Le World Wide Web, création de serveurs sur Internet. Paris : Addison-Wesley, 1995. ISBN 2-87908-102-5

INDEX